



**IQ-2001**

**Cambridge, Massachusetts, USA**

**Proceedings of  
the Sixth International  
Conference on  
Information Quality**

Edited by

**Elizabeth M. Pierce**

**Indiana University of Pennsylvania**

and

**Raïssa Katz-Haas**

**Ingenix/United Health Group**

## **WELCOME FROM THE CONFERENCE CO-CHAIRS**

The MIT Conference on Information Quality (IQ) has been instrumental in establishing a premier forum for researchers and practitioners of information quality. The conference started in the classrooms at MIT, Cambridge in 1996. The conference grew along with and because of the participants from many international communities. Today's gathering marks the conference's 6<sup>th</sup> year as well as the adoption of its new name: The 6<sup>th</sup> *International Conference on Information Quality*, commonly known as IQ-2001. The conference provides a stimulating and unique forum for promoting the exchange of experience and knowledge about IQ research and practice. Its importance is reflected in the strong support each of the previous five conferences has received.

Acknowledgments are extended to all conference participants. An applied field such as information quality demands active interaction and collaboration between practitioners and researchers for its growth. This collaboration has been a major driving force in advancing the IQ field. We thank you for your contribution in establishing IQ as a multi-disciplinary field and pushing the research to become even more useful and practice-oriented. Accordingly, the newly established practitioner Program Co-chairs, RaïssaKatz-Haas for IQ-2001 and Bruce Davidson for IQ-2002, will begin a new tradition of actively encouraging practitioners to report their organization's experiences.

We also wish to acknowledge the support of many sponsors, particularly the MIT Total Data Quality Management Research Program, the Center for Information Technologies and Market Transformation at Berkeley, Information Quality Coalition, Marist College, Georgia Public Health Division, Data Quality Journal, FirstLogic Inc., Group 1 Software, and Cambridge Research Group. We also greatly appreciate Mr. Dane Iverson of Ingenix Inc. who will give the keynote speech.

Special thanks are due to Professor Elizabeth Pierce and Ms. Raïssa Katz-Haas for producing an outstanding conference program for IQ2001 and responding to the never-ending e-mails since early spring. We appreciate Professor Leo Pipino for arranging the plaques and helping with many other aspects of the conference. Thanks are also due to Professors Richard Wang and Stuart Madnick for overseeing the process of preparing for the conference and answering emails at lightning speed. We would also like to thank Tom Maglio, John Maglio, Jerry Cai, Kerry Brennan, and Fori Wang, for their assistance in producing the proceedings, managing the MIT TDQM website, and assisting the conference site operation.

As the Co-Chairs of the conference, we welcome you to *the 6<sup>th</sup> International Conference on Information Quality* at MIT, Cambridge.

Yang Lee and James Funk  
Conference Co-Chairs

## **WELCOME FROM THE PROGRAM CO-CHAIRS**

Welcome to the *6th International Conference on Information Quality*. This year we are pleased to showcase over 36 presentations by practitioners and academics from ten different countries covering all aspects of information quality. Conference sessions showcasing the rigorously reviewed research papers and practice-oriented papers are organized by topic into parallel sessions. A sample of the information quality topics covered in this conference include IQ organizational dynamics, measuring and improving information quality, the Internet, data warehousing and data mining, and decision support. We hope that all participants find the conference to be a worthwhile and educational experience.

Members of the program committee reviewed the papers submitted to the conference and provided feedback to the authors of the papers. We would like to thank the members of the program committee for their excellent reviews. The contributions of the program committee made our work on the conference program a pleasure.

### IQ-2001 Program Committee

Adenekan Dedek, Suffolk University  
Andreas Neus, IBM Germany  
Barbara Klein, University of Michigan at Dearborn  
Beverly Kahn, Suffolk University  
Bruce Davidson, Cedars-Sinai Health System  
Burton Cutting, Caxton Associates  
Craig Fisher, Marist College  
Diane Strong, Worcester Polytechnic Institute  
Don Ballou, State University of New York at Albany  
Don Rossin, University of Michigan at Dearborn  
Felix Naumann, Humboldt University Berlin  
Frank Dravis, Firstlogic  
Ganesan Shankar, Boston University  
Richard McCarthy, Central Connecticut State University  
Giri Kumar Tayi, State University of New York at Albany  
InduShobha Chengalur-Smith, State University of New York at Albany  
James Funk, S. C. Johnson  
Jennifer Long, Canadian Institute for Health Information  
Leo Pipino, UMASS, Lowell  
Marc Rittberger, University of Konstanz  
Mike Hanosh, Intel - Data Quality Program  
Melissa Tzourakis, Ingenix / United Health Group  
Olayele Adedokun, DePaul University  
Tamraparni Dasu, AT&T Labs  
Vassilios Verykios, Drexel University  
Yang Lee, Northeastern University

The final conference program and other information about the International Conference on Information Quality are available at <http://web.mit.edu/tdqm/>.

Elizabeth M. Pierce, Indiana University of Pennsylvania  
Raissa Katz-Haas, Ingenix / United Health Group  
Program Co-Chairs, the 6<sup>th</sup> International Conference on Information Quality (IQ 2001)

**IQ-2001 Conference Program Schedule**

**FRIDAY, November 2**

**5:00 - 6:30**      **Registration and Reception at MIT Faculty Club**      **E52**  
 Informal Discussions      (6<sup>th</sup> Floor)

**SATURDAY, November 3**

**8:30 - 9:00**      **Registration and Continental Breakfast**      **E51-345**

**9:00 - 9:20**      **Chairs' Welcome and Opening Remarks**      **E51-345**  
 Yang W. Lee, Conference Co-Chair,  
 Northeastern University  
 James D. Funk, Conference Co-Chair, S.C. Johnson  
 Elizabeth M. Pierce, Program Co-Chair,  
 Indiana Univ. of Pennsylvania  
 Raïssa Katz-Haas, Program Co-Chair,  
 Ingenix/UnitedHealth Group

**9:20 - 10:00**      **✓ Keynote Speech**      **E51-345**  
 Dane Iverson, Senior Vice President  
 Ingenix/UnitedHealth Group

10:00 - 10:15      Coffee Break

**10:15 - 11:45**      **1A Parallel Paper Session**      **E51-372**  
**IQ and Organizational Dynamics**  
*Session Chair: Bruce Davidson*  
*(IQ-2002 Program co-chair)*  
**Storytelling as a Management Tool** (Practice-Oriented  
 Paper) Bruce Davidson, Cedars-Sinai Health System  
**Generations of Information Quality** (Practice-Oriented  
 Paper) Frank Dravis, Firstlogic, Inc.  
**Organizational Realism Meets Information Quality**  
**Idealism: The Challenges of Keeping an Information**  
**Quality Initiative Going** (Practice-Oriented Paper)  
 Beverly K. Kahn, Suffolk University  
 Raïssa Katz-Haas, Ingenix/UnitedHealth Group  
 Diane M. Strong, Worcester Polytechnic Institute

**10:15 - 11:45**      **1B Parallel Paper Session**      **E51-376**  
**Improving Data Warehouse Source Data**  
*Session Chair: Craig Fisher, Marist College*  
*(IQ-2002 Program co-chair)*

Friday, Nov 2—Saturday, Nov 3

**A Proposed Framework for the Analysis of Source Data in a Data Warehouse (Research Paper)**

M. Pamela Neely, Marist College

**External Data Selection for Data Mining in Direct Marketing (Practice-Oriented Paper)**

Dirk Arndt, Daimler Chrysler, Germany

Wendy Gersten, Daimler Chrysler, Germany

**A Strategy for Managing Data Quality in Data Warehouse Systems (Research Paper)**

Markus Helfert, University of St. Gallen, Switzerland

Eitel von Maur, University of St. Gallen, Switzerland

11:45 - 1:15

**✓ Lunch speeches: TDQM Research Initiatives**

E51-345

*Introduction:* Rich Wang, Boston University & MIT  
TDQM Program

**Re-manufacture and Information Products**

Tom Allen, Howard Johnson Professor,  
MIT Sloan School of Management

**Data Quality Challenges in Enabling eBusiness Transformation**

Arie Segev, Professor and CITM Director, U.C. Berkeley

**Corporate Household Data**

Stuart Madnick: John Norris Maguire Professor,  
MIT Sloan School of Management

1:15 - 2:45

**2A Parallel Paper Session: IQ and the Internet**

E51-372

*Session Chair:* Giri Kumar Tayi, SUNY Albany

**Accessing the Quality of Online Classified Websites: An Empirical Study of the 100 Largest Newspapers (Research Paper)**

Adenekan Dedeke, Suffolk University

Beverly Kahn, Suffolk University

**Managing Information Quality in Virtual Communities of Practice (Practice-Oriented Paper)**

Andreas Neus, IBM Unternehmensberatung GmbH,  
Germany

**The Issue of IQ in Internet-Based Early-Warning Systems for Trend Management (Practice-Oriented Paper)** Daniel Diemers, SFS-HSG, Switzerland

1:15 - 2:45

**2B Parallel Paper Session: Assessing the Value of Information** *Session Chair:* Frank Dravis, Firstlogic Inc.

E51-376

**An In-Depth Investigation into the Impact of Information Quality Upon the Perceived Value of Information** (Research Paper)

Graham Doig, Loughborough University, U.K.

Neil Doherty, Loughborough University, U.K.

Chris Marples, Loughborough University, U.K.

**Data Quality in the Small: Consumer Information** (Research Paper)

Arnon S. Rosenthal, The MITRE Corporation

Donna M. Wood, The MITRE Corporation

Eric R. Hughes, The MITRE Corporation

Mary C. Prochnow, The MITRE Corporation

**Quality Mining: A Data Mining Based Method for Data Quality Evaluation** (Research Paper)

Sabrina Vazquez Soler, University of Buenos Aires, Argentina

Daniel Yankelevich, University of Buenos Aires, Argentina

1:15 - 2:45

**2C Parallel Paper Session: Reconciling Data**

E51-335

*Session Chair: Felix Naumann, IBM Almaden*

*Research Center*

**An Approximate Matching Technology for Database Searching, Linking, and De-Duplicating** (Practice Oriented Paper)

Arthur Goldberg, Choice Maker

Andrew Borthwick, Choice Maker

**Cleaning Up Very Large Databases and Keeping Them Clean** (Practice Oriented Paper)

Priscilla Broberg, Consultant, Agilent Technologies

**Reconciling the Data Warehouse** (Practice Oriented Paper) Jonathan Wu, BASE Consulting Group

2:45 - 3:00

Coffee Break

3:00 - 4:30

**3A Parallel Paper Session: Implementing DQ Processes** E51-372

*Session Chair: Burton Cutting, Caxton Associates*

**Monitoring and Data Quality Control of Financial Databases from a Process Control Perspective** (Practice-Oriented Paper)

Janusz Milek, Predict AG, Switzerland

Martin Reigrotzki, Predict AG, Switzerland

Holger Bosch, Predict AG, Switzerland

Frank Block, Predict, AG, Switzerland

**The Implementation of Information Quality for the Automated Information Systems in the TDQM Process: A Case Study in Textile and Garment Company in Thailand** (Research Paper)

Athakorn Kengpol, King Mongkut's Institute of Technology, Thailand

**A Methodological Approach to Data Quality Management Supported by Data Mining** (Research Paper)

Udo Grimmer, DaimlerChrysler, Germany  
Holger Hinrichs, Oldenburg Research and Development Institute for Computer Science Tools and Systems, Germany

3:00 - 4:30

**3B Parallel Paper Session: Improving Quality of Searches & Queries**

E51-376

*Session Chair: Beverly Kahn, Suffolk University*

**A Data Quality Browser** (Research Paper)

Theodore Johnson, AT&T Labs  
Tamraparni Dasu, AT& T Labs

**From Databases to Information Systems - Information Quality Makes the Difference** (Research Paper)

Felix Naumann, IBM

3:00 - 4:30

**3C Parallel Paper Session: IQ and the Decision Making Process**

E51-335

*Session Chair: Diane Strong, WPI*

**Conceptual Ideas Underlying the Information Engineering Approach for Decision Making in Textiles** (Research Paper)

Yatin Karpe, North Carolina State University  
George Hodge, North Caroline State University  
Neil Cahill, Institute of Textile Technology  
William Oxenham, North Carolina State University

**An Assessment of the Theory Underpinning the Role of Information Quality in the Single-Loop Decision Making Model** (Research Paper)

Raul M. Abril, Brunel University, U.K.

**Information Envelope and its Information Integrity Implications** (Research Paper)

Vijay Mandke, Unitech Systems, India  
Madhavan K. Nayar, Unitech Systems, India  
Kamna Malik, Institute of Management Technology, India

4:30 - 6:00	<p><b>✓ Plenary Panel: Progress in Information Quality: Why So Slow?</b>  <i>Moderator: James D. Funk, SCJ</i></p> <p><b>Panelists:</b>                  Thomas C. Redman, Navesink Consulting                  Larry English, Impact International                  Tony Tortorice, Predictive Modeling LLC                  Ken Orr, The Ken Orr Institute                  Stuart E. Madnick, MIT Sloan School of Management</p>	51-345
-------------	---	--------

**SUNDAY, November 4**

8:30 - 9:00	<p><b>Registration and Continental Breakfast</b></p>	E51-345
9:00 - 10:30	<p><b>4A Parallel Paper Session: Defining Information Quality</b>  <i>Session Chair: Leo Pipino, UMASS, Lowell (IQ 2002 Publicity Chair)</i></p> <p><b>A Conceptual Framework and Belief Function Approach to Assessing Overall Information Quality</b>                  (Research Paper)                  Matthew Bovee, The University of Kansas                  Rajendra P. Srivastava, The University of Kansas                  Brenda Mak, The University of Kansas</p> <p><b>A Generic Framework for Information Quality in Knowledge Intensive Industries</b> (Research Paper)                  Martin Eppler, University of St. Gallen, Switzerland</p> <p><b>A College Course: Data Quality in Information Systems</b> (Research Paper) Craig Fisher, Marist College</p>	E51-372
9:00 - 10:30	<p><b>4B Parallel Paper Session: Improving IQ in the Healthcare Industry</b>  <i>Session Chair: Yang Lee, Northeastern University</i></p> <p><b>Overview of Deloitte &amp; Touche's Approach to Information Quality</b> (Practice Oriented Paper)                  John Gimpert, Deloitte &amp; Touche                  Tim Krick, Deloitte &amp; Touche</p> <p><b>Data Quality and Medical Record Abstraction in the Veterans Health Administration's External Peer Review</b> (Practice Oriented Paper)                  James H. Forsythe, West Virginia Medical Institute                  Jonathan Perlin, VHA Office of Quality and Performance                  John Brehm, West Virginia Medical Institute</p>	E51-376

Saturday, Nov 3— Sunday, Nov. 4



- The Canadian Institute for Health Information Data Quality Framework, Version 1: A Meta-Evaluation and Future Directions** (Practice Oriented Paper)  
J. A. Long, Canadian Institute for Health Information  
J. A. Richards, Canadian Institute for Health Information  
C. E. Seko, Statistics Canada
- 10:30 - 10:45 Coffee Break
- 10:45-12:15 **5A Parallel Paper Session: Assessing Information Quality** E51-372  
*Session Chair: Mostapha Ziad, Suffolk University*  
**Tracking the Physical and Information Product Flows in Mobile Patient Service Supply Chain: A Real Vision Lab Approach** (Research Paper)  
P. Balasubramanian, Boston University  
G. Shankaranarayan, Boston University  
R. Wang, Boston University  
**Non-Intrusive Assessment of Organizational Data Quality** (Research Paper)  
Binling Jin, University of Manchester, U.K.  
Suzanne M. Embury, University of Manchester, U.K.  
**Using Control Matrices to Evaluate Information Production Maps** (Research Paper)  
Elizabeth Pierce, Indiana University of Pennsylvania
- 10:45 - 12:15 **5B Parallel Paper Session: Issues in Information Quality** E51-376  
*Session Chair: Jim Hurysz, Data Quality*  
**Data Quality Issues in Service Provisioning & Billing** (Research Paper)  
Tamraparni Dasu, AT&T Labs  
Theodore Johnson, AT&T Labs  
**Introducing Data Quality in a Cooperative Context** (Research Paper)  
Paola Bertolazzi, IASI-CNR, Italy  
Monica Scannapieco, University of Rome, Italy  
**Information Quality and Large Scale Project Budget Tracking** (Practice Oriented Paper)  
Viktor Dvurechenskikh, Comptroller of Moscow, Russia  
Vladimir Baranov, Comptroller of Moscow, Russia  
George Huntington, Consultant
- 12:15 **✓ End of IQ-2001 Conference**

Sunday, Nov. 4

# Meta-Information Quality

Abstract For Keynote Address

Dane S Iverson

Sr. Vice President Enterprise Information Solutions - Ingenix

Data and information quality is based on several factors. Some of the basic requirements for information quality are comprised of data that must be accurate, complete, timely, accessible and understood for use by knowledge workers. A prerequisite for these factors is the quality of the meta-data. The focus of the keynote address will discuss how metadata enables information quality.

As a foundation metadata is the semantic structure, definitions, and descriptive attributes about the data that need to be communicated and used to enable information quality. To improve information quality the metadata attributes must go beyond the typical data dictionary names and definitions with technical details. The metadata should include process and procedure documentation such as data capture, storage, and transformation rules for technical support. To support end user's data understanding and information quality initiatives it should document quality and usage metrics, examples and data use tips and many others meta-data attributes that we will discuss. The purpose of metadata is to communicate, educate, and facilitate the use of the data.

The challenge is to make metadata useful and available when needed by information handlers and knowledge workers. The metadata must be concise, rich in detail but not overwhelming. The access to metadata needs to be layered to support better communication and education about the data to a wide audience. The message is metadata quality supports information quality as the foundation to enable the use of data.

**Storytelling as a Management Tool:  
Institutionalizing the Data Quality Function  
at Cedars-Sinai Medical Center  
(Practice-Oriented Paper)**

Bruce N. Davidson, Ph.D., M.P.H.  
Director, Resource & Outcomes Management  
Cedars-Sinai Medical Center

**Executive Summary**

Especially in the hospital setting, motivating support for the implementation of information quality (IQ) standards and policies can be a daunting task. Priorities for both leadership and employees focus first on direct patient care, second on meeting budgets, third on patient care improvement, and only at a much more subterranean level on managing IQ. This is true despite the fact that higher-level priorities in many ways depend upon the timely availability of accurate information for administrative and clinical decision-making. While there are a number of current legislative and regulatory initiatives that would impose IQ standards upon healthcare organizations, at the same time, the healthcare industry is experiencing decreasing revenue per unit of output as purchasers and payers try to reduce their costs.

In this type of environment, how can leaders be motivated to prioritize IQ work, and how can employees be motivated to allocate the time and energy required to carry out that work? One way, certainly, is to quantify the impact of poor quality information, for example in terms of lost revenue or lost customers. However, it is generally agreed that this is conceptually difficult, and furthermore, dry lists of numbers do not always succeed in sufficiently mobilizing the commitment needed to squeeze yet another high priority task into a day that is already full to overflowing with high priority tasks. This Practice-Oriented Paper explores storytelling as a management tool that can serve to capture the interest and support needed to successfully establish and implement IQ standards and policies.

**Storytelling as a Management Tool**

Institutionalizing the Data Quality Function  
at Cedars Sinai Medical Center

The 6th International Conference on Information Quality  
November 3, 2001

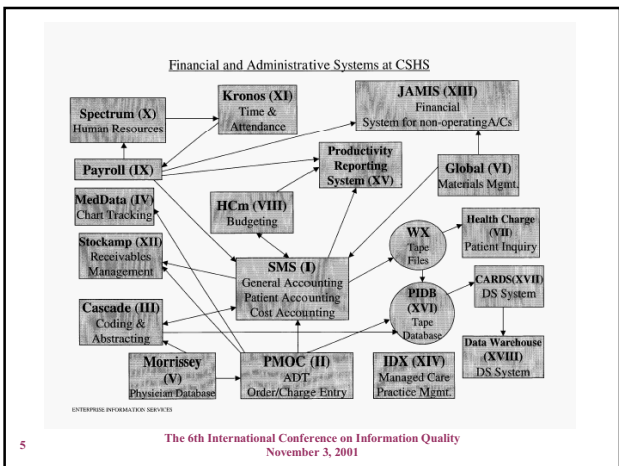
## Large Urban Teaching Hospital

- 875 Beds
- Almost 10,000 personnel
  - 8,000 Employees
  - 1,750 Physicians
- \$2 Billion in gross revenue
- “Illuminated manuscript” as gold standard for information transmission

4 The 6th International Conference on Information Quality  
November 3, 2001

Bruce N. Davidson, Ph.D., M.P.H.  
Director, Resource & Outcomes Management  
Cedars-Sinai Medical Center  
Los Angeles, California

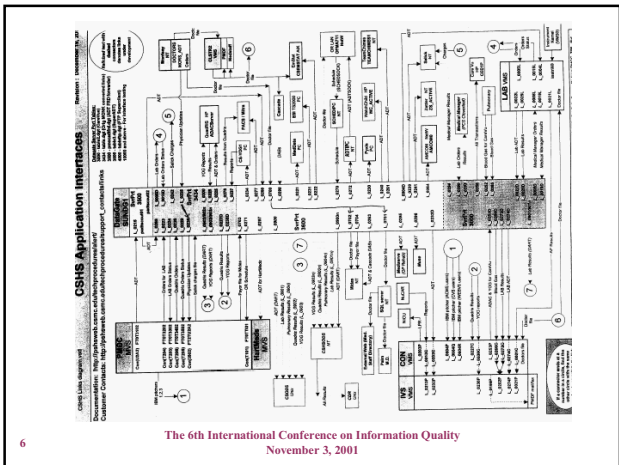
The 6th International Conference on Information Quality  
November 3, 2001



**CEDARS-SINAI MEDICAL CENTER**

- Academic Medical Center/Health System
- Largest Non-Profit Hospital in the Western US
- Basic Annual Statistics
  - 50,000 inpatients
  - 90,000 outpatients
  - 60,000 ER visits
  - 7,000 deliveries

3 The 6th International Conference on Information Quality  
November 3, 2001



## Data Quality Initiative Emerges from Resource & Outcomes Management Dept.

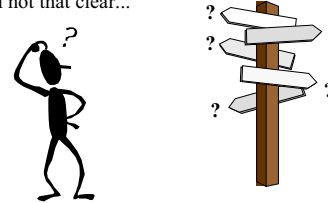
- Internal Consulting Department
- Produces Clinically Oriented Information Products to Support Systematic Patient Care Improvement
- Assures Availability and Reliability of Required Data and Validity of Methods
- Manages Information Flow Through Institution-Wide Quality Management and Medical Staff Database/Reporting Applications

7

The 6th International Conference on Information Quality  
November 3, 2001

## It may look like a lot is going on,

but where we're really going is still not that clear...



10

The 6th International Conference on Information Quality  
November 3, 2001

## Resource & Outcomes Management Dept.'s Vision, Mission, & Objectives

- Vision
  - To be the trusted source for reliable information
- Mission
  - To deliver information products that meet our customers' desired standards of integrity, completeness, accuracy, timeliness, and usability
- Objectives
  - Delivery of specific information products
  - Information management for specific applications

8

The 6th International Conference on Information Quality  
November 3, 2001

## Since, for example...

- DPG Charter was renewed this year by senior leadership, but its meeting frequency dropped from once per month to once every two months, and it has been cancelled once, so it has only met 3 times this year.
- DQMWG Charter was renewed this year by DPG, but key departments, such as IS and Finance, rarely participate.

11

The 6th International Conference on Information Quality  
November 3, 2001

## History of DQ function at CSMC

- |  |   |
|--|---|
| <ul style="list-style-type: none"> <li>■ 1997                             <ul style="list-style-type: none"> <li>- DPG Convened</li> </ul> </li> <li>■ 1998                             <ul style="list-style-type: none"> <li>- TDQM Summer Course</li> <li>- DQMWG Spun Off of DPG</li> <li>- IQ Survey, round 1</li> </ul> </li> <li>■ 1999                             <ul style="list-style-type: none"> <li>- DQ Concept Kick-Off</li> <li>- IQ Survey, round 2</li> <li>- DQ Mgmt Objectives first in FY 99-00 Annual Plan</li> </ul> </li> </ul> | <ul style="list-style-type: none"> <li>■ 2000                             <ul style="list-style-type: none"> <li>- Big DQ Improvement Project</li> <li>- DQ Mgmt Objectives again in FY 00-01 Annual Plan</li> <li>- ROM Dept reorganization to capitalize on DQ framework</li> </ul> </li> <li>■ 2001                             <ul style="list-style-type: none"> <li>- DPG &amp; DQMWG Charters reviewed and renewed</li> <li>- DQ Mgmt Objectives re-emphasized in FY01-02 Annual Plan</li> <li>- IQ Survey, round 3</li> </ul> </li> </ul> |
|--|---|

9

The 6th International Conference on Information Quality  
November 3, 2001

## The Sad Realities of Life

- It's not enough to make a logical argument and advocate for it vigorously.
- Senior leadership has many pressing issues to address and does not yet recognize the relative importance of this function.
- Operational departments aren't yet budgeted to accommodate this function so staffing levels don't reflect required effort.

12

The 6th International Conference on Information Quality  
November 3, 2001

## Leading to One Point of View...

### Institutional Midwife

- Trying to help an organization give birth to a new function
- It's a painful process and there are risks
- Can we play a role that will soothe the mother's pain and ensure the baby's survival?

13

The 6th International Conference on Information Quality  
November 3, 2001

## Institutionalizing through Change



- Efforts to advocate vigorously for a logical argument are not having the desired impact.
- What changes do I need to undertake in my approach that might increase the impact?

16

The 6th International Conference on Information Quality  
November 3, 2001

## And Another Point of View

### Agent of Change

- All improvements are changes, but not all changes are improvements.
- How many psychiatrists does it take to change a light bulb?
- Can we help the organization want to change?

14

The 6th International Conference on Information Quality  
November 3, 2001

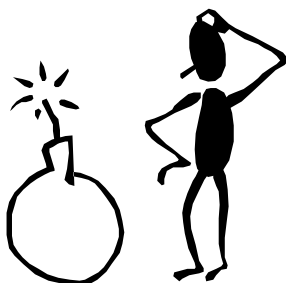
## Improvement in Institutionalization

- Increase the degree to which the data quality function is embraced by leadership and line staff alike.
- Help leadership recognize the relative priority of the data quality function and allocate resources so staffing levels will reflect required effort.

17

The 6th International Conference on Information Quality  
November 3, 2001

## What to do...? What to do...?



15

The 6th International Conference on Information Quality  
November 3, 2001

## Preview of Key Points

- It's hard to catch their attention, and when you do, they focus on little stuff
- Framing and communications strategies dominate the game plan at this stage of institutionalization.
- How to address it? Here's my example.

18

The 6th International Conference on Information Quality  
November 3, 2001

## So how do you catch their attention?

### Numbers?

Yes, you've gotta have numbers. But it's hard to measure the overall impact of poor data quality, so when you do measure, you end up focusing on all the little fires.

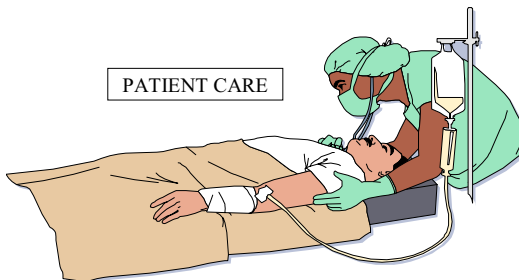
- How much money will we lose because the diagnosis codes in the contract aren't the same as the patients' actual codes?
- How many days will we incur a fine because our required data submission to the state doesn't meet their specs?

19

The 6th International Conference on Information Quality  
November 3, 2001

## #1 PRIORITY

PATIENT CARE



22

The 6th International Conference on Information Quality  
November 3, 2001

## But numbers like that aren't enough...

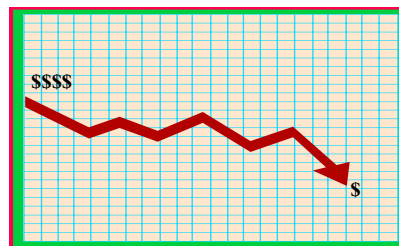
- They ask for numbers so you bring them numbers, and then they focus on putting out the all little fires instead of on why there are so many little fires that need to be put out to begin with.
- Perhaps they think that if we put out the little fires quickly enough we can keep the whole thing from going up in flames...

20

The 6th International Conference on Information Quality  
November 3, 2001

## #2 PRIORITY

CONTROLLING COSTS



23

The 6th International Conference on Information Quality  
November 3, 2001

## And as they put the little fires out ...one by one...

- You'd think attention would automatically move in the direction of investing in the data quality function
- But instead, they quickly turn their attention back to the REAL priorities...

21

The 6th International Conference on Information Quality  
November 3, 2001

## #3 PRIORITY

PATIENT CARE IMPROVEMENT TEAMS



24

The 6th International Conference on Information Quality  
November 3, 2001

#4 priority?

25

The 6th International Conference on Information Quality  
November 3, 2001

...without forgetting  
**THE REAL PRIORITIES!!!**

28

The 6th International Conference on Information Quality  
November 3, 2001

How do you convince them...?

26

The 6th International Conference on Information Quality  
November 3, 2001

...and always keeping in mind that...

“All data are wrong.  
Some data are useful.”

-- W. Edwards Demming

29

The 6th International Conference on Information Quality  
November 3, 2001

...that it's worth the cost?

27

The 6th International Conference on Information Quality  
November 3, 2001

Measurement for Improvement

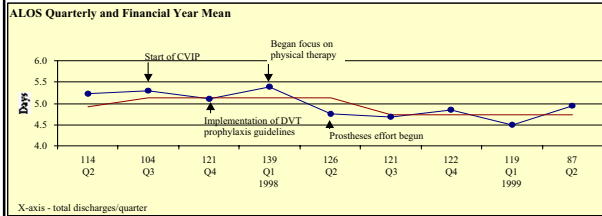
- What are we trying to accomplish? (Aims)
- How do we know that change is an improvement?
- What changes can we make that will result in an improvement?

30

The 6th International Conference on Information Quality  
November 3, 2001



## Documenting Improvement Over Time Total Hip Replacement LOS



31

The 6th International Conference on Information Quality  
November 3, 2001

## Relative Priority

- How to influence relative priority when there are many competing high priority projects
- Taking what's important to you and making it important to them

34

The 6th International Conference on Information Quality  
November 3, 2001

## Storytelling...

- Offers context, relevance, emotion, and a lasting mental image
- May be most powerful way of delivering a compelling and memorable message
- Strategic Issues
  - Framing
  - Communications

32

The 6th International Conference on Information Quality  
November 3, 2001

## Salience and Internalizability

- Choosing an image that already has innate meaning to most audiences
- Getting people to respond to the message through a self-motivated mechanism
- Can they tell the story too?

35

The 6th International Conference on Information Quality  
November 3, 2001

## Framing Strategy

- Relative Priority
  - Opening up the door
- Salience
  - Innate meaning - powerful response
- Internalizability
  - Translating ideas into action

33

The 6th International Conference on Information Quality  
November 3, 2001

## Communications Strategy

- Simplification and Repetition
  - Huh?
- Persuasion
  - OK, why should I do it?
- Role
  - My role as I see it and as they see it

36

The 6th International Conference on Information Quality  
November 3, 2001

## Simplification and Repetition

- It may take 25 repetitions per listener for a message to be effectively delivered due to:
  - Information Overload: We're bombarded with new information every day.
  - Rate of Absorption: New information is absorbed and processed slowly.
  - Searching for Consistency: Consistency of message is used by listeners to judge sincerity.

37

The 6th International Conference on Information Quality  
November 3, 2001

## So What Does This Say About How to Take Action?

- If improvement = change, then
- The change hypothesized to improve the degree to which the data quality function is embraced by leadership and line staff alike is to
- Develop a powerful, simple, easily remembered story and repeat it over and over again

40

The 6th International Conference on Information Quality  
November 3, 2001

## Persuasion

...persuasion becomes a negotiating and learning process through which a persuader leads colleagues to a problem's shared solution...

- Conger, JA. The Necessary Art of Persuasion.  
Harvard Business Review, May-June 1998.

38

The 6th International Conference on Information Quality  
November 3, 2001

## Game Plan - The Test of Change

- Develop story or analogy
- Practice talking about it
- Have it ready to use - you never know when the opportunity might arise.
- Notice when it can be worked into a discussion.
- Use it often - Remember! 25xpp

41

The 6th International Conference on Information Quality  
November 3, 2001

## Role

- From MY Point of View
  - Institutional Midwife
  - Agent of Change
- From THEIR Point of View
  - Reliable Source of Technical Expertise
  - Trusted Advisor

39

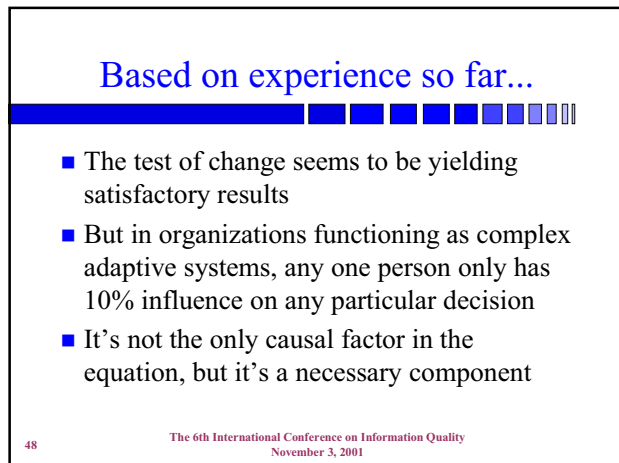
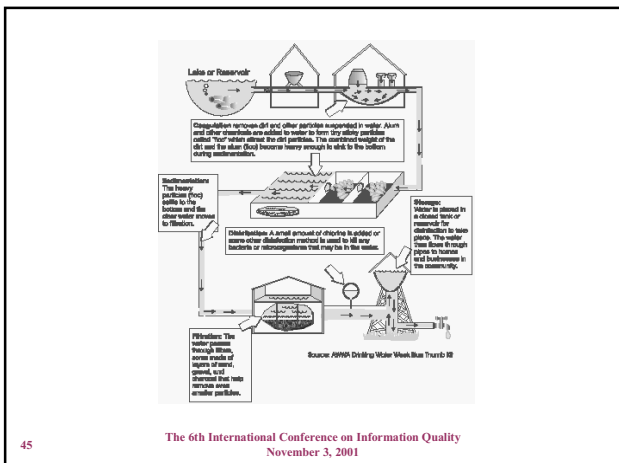
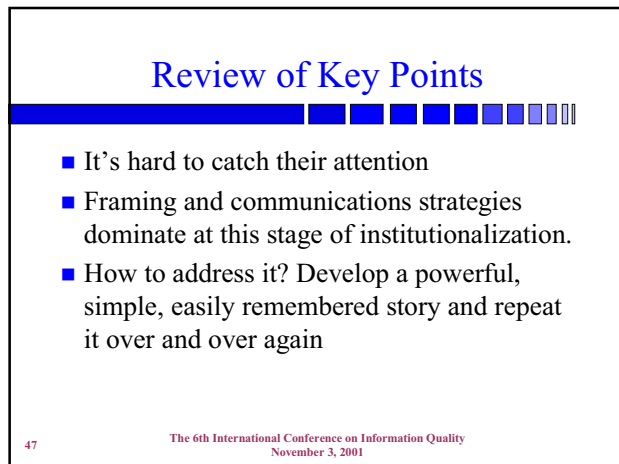
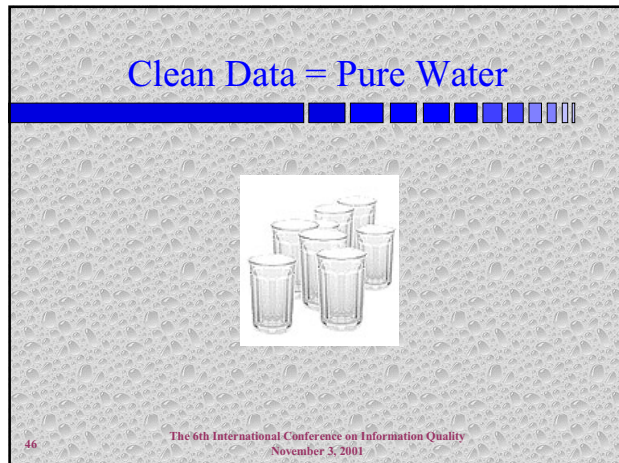
The 6th International Conference on Information Quality  
November 3, 2001

## My Story

- Info Systems = Plumbing
- Data = Water
- Data Quality = Sewage Treatment & Water Purification
- Clean Data = Pure Water

42

The 6th International Conference on Information Quality  
November 3, 2001



## Storytelling as a Management Tool

- Develop story or analogy
- Practice talking about it
- Use it often

49

The 6th International Conference on Information Quality  
November 3, 2001

## Institutionalization of the Data Quality Function may Benefit from Storytelling

- Increases the degree to which the data quality function is embraced by leadership and line staff alike.
- Helps leadership recognize the relative priority of the data quality function and allocate resources so staffing levels will reflect required effort.

50

The 6th International Conference on Information Quality  
November 3, 2001

## Clean Data = Pure Water



The 6th International Conference on Information Quality  
November 3, 2001

## **Generations of information quality A practice paper**

Frank Dravis  
Firstlogic, Inc.

---

### **Executive summary**

Too often beleaguered business and IT and managers struggle with communicating to executives that the organization is suffering from data quality problems. What the managers don't realize is organizational *immaturity* is the root cause of the communication struggle. The executives, the organization as a whole, are unprepared and incapable of hearing the data quality message as presented. A gulf, a perception gap exists between the levels of management and the functions in the organization that must be bridged at the awareness level of senior management, and that of the functional managers – business and IT. The purpose of this presentation is to educate the IT and business managers to the concept of organizational maturity, specifically in regards to information quality. Once the managers understand there are roughly five levels of information quality maturity, and that the messaging, actions, and behaviors change with each level, the managers will be ready to “tune” their communication for the proper reception at the level of their audience.

In support of the above argument we explore the industry maturity cycle, technology diffusion curve, a sampling of existing information quality maturity models, and present a maturity assessment case study. Additionally, we draw parallels to TQM concepts, and touch on the components to an information quality initiative. The ultimate goal of the presentation is to educate business and IT managers as to the cultural issues surrounding information quality, and thus equip them to cope and then change their organization's attitude towards information quality.

INNOVATIONS 2001  
Research. Intelligence. For. Success.

## Generations of Information Quality

© Firstlogic, Inc 2001  
Rev 1.3

INNOVATIONS 2001

## Your Speaker...

Frank Dravis, VP Information Quality  
*Research and Practice*

- 16 years in IT and S/W Development, 13 years specializing in information quality solution design and implementation.
- Responsible for identifying and pursuing strategic information quality opportunities for Firstlogic, in terms of external markets, and internal practices.
- Started in DQ by writing address parsing, assignment, and standardization routines

© Firstlogic, Inc 2001

INNOVATIONS 2001

## What You Will Learn

- Concept of industry maturity levels and adopters
- Information quality (IQ) maturity models
  - Sampling of various models
- IQ maturity level indicators
- Methods for assessing IQ maturity levels
- What to do if you find yourself in a level 1
- Continued evolution of IQ

© Firstlogic, Inc 2001

INNOVATIONS 2001

## Value of Information Quality

- Decreased Operational Costs, Decreased Rework: Greater Efficiency
- Faster Decision Making
- More Accurate Decisions
- Increased Employee Satisfaction
- Increased Customer Satisfaction
- Increased Shareholder Satisfaction
- Greater Effectiveness

**Equate to increased productivity, revenue, and profits**

© Firstlogic, Inc 2001

INNOVATIONS 2001

## British Telecom Understands IQ

- "Physical assets are increasingly becoming less important in determining the success and valuation of companies. Instead intellectual capital, including the value of information and knowledge assets is becoming the critical determinant of perceived worth of future profitability."

"The Drive to High IQ in British Telecommunications. Deploying Information Quality Tools in a Federated Business." MIT IQ 2000 conference

© Firstlogic, Inc 2001

INNOVATIONS 2001

## Famous Failures

- NASA Challenger: o-ring seals out of tolerance
- Ford Pinto: poorly design gas tanks
- Exxon-Valdez: single-hull tanker grounding
- Three Mile Island: inadequate emergency response training
- Piper Alpha oil rig: lack of blast wall protection

© Firstlogic, Inc 2001

INNOVATIONS 2001 firstlogic

## Organizations That Did Not Listen

- All had people in their organizations that warned
- Why did they fail?
  - Because their organization, management and culture was not **ready** (willing) to hear and act what they said
- The organization was immature in some regard
  - Whether it was a safety, environmental, quality, or cost issue
- They had not learned when one of their own tried to teach

*Have you ever felt like you were the one being ignored?*

© Firstlogic, Inc 2001

INNOVATIONS 2001 firstlogic

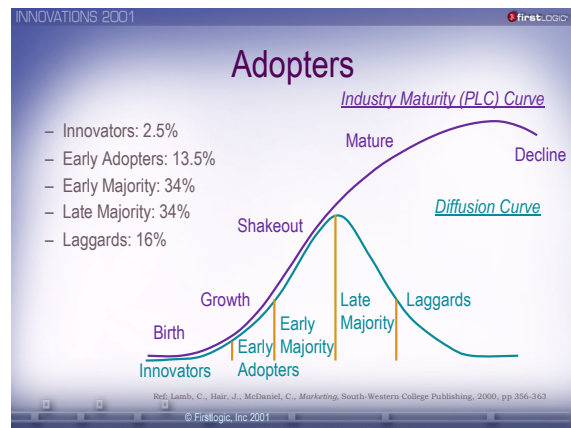
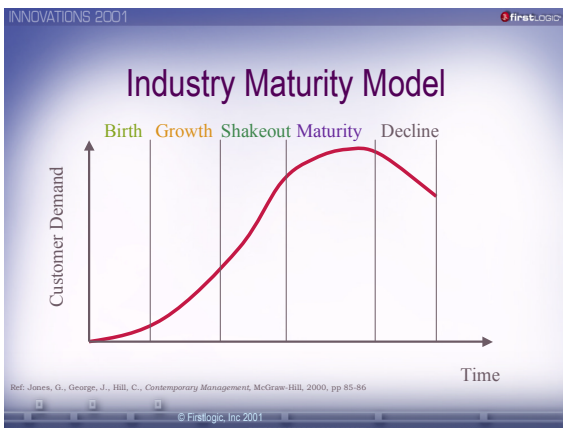
## The Benefits of Knowing

IQ maturity progression

You will...

- Know why senior management has not been listening
- Be aware some organizations are farther ahead
- Understand the behavior of an organization
- Have a framework to change your organization
- Know the actions to pursue as your IQ evolves

© Firstlogic, Inc 2001



INNOVATIONS 2001 firstlogic

## IQ Maturity Models

- CIO Magazine
  - Stage 1, Denial
  - Stage 2, Acceptance
  - Stage 3, Leverage
  - Stage 4, Webification
- Tom Redman, Navesink Consulting
  - **First Generation:** Inspection and rework to find and fix defects
  - **Second Generation:** Process Management to prevent defects
  - **Third Generation:** Design renders defects "impossible"

Ishikawa notes that Quality Systems Evolve

Ref: Stackpole, B., Wash Mc, CIO Magazine, February 15, 2001, pp 101-112

© Firstlogic, Inc 2001

INNOVATIONS 2001 firstlogic

## IQ Maturity Model

Philip Crosby, Larry English

- Level 1: Uncertainty
- Level 2: Awakening
- Level 3: Enlightenment
- Level 4: Wisdom
- Level 5: Certainty

A software development version:

- Carnegie and Mellon's SEI CMM

Crosby, P., Quality is Still Free, McGraw-Hill, 1996, pp 31-55; English, L., Improving Data Warehouse and Business Information Quality, Wiley, 1999, pp 427-450

© Firstlogic, Inc 2001

INNOVATIONS 2001 firstlogic

## A New Maturity Scale


Level 1: Asleep. House is on fire.

Level 2: Awake. Smell smoke.

Level 3: Panic. Put the fire out!

Level 4: Fire's out. Don't want another.

Level 5: Fire resistant. Won't have another.



5

Adapted from Philip Crosby and Larry English  
© Firstlogic, Inc 2001

INNOVATIONS 2001 firstlogic

## Level 1: Asleep

Feeling no pain

- Management Perceptions
  - No awareness of IQ value
  - They have NO data problems
  - Believe information is domain of IT
- Infrastructure
  - No quality org, except in IT dev.
  - No IQ metrics taken or published

- Management Behaviors
  - Faults IT when problems are exposed
  - IT faults business for app failures
  - Finger pointing
- Actions
  - Cover ups, criticisms, and back-biting
  - Information workers frustrated to point of apathy
  - 20% of revenue spent on scrap and rework

Adapted from Philip Crosby and Larry English  
© Firstlogic, Inc 2001

INNOVATIONS 2001 firstlogic

## Level 2: Awake

Felt some pain, might go away on its own

- Management Perceptions
  - Aware of an IQ problem; caused by a catastrophic failure
  - Unsure of size of problem, or persistence
  - Poor IQ has cost them something
- Infrastructure
  - adhoc team established to clean up
  - An IQ manager may be appointed

- Management Behaviors
  - Want to fix problem
  - Reluctant to spend on problem
  - IT assigned to fix problem, but within current budget
- Actions
  - No change in processes or mgt. systems
  - Clean up of specific problem
  - Cost of clean up effort is tracked
  - Some scrap/rework eliminated
  - 18-16% of rev spent on scrap and rework

Adapted from Philip Crosby and Larry English  
© Firstlogic, Inc 2001

INNOVATIONS 2001 firstlogic

## Level 3: Panic

Oh boy, we're feeling some pain now

- Management Perceptions
  - Information problems will not go away
  - Must re-evaluate corporate position on value of IQ
  - Everyone accountable for IQ
- Infrastructure
  - Formal, cross-functional group(s) established for IQ
  - Adhoc business group focused on data standards

- Management Behaviors
  - Actively learning about IQ
  - IQ initiatives are sponsored
  - Funding established for IQ initiatives
  - Directs processes to be permanent
- Actions
  - Business and IT are coordinating on information issues
  - Data quality assessments conducted
  - Root cause of problems sought
  - Long-term solutions implemented
  - 15% of rev spent on scrap and rework

Adapted from Philip Crosby and Larry English  
© Firstlogic, Inc 2001

INNOVATIONS 2001 firstlogic

## Level 4: Fire's Out

Felt enough pain, and are tired of it

- Management Perceptions
  - Significant benefits come from IQ
  - Substantial impact of IQ on bottom line
  - Sr. Mgt. is accountable for IQ
  - IQ is tied to customer satisfaction
- Infrastructure
  - Everyone in the organization is involved in a formal or informal IQ activity
  - CIO is accountable for technical enablement of IQ

- Management Behaviors
  - Ensures continued implementation and maturation of IQ processes
  - Consumers of information are considered customers
  - Cultural obstacles to IQ are addressed
  - IQ metrics are added to KPIs
- Actions
  - Business/IT partnerships are defacto
  - App., data, and business processes are designed with IQ as a requirement
  - Defect prevention is a norm
  - 10% of rev spent on scrap and rework

Adapted from Philip Crosby and Larry English  
© Firstlogic, Inc 2001

INNOVATIONS 2001 firstlogic

## Level 5: Fire Resistant

Feeling little pain, and want to keep it that way

- Management Perceptions
  - Folly to conduct business without IQ management in place
- Infrastructure
  - IQ management mentors and trains Bus/IT teams
  - Assures new systems are design with quality in mind

- Management Behaviors
  - New IQ problems fixed immediately
  - Employees incentivized to look for issues
  - Compensation elements tied to IQ
- Actions
  - Measures lifetime value of customers
  - Most IQ failures caused by external events
  - Audits performed on process/sys design
  - 5% of rev spent on scrap and rework

Adapted from Philip Crosby and Larry English  
© Firstlogic, Inc 2001



INNOVATIONS 2001 firstlogic

## Maturity Assessment

- Don't need metrics or measurements
- Do need cross-functional input and perceptions
  - Acct., Mktg., Admin., Sales, Mnfg, Shipping, R&D, *and* IT
- Survey, who's purpose is to determine:
  - Perceived importance of data quality to organization
  - Data steward's perceptions of current data quality
  - Data consumer's perceptions of current data quality
  - Policies and responsibilities to cleanse operational data
- Don't use personal attribution in the findings


© Firstlogic, Inc 2001

INNOVATIONS 2001 firstlogic

## Awakening A Level 1

Level 1s need the most help, and they predominate

- The survey starts the process
- Document the information issues
- Pick the top issue and assess impact
- Educate management that they are feeling pain
- Be ready with a proposed solution
- Appoint an IQ smoke detector



© Firstlogic, Inc 2001

INNOVATIONS 2001 firstlogic

## IQ Maturity Case Study: FGV

- Personal interviews of 8 people
- 2 Senior managers
- 1 IT manager, 1 IT analyst
- 3 Business managers
- 1 Customer support manager
- Wanted a strong cross-section to smooth anomalies, agendas
- Asked 80 questions
- Questions mapped against 59 individual maturity indicators
- Questions constructed to show continuity and affirmation

© Firstlogic, Inc 2001

INNOVATIONS 2001 firstlogic

## FGV Assessment

FGV assessed at Level 3 with caveats

- No formal, **cross-functional** IQ groups established
- Root cause of problems not always sought
- Long-term solutions not always implemented
- Cost of clean-up efforts not tracked

Average estimated % of time spent on rework: 19.7%

- Percent of time working with information

Weak indicators leading to the next level were positive signs of maturity growth, rather than maturity regression

© Firstlogic, Inc 2001

INNOVATIONS 2001 firstlogic

## Assessment Insights

- In some cases existing cynicisms prompted overly critical judgments.
- Some perceptions were completely wrong according to facts.
- Perception gaps existed between mgt levels, and business and IT.
- The gap hindered effective and joint planning of IQ initiatives.
- Lower mgt. was often unaware of senior mgt. Intentions.
- Three types of personnel found: business, IT, and boundary.
- Personnel involved in long-standing IQ activities no longer see them as such (SOP).
- Assessments identify important, future cross-functional IQ initiatives.

© Firstlogic, Inc 2001

INNOVATIONS 2001 firstlogic

## Matrix Perception Gap

Attitude Tendencies

	Business	Boundary	IT
Senior Mgt	Learning	Realistic	Aware
Line Mgt	Critical	Realistic	Defensive
Workers	Negative	Realistic	Positive

© Firstlogic, Inc 2001

INNOVATIONS 2001 firstlogic

## Actions

to Speed Maturity Progression

- Establish forums for business and IT management to share perceptions and engage in dialog. **Close the perception gap.**
- Establish regular, multi-channel, wide-spread communication of IQ activities. **Eliminate perception inaccuracies.**
- Senior mgt. vigorously participate in communication. **Remove any doubt as to position and intentions.**
- Identify and increase number of liaison personnel. **Promotes communication, flow of information, and common perceptions**

© Firstlogic, Inc 2001

INNOVATIONS 2001 firstlogic

## IQ Initiative Fundamentals

- Be aware that IQ is a cultural issue
- Start with a **pilot**
- Pick an information issue where pain is apparent
- Research the problem, and then the solution
- Find your sponsor
- Understand the perception gap
- Don't assume you need a hard ROI

© Firstlogic, Inc 2001

INNOVATIONS 2001 firstlogic

## IQ Initiative Specifics

- Determine importance of data quality to organization (*accomplished via maturity survey*)
- Determine data owners' perceptions of current data quality (*maturity survey*)
- Determine down-stream users' perceptions of current data quality (*maturity survey*)
- Determine policies and responsibilities to cleanse operational data (*maturity survey*)
- Establish a cross-functional IQ team to resolve disputes

Moss, L., Method Focus Inc., Dirty Data presentation to Firstlogic, Inc., 2001  
© Larissa Moss, Method Focus Inc.

INNOVATIONS 2001 firstlogic

## IQ Initiative Specifics (cont)

- Create a process to include down-stream data users in operational system requirements and analysis sessions
- Create a policy for logical data modeling
- Create a policy for meta data capture (business & technical)
- Create a policy for a central DW staging area
- Assemble and train a team to regularly assess the quality and the consistency of operational and DW data

Moss, L., Method Focus Inc., Dirty Data presentation to Firstlogic, Inc., 2001  
© Larissa Moss, Method Focus Inc.

INNOVATIONS 2001 firstlogic

## IQ Initiative Specifics (cont)

- Establish procedures for prioritizing which data to cleanse first (and where)
- Establish procedures for rejecting or suspending dirty data
- Review and revise existing data standards
- Incorporate new standards into the development methodology
- Change incentive policy to include accountability for data quality
- Manage data like any other resource in the company

Moss, L., Method Focus Inc., Dirty Data presentation to Firstlogic, Inc., 2001  
© Larissa Moss, Method Focus Inc.

INNOVATIONS 2001 firstlogic

## IQ Parallels to TQM

1. Build organizational commitment to quality:  
**Includes your information**
2. Focus on the customer:  
**Information consumer**
3. Find ways to measure quality:  
**In your information**
4. Set goals and create incentives:  
**For the management of information**
5. Solicit input from employees:  
**Uncover their issues and ideas for information**

Ref: Jones, G., George, J., Hill, C., Contemporary Management, McGraw-Hill, 2000, pp 655-659  
© Firstlogic, Inc 2001

INNOVATIONS 2001 firstlogic

## IQ Parallels to TQM

6. Identify defects and trace them to their source:  
**Where did the defective data come from?**
7. Introduce just-in-time inventory:  
**Information when people need it**
8. Work closely with suppliers:  
**Those who produce your information**
9. Design for ease of manufacture:  
**Ensure information accuracy first**
10. Break down barriers between functions:  
**Information transcends functions**

Ref: Jones, G., George, J., Hill, C., Contemporary Management, McGraw-Hill, 2000, pp 655-659  
© Firstlogic, Inc 2001

INNOVATIONS 2001 firstlogic

## Favorable IQ Factors

- When the Company Has a Commitment to Quality
  - Malcolm Baldrige, TQM, CMM, ISO
- When Data is a Success Factor to an Important Project
  - Data creation, usage, integration, and reporting
    - Business intelligence (BI), analysis, decision support
    - Customer Relationship Management (CRM), Enterprise Application Integration (EAI)
    - B2B, E-commerce
    - Merger/Acquisition
    - Project planning, application development, configuration and change management
- Quality Control is a Mandate
  - Government Inspection (FDA, SEC, DOD, USPS: *CASS, PAVE, SERP*)

Ref: Fortino, R., DMR Consulting  
© Firstlogic, Inc 2001

INNOVATIONS 2001 firstlogic

## Demands of Your Customers

- Your customers, patients, or passengers will drive you harder for information than you will drive yourself. Use those demands to propel your organization's IQ initiative.





© Firstlogic, Inc 2001

INNOVATIONS 2001 firstlogic

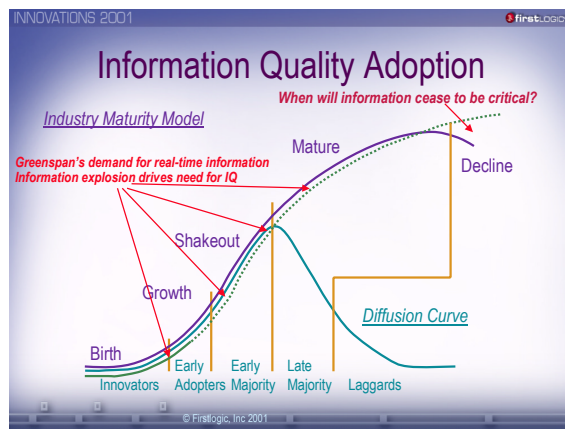
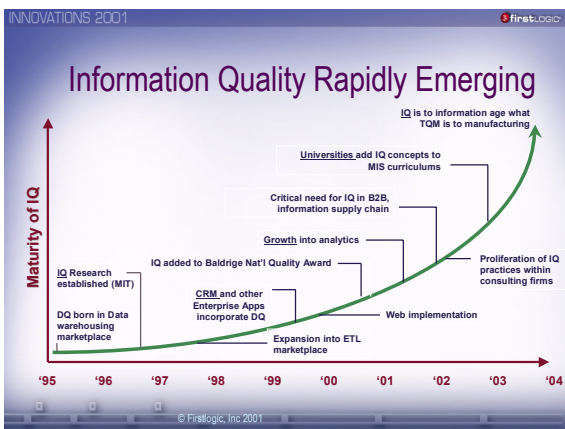
## Establish Information Goals

*What is your key business driver?*

Examples:

- Achieve a single view of the customer
  - More effective communication, better relations, better management of accounts
- Calculate the life-time value of any customer
  - Who are the top 20% of your customers generating 80% of your revenue?
- Accurately segment revenue/earnings per your vertical markets
- Information reporting done at a click of a button
  - Get operational reports within minutes of when you need them
  - Eliminate subordinate "scurry" as they scramble to acquire data and build reports
- Reduce operational costs
  - Consolidate redundant data and data stores. Reduce rework via better production processes.

© Firstlogic, Inc 2001



INNOVATIONS 2001 firstlogic

## The Benefits

of knowing IQ maturity evolution

You...

- Understand why senior management has not been listening
- Are aware organizations function at different levels
- Understand the behavior of your organization
- Have a framework to change your organization
- *Know the actions to pursue as your IQ evolves*

"Seek first to understand, then to be understood."  
--Stephen Covey

1


© Firstlogic, Inc 2001

INNOVATIONS 2001 firstlogic

## In The End

Your information can be either your competitive advantage, or disadvantage.

*It will be one or the other.*



1

© Firstlogic, Inc 2001

INNOVATIONS 2001 firstlogic

## References

- [1] Stackpole, B., *Wash Me*, CIO Magazine, February 15, 2001, pp 101-112
- [2] Lamb, C., Hair, J., McDaniel, C., *Marketing*, South-Western College Publishing, 2000, pp 356-363
- [3] Jones, G., George, J., Hill, C., *Contemporary Management*, McGraw-Hill, 2000, pp 85-86, 655-659
- [4] English, L., *Improving Data Warehouse and Business Information Quality*, Wiley, 1999, pp 427-450
- [5] Huang K., Lee Y., Wang R.: *Quality Information and Knowledge*; Prentice Hall: 1999, pp 60-62, 85-90
- [6] Forino, R., DMR Consulting
- [7] Moss, L., Method Focus Inc., *Dirty Data* presentation to Firstlogic, Inc., 2001
- [8] Redman, Tom, *Definitions of Data Quality*, <http://www.navesinconsulting.com/Definitions%20of%20Data%20Quality.html>
- [9] Crosby, P., *Quality is Still Free*, McGraw-Hill, 1996, pp 31-55
- [10] Dowie, M., *Pinto Madness*, Mother Jones, Sept Oct 1997, [http://www.motherjones.com/mother\\_jones/SO77/dowie.html](http://www.motherjones.com/mother_jones/SO77/dowie.html)
- [11] Vaughan, D., *The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA*, University of Chicago Press, Chicago, 1996
- [12] various, *Famous Engineering Disasters*, University of Guelph, School of Engineering, Ontario, <http://www.oes.uoguelph.ca/webfiles/james/homepage/Teaching/FamousEngnDisasters.htm>
- [13] TeraQuest Inc, CMM Assessment Final Findings presentation to Firstlogic, Inc., March 1999

1

© Firstlogic, Inc 2001

INNOVATIONS 2001 firstlogic

## Questions



1

© Firstlogic, Inc 2001

INNOVATIONS 2001 firstlogic

## Capability Maturity Model

1

© Firstlogic, Inc 2001

## **Organizational Realism Meets Information Quality Idealism: The Challenges of Keeping an Information Quality Initiative Going**

### **Beverly K. Kahn**

Suffolk University  
Sawyer School of Management  
8 Ashburton Place  
Boston, MA 02108  
Tel: (617) 573-8642  
Fax: (617) 573-8345  
[bkahn@acad.suffolk.edu](mailto:bkahn@acad.suffolk.edu)

### **Raïssa Katz-Haas**

Ingenix  
450 Columbus Blvd  
Hartford, CT 06115  
Tel: (860) 702-8721  
[rkatzh@uhc.com](mailto:rkatzh@uhc.com)

### **Diane M. Strong**

Worcester Polytechnic Institute  
Management Department  
100 Institute Road  
Worcester, MA 01609  
Tel: (508) 831-5573  
Fax: (508) 831-5720  
[dstrong@wpi.edu](mailto:dstrong@wpi.edu)

### **Abstract**

While many seem to understand, in theory, the value of high-quality information, the realities of daily organizational dynamics may cause information quality projects to falter. In our conference paper last year, we described how Ingenix moved from ‘theory’ to an established Information Quality initiative, and how it laid the groundwork for its first and future information quality improvement projects. This paper describes the challenges and successes of keeping information quality initiatives going in the presence of development deadlines, the call of daily work activities, organizational dynamics, etc. The lessons learned from Ingenix provide insights for other organizations as they seek to keep information quality initiatives strong and healthy. One of the lessons/insights is that multiple, even dissimilar IQ (*See acronyms, p. 12*) efforts, carried out simultaneously can be as effective as a single overarching initiative.

*This paper is dedicated to the memory of Tim B. Kaufman  
Galaxy Architect. Mentor. Friend.*

## **INTRODUCTION**

In our IQ 2000 conference paper, "How to Get an Information Quality Program Started: The Ingenix Approach", we described one company's approach to starting an information quality improvement program [Kahn, Katz-Haas, and Strong, 2000]. While the essential first step of an information quality (IQ) improvement endeavor—getting the project started—is more difficult than one might expect, keeping it going is just as challenging. In this paper, we look at a year in the life of what began as a corporate-sponsored information quality initiative.

What we see are actually several information quality efforts operating simultaneously and interacting with each other. Some of the efforts operate at a local level, others under corporate sponsorship, some as part of information systems development projects, and some as more independent information quality efforts. As is typical of the dynamics of any organization, these efforts start and then may flounder, and then re-gain momentum in a different form—or several forms.

## **OVERVIEW OF INGENIX**

Ingenix is a wholly owned subsidiary of UnitedHealth Group (UHG), a multi-billion dollar diversified health and wellness company headquartered outside Minneapolis. Ingenix is one of the six business segments of UHG. (For more information on UnitedHealth Group, see <http://www.unitedhealthgroup.com>.) Ingenix has become one of the largest health care information and research companies. At the UHG web site, Ingenix is described as follows:

“ Improved knowledge and information are key to achieving improved health and well-being. Access to accurate, unbiased research and information helps improve the effectiveness of care by supporting fact-based clinical and financial decisions. Ingenix is uniquely positioned to fulfill this need.”  
(<http://www.unitedhealthgroup.com/ingenix/index.html>)

Ingenix has become one of the largest health care information and research companies in the industry, providing a comprehensive line of products and services—many of which are rooted in proprietary databases. Therefore high quality data is essential to Ingenix's continued growth, and Ingenix is well aware of this. To quote from the CEO of Harvard Pilgrim Health Plan,

“ If you are in the health [care] business, your data better be damn good because you're using that data to make decisions that are going to impact your company six, 12, and 18 months from now. If you don't have full faith in the comprehensiveness and precision of the data you're using, you're taking some huge and unacceptable risks.”

[McCue, 2001]

## THE SHARED DATA WAREHOUSE

The Shared Data Warehouse (SDW) is a department within Ingenix's Information Systems (IS) unit. The SDW is made up of six interdependent groups: 1) Information Quality & Strategies (IQS), 2) Business Analysis, 3) Software Engineering, 4) Training and Support, 5) Project Management, and 6) Applications Development.

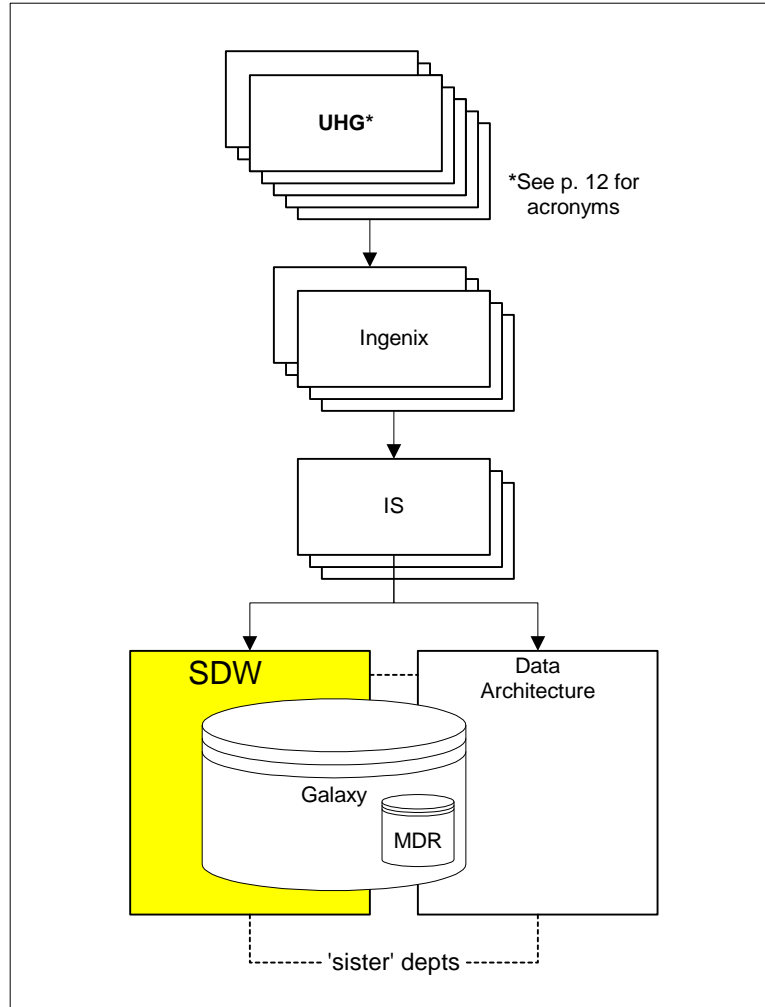


Figure 1 SDW Relationships

The SDW works closely with another IS department (which, up until a few months ago was part of the SDW), the Data Architecture Department. This department includes six groups: 1) Metadata Repository (MDR) group, 2) Data Administration group, 3) Modeling group, 4) Database Administration group, 5) the ETL (extraction, transformation, and load) group, and 6) the Technical Infrastructure group.

In the summer of 1998, UnitedHealth Group charged the SDW with developing a new data warehouse (now known as Galaxy) to integrate data from UnitedHealth Group's two major data stores (MARS and DSS) into one repository. This would make reporting across the enterprise easier for customers/users. As part of the Galaxy charter and to support the needs of its direct and indirect users, the Shared Data Warehouse committed to deliver data of equal or better quality as MARS/DSS data. In addition, Galaxy was purposefully architected to minimize data quality problems. Thus from the beginning of the Galaxy project, Ingenix recognized that information quality issues are an integral part of data warehouse development.

## **QUALITY DATA SOURCES FOR THE DATA WAREHOUSE**

The Galaxy data warehouse is a significant project, which was completed on schedule on August 1, 2001. (Completion means that all the planned subject areas are in the data warehouse and accessible to its users). Each Subject Area team decided what data to include and from which sources. Data quality issues were among the considerations in selecting data and their sources. This enabled subject area teams to make best choices about what sources to use for the warehouse.

## **GALAXY INFORMATION AND METADATA QUALITY STRATEGY**

SDW Management and IQS developed Galaxy's data quality strategy in December of 1999. The strategy included eventual corporate sponsorship, the not yet formed IQPI initiative/group, continuous process improvement (using a Six Sigma methodology [Pande, Neuman, and Cavanaugh, 2000]) and the ultimate goal: prevention of data and metadata quality problems.

Very little cleansing or editing of source data is done for the data warehouse. This is by design. IQS and SDW management decided that the best way to improve the quality of data over the long run was to use most data 'as is' rather than attempting to hide or cover up every data quality problem by cleansing and filtering/editing. This way, root causes of data problems can be located, analyzed, and remedied, preventing future problems. Additionally, an overabundance of cleansing/editing routines degrades warehouse performance, a highly significant data quality issue to customers.

Users need high quality metadata as well as quality data. The metadata repository (MDR) group and the modelers within the Data Architecture department built Galaxy's metadata repository based on business requirements gathered by IQS. IQS and the MDR group work closely to ensure high quality metadata for Galaxy's customers and SDW staff.



The MDR group maintains the metadata and ensures that metadata is accessible by the SDW staff for various purposes. One of these purposes is to allow extraction of appropriate metadata for delivery to Galaxy's customers. The 'web team' (an extension of IQS) publishes the extracted metadata to the web for easy access by users. Below is an example of metadata as the user sees it. This is from a page of the Galaxy Data Dictionary:

The screenshot shows a web interface for a data dictionary. At the top, it displays 'Table Name: Diagnosis Code' and 'DB2 Name: DIAGNOSIS\_CODE'. Below this are navigation links for 'Table Descriptions', 'Columns', 'Indexes', and 'Joins', along with a search icon and a 'Dictionary Home' link. A 'Skip to' menu lists letters A, R, G, L, M, U. The main content area is titled 'A' and lists the following details for the 'Diagnosis Code' table:

Business Name: AHRQ Diagnosis Detail Category	Data Type: CHARACTER
Code	Length: 3
DB2 Name: AHRQ_DIAG_DTL_CATGY_CD	
Identifies the AHRQ (Agency for Healthcare Research and Quality) detailed grouping of diagnosis codes that describe a disease (e.g. tuberculosis).	

Figure 2. Meta Data Example

Metadata quality improvement is a core part of building and operating the data warehouse at Ingenix. Its business is to supply healthcare data and information to other segments within UHG and to external customers. For Ingenix to provide a high quality product to its customers, it must ensure both high quality metadata and data.

## CORPORATE HEADQUARTERS TAKES ON INFORMATION QUALITY

To dramatically improve information quality in a company, improvement initiatives must go beyond its IS group(s) because IS groups do not 'own' business processes that create, capture, and gather data for data entry, processing, or other manipulation of data. There are many reasons for particular quality problems. For instance, many information quality problems in organizations are introduced at the original source (i.e., first business event/transaction). Others are introduced as a side effect of policy decisions about the business. Thus, corporate-wide information quality initiatives, sponsored or lead by Senior Management, are critical to the success of IQ initiatives.

In our IQ 2000 conference paper, we described the beginnings of the Information Quality Process Improvement (IQPI) initiative, a UHG-wide information quality initiative [Kahn, Katz-Haas, and Strong, 2000]. Planning and discussion of information quality began in earnest mid-1999. Ingenix sponsored an enterprise-wide IQ seminar that took place about a year later in June 2000. From this seminar, an initiative and program group consisting of fifteen members from all major business segments of UHG were formed in July. The name of the group and initiative was IQPI (Information Quality Process Improvement).

At the time of last year's IQ conference in October, this corporate-wide information quality initiative was off to a good start. It had support from the President of UHG, involvement from the appropriate groups, and a list of information quality issues that were the focus of the group's work. Unfortunately, the momentum was not maintained, and the last time the IQPI group met was in December 2000. *However*, consciousness was raised across the enterprise of

the criticality of good quality data. Other IQ efforts emerged as a result of this and of the IQ seminar and IQPI. In addition, the UnitedHealth Group president created a position, which is wholly dedicated to data quality across the company with a focus on improving the quality of the operational databases and their data. The position is high enough in the organization to carry the clout needed to work across segments/divisions. The President filled this position in the summer of 2001.

## **CORPORATE QUALITY IMPROVEMENT INITIATIVE**

Planning for a corporate-wide focus on a ‘TQM-type’ quality and process improvement initiative began at UHG corporate headquarters in January 2000. The plan is to phase in this quality program, which we will call the Quality Excellence (QUE) program (a fictitious name), to all of UHG within three years. The QUE program has several general purposes: 1) support UHG’s business strategy, 2) improve results, 3) exceed customer expectations, and 4) achieve performance excellence.

Many business segments have begun at least one QUE project. Ingenix selected Galaxy’s IQ as one of its QUE projects and provided two sponsors for assistance, if needed. (These sponsors also report project progress to corporate sponsors.)

The QUE initiative is similar to TQM, CPI, and other quality/improvement programs. QUE essentially follows a Six Sigma methodology, including SPC (Statistical Process Control) and the DMAIC model. DMAIC is an essential part of Six Sigma and stands for **Define, Measure, Analyze, Improve, and Control**. Defining what “quality” means, according to the customers, was part of the 2<sup>nd</sup> quarterly QUE report. For its project, IQS provided the following definition:

“The quality of data = its fitness for use. In addition, the quality of the data meets or exceeds customer requirements.”

## **MEMBERSHIP DATA QUALITY IMPROVEMENT INITIATIVE**

IQS had already been using the DMAIC model in its focus on information quality. Its first project involved looking at, and improving membership data. For many businesses, membership data is often difficult to improve and is an enormous task. Clearly this was not a one-person job. IQS formed a small, focused QUE team consisting of:

- representatives from four business areas (customers)
- two source system representatives
- four staff members from the SDW

Working in small teams is part of the DMAIC methodology. What is a little unusual about this team is that it is cross-functional and must work through a variety of agendas, locations, and schedules. However the advantages of a cross-functional team far outweigh any disadvantages.

Membership data refers to data about the people enrolled in a healthcare plan and who are eligible for certain benefits. Membership data passes through many business processes, paper forms, eligibility systems, claims processing systems, and the all business processes involved in installing, canceling, and re-installing policies and their attendant membership. Improving the reliability of membership data is a huge undertaking due to the number of members, the number of systems (mostly 'stovepipe'), the number and complexity of business processes involved (also often stovepipe), and the interrelationships between all of the above.

To understand why an accurate member count is difficult, consider the following:

- When policies are cancelled, members' claims, which were incurred before policy termination typically remain eligible for a period of time called 'runout'
- When a member dies, coverage may continue for the spouse. A business decision was made that, out of consideration for the spouse, coverage would continue under the social security number of the deceased. As a result, there remains a record in the data warehouse for the deceased, as well as a record for the surviving spouse.
- Contract cancellation dates that are not in proper relationship to service dates cause innumerable problems. These problems (i.e., symptoms) are fairly easy to spot, but finding their root causes is complex and difficult.
- Most significantly, membership data are housed redundantly in numerous systems across the enterprise (See Figure 3.)

These are the types of issues that can skew member counts, causing departments to develop expensive workarounds among other consequences.

The QUE team started with a *relatively* small project: finding out why there are records for 'active' members whose policies have been cancelled and how to remedy this situation.

Thus far, the QUE team has conducted over 40 interviews, exchanged 130+ emails, had numerous 'phone meetings' across the company and have begun data and document collection. The purpose of all this data collection is to understand one membership process: the Case Cancellation process.

The team has mapped data flows and business and systems process related to the Case Cancellation process. These data flows are shown at a high level in Figure 3. The complexity of the data and systems relationships is apparent.

### Case Cancellation: high-level data flow

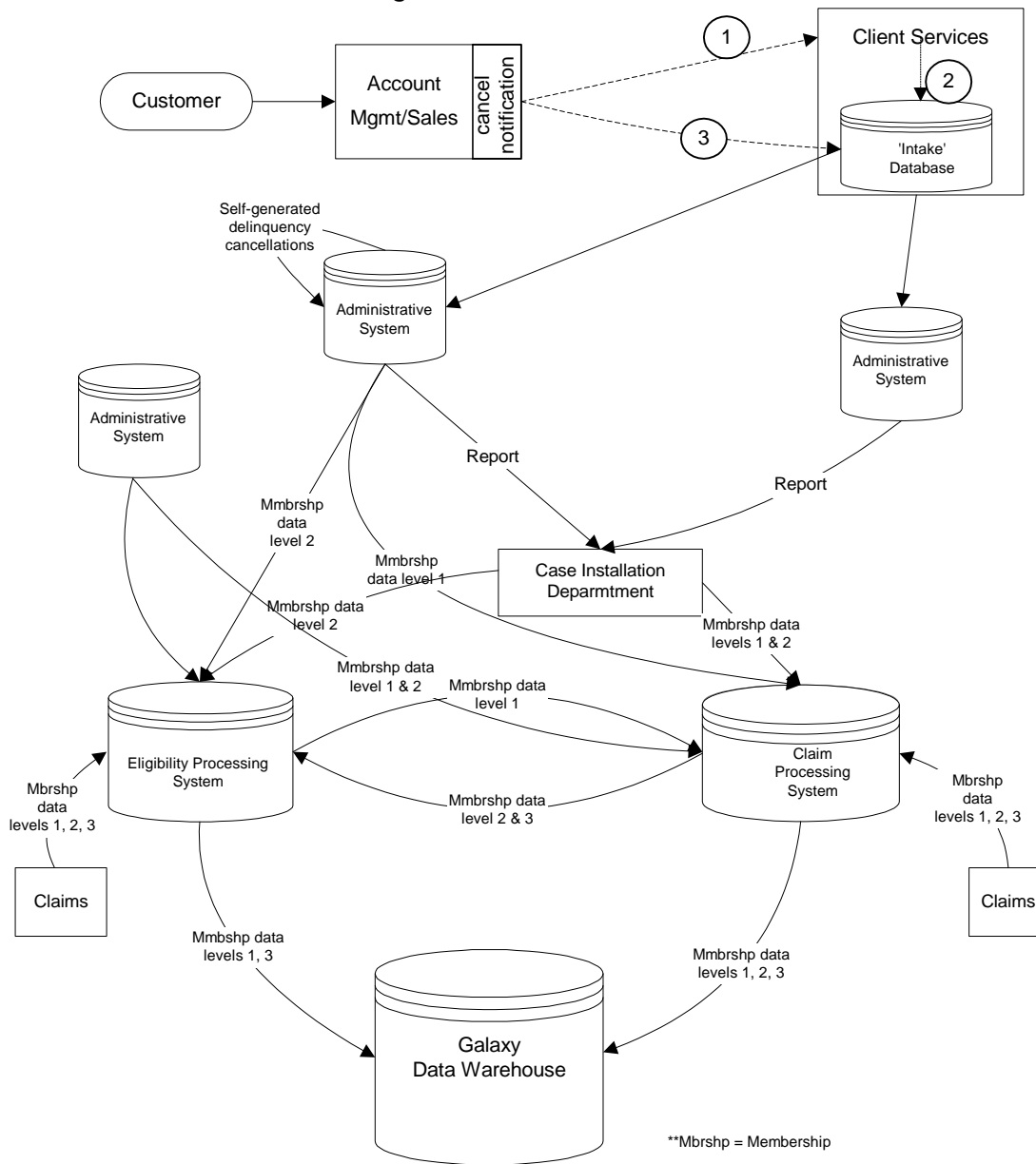
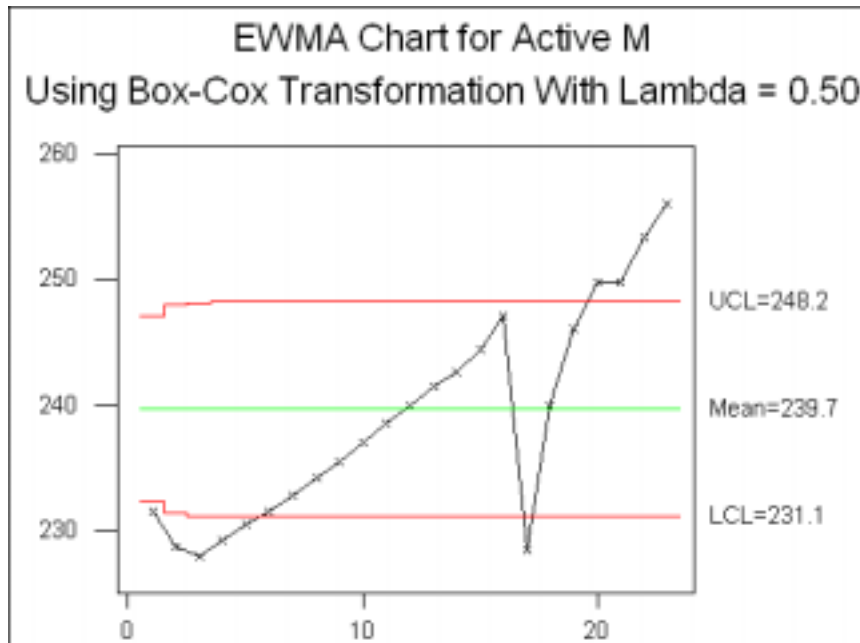


Figure 3 Case Cancellation Process

The team has also put a measurement system in place and looked at past data behaviors. For example, the control chart in Figure 4 plots the number of members who show up as active in the Administrative systems (see Figure 3), but whose policies are, in fact canceled according to data in Galaxy. The chart helps the team estimate the extent of the problems as well as some possible causes.

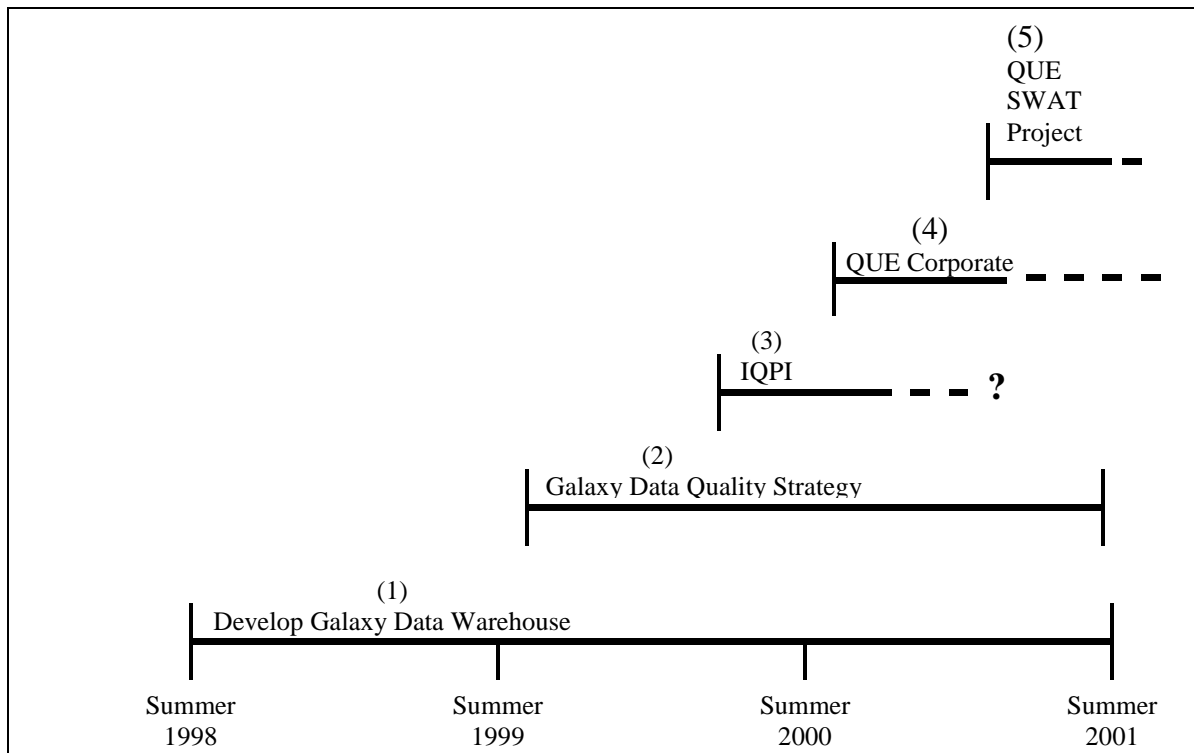


**Figure 4** Control Chart indicates that the Case Cancellation process is not in statistical control.

The team has also begun statistical and in-depth root cause analysis. Information from these activities will form the basis/leverage to request changes to error-prone systems and/or business processes. These changes will result in improved quality of membership data and will be monitored with control charts and other measures. Then another improvement process will begin with a new membership-related data quality project until membership data are highly reliable. While this plan seems fairly straightforward, it is not: it consists of many iterations and its implementation is sometimes organizationally difficult.

## DISCUSSION

Our purpose in this paper is to learn about how organizations keep data and information quality initiatives going in the presence of real organizational dynamics, rather than to focus on the detailed quality definitions, metrics, analysis techniques, etc. of a particular IQ project. To understand better the information quality initiatives at Ingenix and UnitedHealth Group (UHG), we present the time lines shown in Figure 5.



**Figure 5** Time Line of Quality Initiatives

In **Figure 5**, there are five IQ projects, each of a different type:

- The baseline project is the development of the Galaxy Data Warehouse, which is a standard IS development project. As part of this project, the quality of the data is improved by carefully selecting the sources of data to use in the data warehouse. In addition, even before development began, Galaxy was designed to minimize data quality problems.
- Meanwhile IQS interviewed and surveyed users, once about their data quality needs, and again about their metadata needs. These surveys underscored the importance of quality metadata as well as quality data to Galaxy's customers. These quality-related surveys led to broader awareness of the importance of information quality to SDW, Ingenix, and eventually to Galaxy's information quality strategy.

- IQPI started within SDW and Ingenix. It was then moved to a more ‘central’ segment. IQPI was continued for about six more months, then seemingly went into hibernation.
- The QUE initiative is a corporate level quality/process improvement initiative, to be adopted by all business segments.
- Ingenix chose Galaxy data quality as one of its QUE projects.

### **A MORE COMPLEX PICTURE**

This is a more complex picture of information quality initiatives than has been reported in the literature. The paper, “Data Quality in Context”, which studied data quality improvement projects in three organizations, reports three contexts in which data and information quality projects are usually initiated and performed [Strong, Lee, and Wang, 1997]:

- As part of an IS development project
- As an information quality project typically initiated by an information quality champion
- As a TQM project initiated as part of the company’s TQM focus

In that study, each organization used only one of these contexts for information quality improvement. At Ingenix, all three contexts are present at the same time:

1. Data quality improvement is taking place as part of the data warehouse development project (Galaxy).
2. IQS and SDW management ensured progress toward both metadata and data quality. The head of IQS is an information quality champion who was instrumental in planning Galaxy’s IQ Strategy, as well as being instrumental in initiating the IQ seminar and the IQPI initiative. This initiative became a corporate sponsored information quality project.
3. The QUE initiative provides an overall TQM-type context in which IQS formed a cross-functional team to work on a membership data quality project.

Maintaining IQ improvement projects in each of the three contexts can be difficult. Data quality initiatives within IS projects can all too easily be put on the back burner in the effort to complete the development project on time. Independent information quality projects may fail to receive adequate support of resources and of managerial attention to improve information quality in a significant way. Even if they are conducted at a corporate level rather than a local level, projects may falter. Corporate TQM-type initiatives may provide excellent contexts for local information quality initiatives--but companies can find such initiatives difficult to sustain over a long time period.

Ingenix is simultaneously maintaining information quality initiatives in all three contexts. It started as part of an IS development project, added a second context, an independent information quality group that initiates information quality projects, and finally a TQM-type context in which local quality projects (including IQ projects) are developed. These efforts demonstrate Ingenix's growing expertise in information quality improvement.

Ingenix's experiences with IQ improvement projects provide lessons to other companies. The most significant one is that IQ improvement projects/initiatives are not always easy, but by initiating projects in more than one context, these projects have a greater opportunity to succeed. Ingenix's view that IQ is important to its business has given it the vision to see opportunities for IQ improvement within IS projects, within corporate initiatives, and as purposeful information quality improvement projects. All of the three contexts can provide support for improving information quality.

## CONCLUSION

This paper takes the next step beyond our paper at last year's conference, "How to get an Information Quality Program Started" to focus on how to keep an information quality program going. Although the project reported in last year's paper, the IQPI initiative, has faltered somewhat, Ingenix has made significant progress in keeping its information quality efforts alive by using several approaches, each different than the other, and some interacting with each other.

Inherent in this paper are some larger questions: "Why are 'quality' initiatives so difficult to sustain over time?" "Why was the momentum of the original IQPI initiative not maintained? Why are data quality efforts so difficult?" While there may be many reasons, one explanation lies in the very make-up of organizational systems. Looking at it from a 'systems perspective', *events* often are thought to have had a variety of causes: "This and this happened", "they caused it to happen . . ." and so on. Focusing on *events* leads to reactivity. The explanations may be true, but they are not necessarily useful. Systems thinking is *generative*. The fact that the original IQPI initiative was not maintained intact is a symptom, not a cause. Fixing symptoms doesn't help; in fact, it often worsens situations. Root causes would be difficult to arrive at without modeling the organizational system and the dynamics therein.

According to Peter Senge,

"There seems to be a particular lack of appetite in many American [organizations] for the hard work of articulating our mental models conceptually. Developing explicit models. . . of complex [organizational systems] to test alternative processes and strategies strikes many action-oriented managers as too theoretical. This is especially troubling in light of the widely recognized difficulties . . . in transferring [knowledge, information, data] from one group to another."<sup>1</sup>

Another explanation involves the nature and current frequency of change. Macroeconomic forces have made constant change an imperative for organizations that want to survive. However, change in organizations is often met with inwardly focused cultures

---

<sup>1</sup> Peter Senge. *The Fifth Discipline*. Doubleday, New York. 1994



(organizational and/or departmental), complacency if not outright resistance, and most of all, the genuine and natural human fear of the unknown.

It is not within the scope of this paper to address these two issues—1) modeling the organization and its dynamics, including IQ initiatives and 2) resistance to change within organizations as an IQ challenge—except to bring them to light. In addition these issues strongly suggest directions for future studies.

## **ACRONYMS**

CPI	Continuous Process Improvement
DMAIC	Define, Measure, Analyze, Improve, Control
ETL	Extraction, Transformation, Load
IQ	Information Quality
IQPI	Information Quality Process Improvement
IQS	Information Quality & Strategies
IS	Information Systems
MDR	MetaData Repository
SDW	Shared Data Warehouse
SPC	Statistical Process Control
TQM	Total Quality Management
UHG	UnitedHealth Group

## **REFERENCES**

- Kahn, Beverly K., Katz-Haas, Raïssa, and Diane M. Strong, "How to Get an Information Quality Program Started: The Ingenix Approach", *Proceedings of the 2000 Conference on Information Quality*, October 20-22, 1999, Cambridge, MA, pp. 28-35.
- McCue, Michael T. "The Best of Both Worlds: Executive Profile". *Managed Healthcare Executive*. May 2001. p. 22
- Pande, Peter S., Neuman, Robert P., and Roland R. Cavanaugh. *The Six Sigma Way: How GE, Motorola, and Other Top Companies are Honing Their Performance*. McGraw-Hill, New York, 2000
- Senge, Peter. *The Fifth Discipline*. Doubleday, New York. 1994
- Strong, Diane M., Yang W. Lee, and Richard Y. Wang, "Data Quality in Context", *Communications of the ACM*, Vol. 40, No. 5, May 1997, pp. 103-110.

# **A Proposed Framework for the Analysis of Source Data in a Data Warehouse**

M. Pamela Neely  
Marist College  
[Pamela.neely@marist.edu](mailto:Pamela.neely@marist.edu)

## **Abstract**

This paper introduces a framework for the analysis of data quality in source databases, prior to migration to the data warehouse. Additionally, as part of the framework, a tool for the collection of meta-data is proposed. The use of the framework, in conjunction with the tool, will allow the developer of a data warehouse to allocate the scarce resources available for data cleansing. This is accomplished by identifying the data fields that yield the greatest benefit to the warehouse and focusing cleansing efforts on those fields. Additionally, it is proposed that when the meta-data tool is completed, it is possible to assign the task of specific data field identification to novices on the data warehouse development team.

## **Introduction**

In a study by Wixom and Watson (2001), examining the factors that affect data warehouse success, it was concluded that the quality of data in a data warehouse is a critical factor in the success of the warehouse. Wixom and Watson's research supports the previous data warehousing literature, showing that high quality data creates value for the organization. However, although they pose the question, "...can a data warehouse even exist without data quality? (Wixom and Watson 2001, pg. 35)", their research does not show how data quality is achieved.

In this paper, I propose a framework, the Data Quality Analysis Framework (DQAF), for the analysis of source data, prior to migration to a data warehouse. The use of this framework, and a related relational database tool for the collection of meta-data (data about the data), can begin to address the question of how to achieve data quality in a data warehouse. A data warehouse is a dynamic system, growing and changing as user needs grow and change. Data quality is an ongoing concern within the data warehouse and the framework provides a platform for the analysis of source databases throughout the life of the system.

## **Background**

Integrated data repositories, also known as data warehouses, are regularly used to support management decision-making (Goodhue and Wybo 1992) and data mining activities (Forgionne and Rubenstein-Montano 1999). These integrated repositories consist of data from many source databases, which have been designed to support on-line-transaction-processing (OLTP) systems

for day-to-day activities. The data is brought together in one structure to support on-line-analytical-processing (OLAP) systems, which includes multi-dimensional views of data for decision-making, and data mining.

It is essential to both management decision-making and data mining that the data in these repositories are of high quality. Many of the dimensions of data quality, as defined by Wang and Strong (1996), are important in a data-warehousing context. Additionally, fitness for use (Tayi and Ballou 1998) is key. The data must be *accurate* to result in correct decisions. Furthermore, data that is used for secondary purposes, as is the case in a data warehouse, will be judged differently from data that is used for primary purposes, i.e. the transaction processing system. Thus, the *context* of the data becomes a critical determinant in the decision as to the quality of the data. The degree of *completeness* for primary use may be much greater than the degree of completeness required for the data warehouse. Each of these data quality dimensions will contribute to the overall fitness for use as defined by the users of the data warehouse.

The flow of data from source to warehouse is depicted in Figure 1. The data can be examined at multiple points in this flow. Research in source database quality (Storey and Wang 1998) shows us that examining the data at the source is possible. However, much of this research focuses on database design and the necessary steps to take in creating a database that will create an environment where data quality issues will be addressed. For example, validation rules, codes, and date parameters can all be implemented to help ensure quality data in the source. However, at the time of integration it is too late for considering many of these issues. Additionally, many of the challenges associated with integrating multiple data sources are not considered at the source, such as consistent field names across data sources. If data were coming from multiple organizations, then it would have been impossible to address these issues when the databases were created. Thus, although the data can be examined at its source, it is generally not a realistic option to change the source and examining the data elsewhere should be explored.

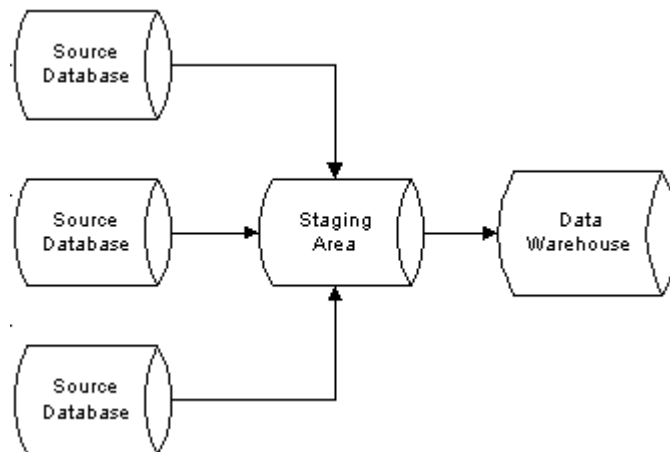


Figure 1- Data Flow from Source to Warehouse

Ideally, the data will be examined in the staging area, prior to migration to the warehouse. Changes (e.g. field names and types) can be made in the staging area that will not affect the source. This allows the developer of the warehouse to analyze data across sources, and determine

exactly what data is needed in the repository. Additionally, the data warehouse developer is in a unique position to evaluate data from a variety of sources and will be able to recommend the best data sources for the warehouse. Finally, the developer can provide feedback to individuals in charge of the source databases regarding the quality of their data.

## Framework Development

The current study involved development of a preliminary framework and related tool for collection of meta-data. The framework was then populated using the results of a series of semi-structured interviews of data warehouse developers and users. In a parallel process, the portion of the framework related to the meta-data tool was tested in two pilot tests. The results of the interview analysis and pilot tests were then used to modify the preliminary framework and create a new framework.

## Preliminary Framework

The research began with the development of a preliminary framework, the Data Quality Audit Process (DQAP). It was built on concepts used in information systems (IS) auditing, database design, and data quality, as well as the financial statement audit framework. This preliminary framework, as shown in Figure 2, consisted of three parts: Planning the Data Quality Audit, Executing the Data Quality Audit Program, and Reporting the Findings.

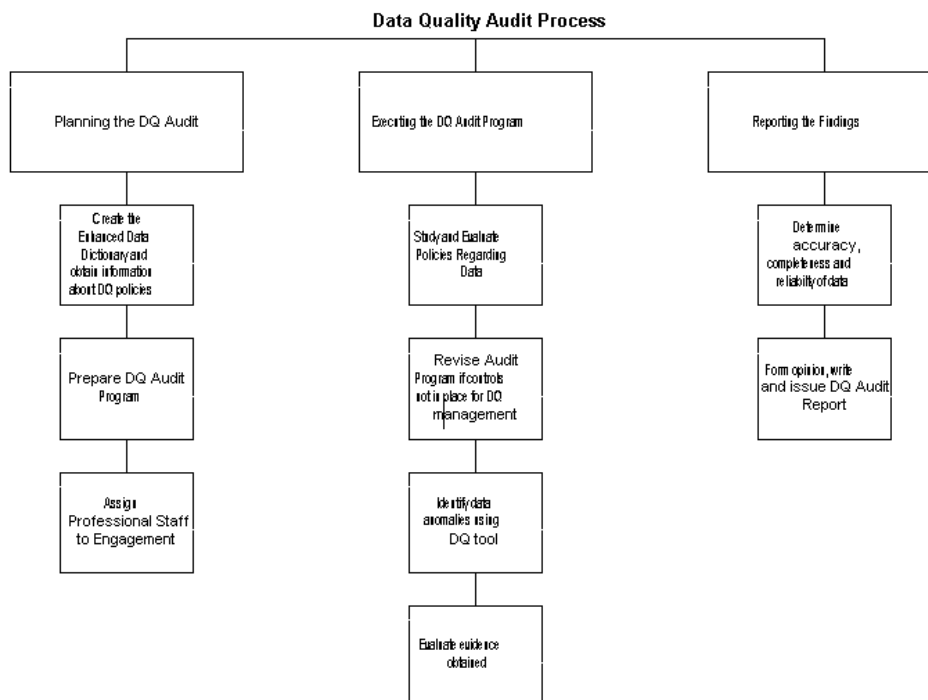


Figure 2- Data Quality Audit Process (DQAP)

The preliminary framework formed the basis for a questionnaire, designed to interview data warehouse developers and users. The results of these interviews were then used to refine and create a new framework.

Embedded in the DQAP is a tool for assimilating meta-data. This tool, known as an Enhanced Data Dictionary (EDD), is an extension of the data dictionary typically found in database documentation. As seen in Table 1, there are columns unique to a data warehouse. For example, the column labeled “Skip?” is used to alert the developer that this data field will not be used in the warehouse and thus, no efforts should be made to determine the quality of the data. The EDD is designed to capture the meta-data that is available from a variety of resources into one document.

Field Number	Column Heading	Field Description	Type/Attribute/Size	Value	Skip?	Field Type
1	Client ID	Unique identifier for each client (Client ID) (internal)	Alpha-numeric /formatted text/15		N	Key
2	Client Name	Client Name	Alpha-numeric /free text/30		N	Text
3	SOURCE_CD	Site	Alpha-numeric/ Coded/15	Fairfield Highgate Other	N	Code
4	Agency ID	Agency ID (this field was used for SSN, then changed to PA#- still also tracking PA# in Identifiers table)	Alpha-numeric/Free Text/15		Y	Text

**Table 1- Enhanced Data Dictionary (EDD)**

It was recognized that implementation of the framework was potentially a multi-year project. Thus, only a portion of the framework was involved in testing. In a parallel process to the interviews, two pilot tests were conducted to determine the value of the EDD and the role it played in the overall framework. The results of these pilot tests further refined the framework.

In the next sections the interview process as well as the pilot test procedure will be discussed.

## Interviews

A series of ten interviews was conducted with individuals responsible for the development and use of a data warehouse. These professionals came from a variety of

backgrounds- industrial, banking, telecommunications, healthcare and government. The questions asked during the semi-structured interview were developed from the DQAP. They addressed the concepts of planning the audit, executing the audit and reporting the findings. Questions were constructed to populate the DQAP, and thus closely followed the nodes of the DQAP. For example, the following questions were asked to elicit data regarding obtaining information about an organization's data quality policies:

- Do you utilize a data dictionary in your process? If so, what purpose does the data dictionary serve in your analysis?
- Who defines your business rules? Are they codified? Who codifies them?

Each interview each lasted approximately one hour. They were taped and transcribed, then analyzed using a qualitative software tool. Key findings from the interviews included:

- Data quality (DQ) was a primary concern for all of the developers
- DQ was considered early in the project, and continued to be an ongoing concern as development progressed
- Data Warehouse developers were not auditors, and did not follow an auditing methodology
- The development of the warehouse generally followed a systems development life cycle (SDLC) approach
- Analysis of DQ had no standard approach. Tools were used when possible, and they ranged from spreadsheets, to programming, to data quality tools
- A data dictionary was unavailable for most of the source data, although attempts were frequently made to collect the meta-data typically found in a data dictionary
- Many of the developers mentioned the criticality of knowing the data suppliers and having a place to go when problems arose

## **Pilot Tests**

In a parallel process, 2 pilot tests were conducted to test the effectiveness of a portion of the DQAP, specifically, the design and use of the EDD. The pilot tests were conducted in two undergraduate classes, one focused on data quality and the other a data management class. The students were instructed to construct a portion of the EDD related to one data source, using the source documents that were available. These documents included a list of field headings and descriptions, as well as a codebook. The students were then asked to analyze a completed EDD for several data sources related to the same warehouse project. Their analysis was designed to highlight "across data source" anomalies such as homonyms and synonyms as well as incompatible codes and field types. Finally the students were asked to look at actual data from the data sources and complete the EDD using this visual inspection of the data.

## ***Data Quality Analysis Framework (DQAF)***

As a result of the interviews and the pilot tests, a proposed framework, the Data Quality Analysis Framework (DQAF) was developed (see Figure 3). This framework differs considerably from the DQAP, principally because of the focus of the developers. They think in terms of systems development, not auditing. Although they found the questions regarding audit

professionals and reporting audit findings intriguing, they did not see them as relevant in the process of data warehouse development. Overall, they felt that the audit considerations should be a concern for another group. Their goal was not to give an opinion on the data, but rather to ensure that it was acceptable to meet the needs of the data warehouse.

The DQAF is designed to fit into the analysis phase of the Systems Development Life Cycle (SDLC), a structured process for developing information systems. It is an iterative process and attempts to incorporate best practices that were elicited from the interviews and to address the deficiencies found in the current practice. The framework is continuing to evolve as the research progresses.

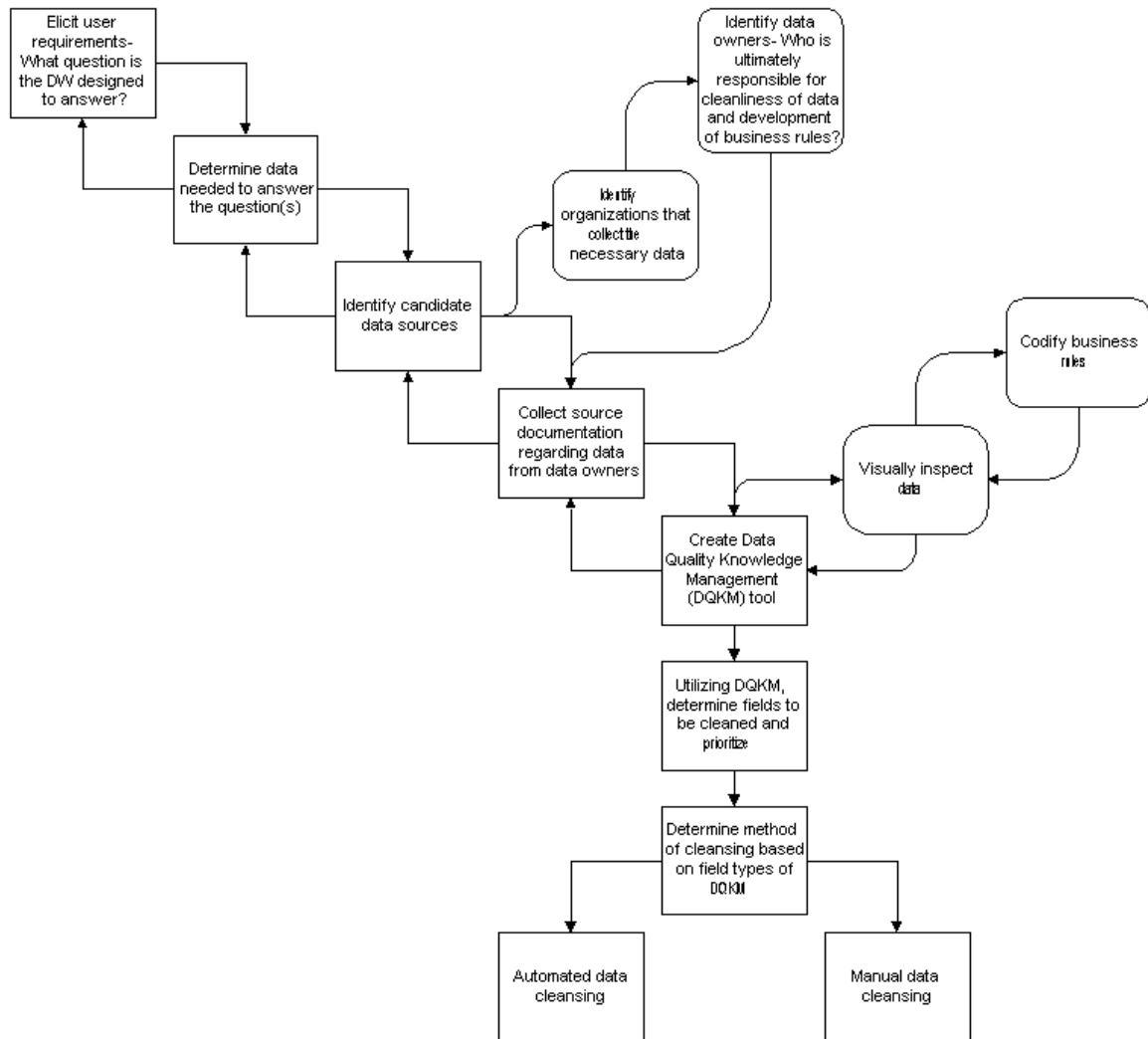


Figure 3- Data Quality Analysis Framework (DQAF)

As noted previously, the Data Quality Analysis Framework (DQAF) is an iterative process. The interviews clearly indicated that not only was data quality considered early in the

development process, but also considered continually throughout the process. Thus, the DQAF shows many loops that resemble the waterfall approach of the SDLC. In the next section I will discuss the various components of the DQAF.

## **Elicit User Requirements**

A key activity in the analysis phase of the SDLC is gathering user requirements. In a data warehousing project the object of this phase is to determine what questions the data warehouse is designed to answer. Thus, this is where the data quality analysis logically begins.

## **Determine Data Needed to Answer the Questions**

In determining the questions that the warehouse should answer, it is critical to know what data will be needed to answer the questions. As an example, consider a system designed for a governmental agency that provides services to the homeless population. They want to know if the services that are provided have an affect on the homeless population. What is the recidivism rate for a given population given a specific mix of services? In order to answer this question, data is needed on length of stay (how long the individual stays in a shelter and how often they return to the shelter once they leave), what services are provided to the individual, and the demographics of the population. Identification of these data is the next step in the analysis of the quality of data.

## **Identify Candidate Data Sources**

Once the needed data is known, the identification of available data sources begins. In a typical data-warehousing situation, the needed data may be available in multiple locations. Identifying candidate data sources encourages the warehouse developer to look “outside the box” and consider sources that may not at first appear to be relevant. This data may vary in quality. However, in the early phases of the framework, it is important to identify all of the available sources. A goal of this framework is to build a repository of meta-data that can be used to facilitate knowledge management. Capturing all of the available sources will yield a richer repository.

## **Identify Organizations**

Identification of the organizations will be a natural extension of the previous step. In the example used earlier of the homeless system, data will be found in the agency as well as homeless shelter providers. In other situations, the data is intra-organizational. In this case the identification of organizations will be identification of departments or subdivisions. This phase should be customized to fit the warehouse being developed.

## **Identify Data Owners**

A critical finding in the interviews was that the data providers must be involved in the process of building the warehouse. As the data was analyzed, exception reports were generated. Automated tools could identify discrepancies between what should be and what was, but it took



human intervention to determine how the data should be changed. For example, an automated tool could determine that an individual left a shelter without ever entering it by comparing the fields associated with admit and depart dates. However, human intervention was necessary to determine if the individual was ever in the system or not.

Identification of the data owners has multiple benefits. First, it involves the people who know the data best to be involved in the process, thus aiding the process of data cleansing. Secondly, the process of building a data warehouse involves integrating data from disparate sources. Previously, these data suppliers would not have had contact with each other. Thus, identification of data owners, and incorporation of this data into the meta-data repository, will allow a better view of the “big picture” of an organization or data-warehousing project.

## **Collect Source Documentation**

Once the data owners are identified, they can then be asked to provide source documentation that will enable the developer to build the repository of meta-data. Source documentation can include, but is not limited to, data dictionaries, codebooks, lists of field names, and other documentation generated by the data suppliers. This source documentation provides the foundation for thorough understanding of the data available for the warehouse. In turn, a complete familiarity with the data will help the developer make better decisions regarding the appropriate actions to take as far as cleansing the data.

## **Create Data Quality Knowledge Management Tool**

A critical component and focus of the DQAF is the Data Quality Knowledge Management (DQKM) tool. This tool, embodied in a relational database, is designed to be a repository for the meta-data associated with a data-warehousing project. The DQKM is designed to collect traditional meta-data, such as field size and type, as well as information about the source data organizations and the individuals who know and understand the source data. The DQKM grew out of the Enhanced Data Dictionary (EDD) that was developed with the DQAP. The EDD went beyond the traditional data dictionary and collected meta-data that is suitable for evaluating data to be included in a data warehouse. It provided information allowing the developer to look “across” databases and determine if there are homonyms (fields that have the same name but store different values, i.e. a field named counselor in two databases- one of the databases refers to the counselor for the patient, another database refers to the counselor that is in charge of the unit) and synonyms (fields that have different names but collect the same values, i.e. SSN and student number).

However, as the amount of meta-data to be collected grew, it became obvious that a flat file method of storing the meta-data was inefficient. Data is collected regarding organizations, data suppliers, business rules, context of use, quality dimensions and appropriate field values, as well as the standard data dictionary components of field size and type. See Figure 4 for the relational schema.

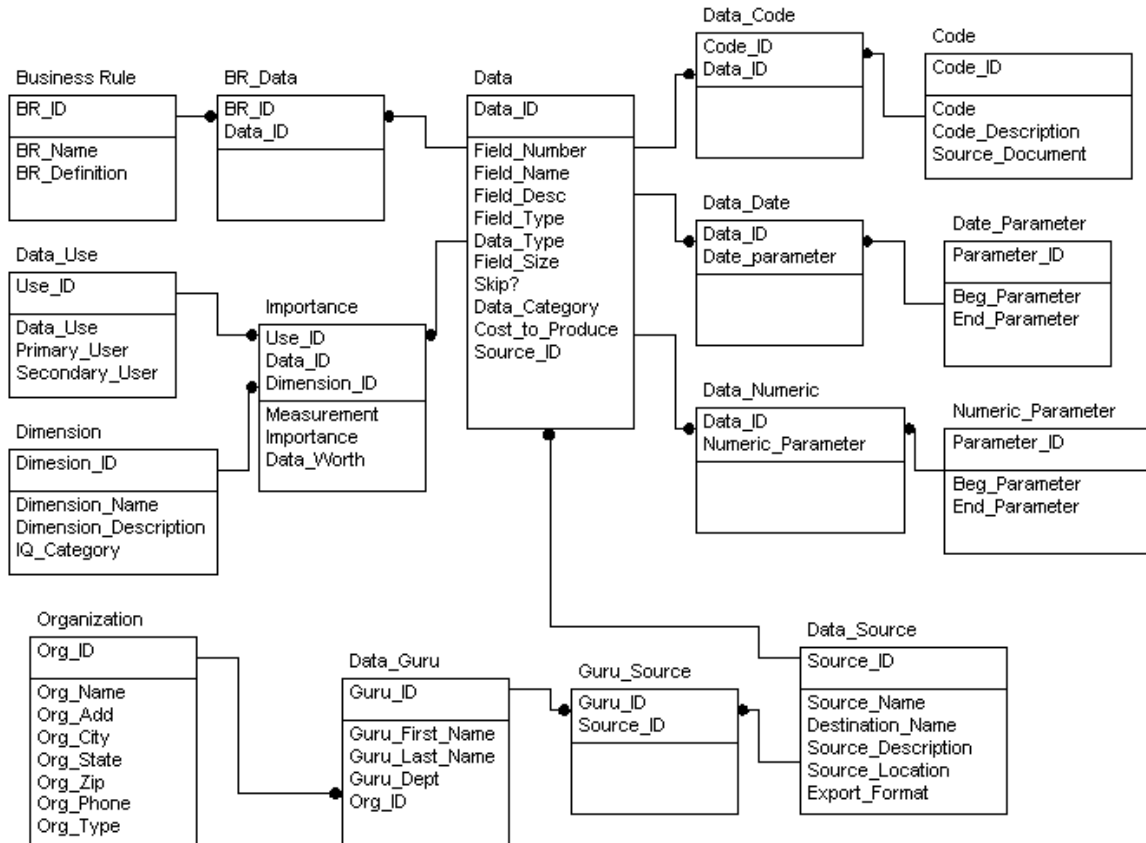


Figure 4- Schema for DQKM

The DQKM requires an enormous amount of effort to construct. Collecting and entering all of the meta-data is very time consuming. My previous experience with construction of the DQKM involved 40 person hours for roughly 5000 records. However, population of the DQKM is easily divided among a number of individuals, each with a different expertise. Thus, the task can be accomplished in a short time frame, as needed. Additionally, making decisions regarding the importance of data requires a great deal of expertise and judgment. The importance table fields are coded based on the data quality dimension being considered, as well as the use of the data. Thus, a particular field may be adequate for use 1 regarding accuracy, but not for use 2. Or a field may be adequate on completeness for use 3 but not on timeliness.

However, the result of all of this effort is a tool to facilitate knowledge networking. The meta-data that has been captured can be used in making decisions regarding the specific data fields on which to spend time and money for cleaning. A given warehouse development team may be composed of individuals who are experts as well as individuals who are novices. Once the DQKM has been constructed, the novices can make the decisions regarding cleansing, as this will be a SQL query. The knowledge is in the tool and can be extracted by novices. For example, a query on the data category will generate the available fields to answer a specific question. If the novices are provided with criteria for deciding the specific fields on which to expend cleansing resources, they can query the database and generate a listing of fields that meet the criteria. Thus, the decision as to exactly which fields should be cleaned has been automated.

Because the DQKM captures so much meta-data, it is also a repository that can be used to provide information when individuals who know the data leave an organization. Additionally, the wealth of information collected in the DQKM allows for a “bigger picture” look of the interrelationships among the data elements. Insights provided by this look at the data may provide new avenues to explore for competitive advantage. For example, querying the database on data category will determine what fields are available detailing a specific category such as gender or ethnicity. These fields will come from a variety of sources and analysis of these sources could alert management to redundancy in data collection efforts. Decisions could then be made as to which sources can be eliminated, thus reducing overall costs.

## **Visually Inspect Data**

The DQKM is an evolving tool. Once the meta-data has been entered from source documents a visual inspection of the data in the sources will enable a greater understanding and clarification of the data. For example, in the absence of a data dictionary, field size and type can only be determined by a visual inspection of the data. The richest DQKM will be built utilizing all of the available resources.

## **Codify Business Rules**

An essential element in determining what data is needed to populate a data warehouse and answer the necessary questions is a definition of the business rules. For example, in the homeless system, how is length of stay defined? Is it admit date to depart date for one shelter? Or is it the length of stay in the system, even if they move from shelter to shelter? Definition of the business rules, and the data needed to support them, is necessary for a well-defined data warehouse and will enrich the DQKM.

## **Utilize DQKM to Determine Which Fields to Clean**

Once the DQKM has been fully populated it can be queried to determine where to focus the data cleansing efforts. After defining standards for cleansing, an analyst can query the database with the criteria set. For example, if it is determined that for use 1, accuracy scores greater than 65% should be cleaned further, a simple query will then determine what fields fall into that category. After determining which fields are candidates for cleansing, then the fields should be prioritized. Thus, a novice, using pre-defined criteria, can analyze and prioritize data fields that will make the best use of scarce resources and offer the greatest benefit to the project.

## **Determine Cleansing Method**

Once it has been determined which fields should be cleaned and they have been prioritized, it is necessary to determine if the fields should be cleansed manually or automatically. Data cleansing tools are appropriate for fields that can be compared to another source, such as a list of codes, range of dates, or postal address lists (Neely 1998). By querying the database, it can be determined if these comparative values are available and what the values are. Conversely, if no values are available for comparison then the data must be cleansed

manually and appropriate steps should be taken to ensure that the data is returned to the data supplier to verify the data.

## **Further Research**

A development team for a data warehouse will consist of both experts and novices. The framework will be used by the team as a whole in the development of the warehouse. However, experts and novices will perform different tasks in the accomplishment of the goal of determining where to focus cleansing efforts.

The framework and tool described in this paper are being tested in a three-phase approach. The first phase consisted of the pilot tests conducted using the DQAP framework and EDD. This phase provided data to construct the DQAF and DQKM. Phase two was conducted after construction of the framework and tool. This phase involved testing the ability of novices to determine which data fields should be cleaned given a completed DQKM. Phase three will be a modification of the second phase, and will involve prioritizing the data fields for cleansing in addition to the identification of them. The goal of the second and third phases is to show that novices can use the DQKM to make decisions regarding the cleansing of data.

Future research will involve more extensive testing of the DQKM. At this point only the portion related to importance has been tested. Testing regarding the ability to determine whether data should be cleaned manually or automatically needs to be done, as well as testing the framework in its entirety.

## **References**

Forgionne, G. and B. Rubenstein-Montano (1999). "Post Data Mining Analysis for Decision Support through Econometrics." Information, Knowledge and Systems Management **1**(2): 145-157.

Goodhue, D. L. and M. D. Wybo (1992). "The Impact of Data Integration on the Costs and Benefits of Information Systems." MIS Quarterly **16**(3): 293-311.

Neely, M. P. (1998). Data Quality Tools for Data Warehousing- A Small Sample Survey. The Conference on Information Quality, Cambridge, MA.

Storey, V. C. and R. Y. Wang (1998). Modeling Quality Requirements in Conceptual Database Design. 1998 Conference on Information Quality, Cambridge, MA.

Tayi, G. K. and D. P. Ballou (1998). "Examining Data Quality." Communications of the ACM **41**(2): 54-57.

Wang, R. Y. and D. M. Strong (1996). "Beyond Accuracy: What Data Quality Means to Data Consumers." Journal of Management Information Systems (JMIS) **12**(4): 5-34.

Wixom, B. H. and H. J. Watson (2001). "An Empirical Investigation of the Factors Affecting Data Warehousing Success." MIS Quarterly **25**(1): 17-38.

# External Data Selection for Data Mining in Direct Marketing (Practice-Oriented Paper)

Dirk Arndt, Wendy Gersten

DaimlerChrysler AG, Research & Technology, Data Mining Solutions, FT3/AD,

PO BOX 2360 89013 Ulm, Germany

{dirk.arndt, wendy.gersten}@daimlerchrysler.com

**Abstract.** Today the purchase of external data is necessary for most direct marketing applications. No company can refer to the internal data alone, especially when targeting new customers. This paper discusses an integrated approach detailing how to select external data sources properly. For that, we try to standardize the selection process, to make it repeatable and to give practical hints in order to overcome handling issues. Therefore, we talk about tools, experiences and perspectives. We start with a detailed problem description and then develop a general process model. In subsequent sections, we discuss how to collect, measure and aggregate the selection criteria without losing too much information quality.

## 1 Introduction

Uniformed products, along with individualization of customers, has brought pressure for change in marketing practices. This implies that additional product benefits are generated by means of communication and services that are designed and delivered to match the customers' individual expectations and needs. This is one of the main goals of direct or database marketing.

“The new *direct marketing* is an information-driven marketing process, made possible by database technology, that enables marketers to develop, test, implement, measure, and appropriately modify customized marketing programs and strategies. [14]” *Data Mining* is the process of data exploration and analysis, which can be used to support these tasks [3]. In order to develop customized or even personalized dialogs and services in direct marketing, marketers use, e.g., attitude, lifestyle, behavioral, and usage information (data). Generally, these data are available from different data sources within and outside the company [1].

Although directly captured data (*internal data*) provides unique information concerning our own customers, brands and products, these data are not always available or of sufficient quality. In many cases, purchasing additional data from outside the enterprise (*external data*) can enhance the overall data situation for the direct marketing tasks on hand [2].

If a company intends to buy external data, it faces several difficulties. Since there are different types of data sources offered by multiple data providers, it is quite hard to find the best choice.

Today's business practice often leans towards convenient and inconsiderate ad hoc decisions. Consequently, many attempts to develop problem solving data sources are sentenced to fail [7]. Therefore, we demand a standardized assessment approach, containing a process model and proper comparison criteria. In this paper, we introduce an approach that tries to fulfill this demand. It was developed and tested for the selection of external data by DaimlerChrysler.

In section 2, we start with a problem description. Next, section 3 introduces the complete process model. The individual steps of this process are described in section 4. Here we discuss for each step, how to execute the tasks, what experiences we have made and what difficulties we were confronted with. In sub-section 4.3, we explain the most intensive process step (close-up-examination). For that reason, it is more detailed and includes a system of comparison criteria.

## **2 Problem Discussion**

In this section, we aim to give a quick overview about the main complications of the overall problem, as we experienced them in practice. In sections 3 and 4, we return to these drawbacks and try to give hints on how to overcome them.

One problem aspect is that the question we want to answer is *not one-dimensional*. If we want to buy external data, we need to make three relevant decisions, which largely influence each other. We have to choose among the different kinds of data (lifestyle, census, etc.), between the diverse providers of these data and, finally, we are required to pick the attributes within the data sources.

In order to do so, we must first describe the primary objective of the project from a business perspective [5]. Often we face *many competing objectives and constraints* that are important for the decision. If we intend long term usage of the data (e.g. creation of permanent fields in the customer database), the situation gets even more complicated. Here we do *not exactly know what future business problems* we will face. But if we want to give the right answers in the future, we have to collect the necessary data today.

After defining the business problem, it has to be transferred to a data mining goal. A data mining goal states the business objective in technical terms [5]. The data mining goal corresponds with the data mining algorithms we plan to use. And, these strongly depend on the data input [4]. As we do *not know in advance whether the data mining results will solve the business problem*, we might have to change the data mining goals and algorithms. This may cause that the chosen data do not fit anymore [8].

Talking about unfitting data, we are confronted with another problem. If we want to measure data quality, we need to find proper measures [11]. For example in [10] Data Quality Mining (DQM) is introduced as a new approach to address data quality issues by means of data mining methods. The overall intention hereby is to *gather the information* without mistakes or errors, to *derive manageable scales* for information measurement and to *aggregate it* for the final assessment. We will address these points in more detail in section 4.

Besides the aspects mentioned above, typically the decision is made *under pressure of time and resources* (mainly human resources, money and hardware). Unfortunately, we are seldomly able to reduce the caused results by means of experience, because we *cannot build on prior*

*knowledge*, due to employees leaving the company and insufficient documentation. These effects are strengthened when there is just one person in charge. Additionally, here we cause high *subjectivity of the decision*.

### **3 The overall process model**

As mentioned before, we now describe the overall process model, developed and tested by DaimlerChrysler. First, we consider the adaptation level of the model to the respective project. Second, we explain the connection between the single steps and why they are created at all and ordered in a particular way. We will have a closer look at the steps in section 4.

When the idea for standardizing the data selection process was born, we aimed to create a detailed user guide. After a short time it became clear that such an approach is not possible. We realized that each project, even within the field of data mining for marketing, is much too specialized and too complex for this approach. Consequently, we changed the goal. Now we aim to provide a generic framework, which has to be adapted for each selection.

The more detailed and the more accurate the adaptation is executed, the more time and budget is needed. The energy spent for that should match the relative importance of the project. There is a wide range of possible solutions.

*E.g. for the evaluation of our approach we had two people working 40% of their time over a period of six months. Additionally, we had a team of experts standing by. But when we helped to choose a data provider for a large but single acquisition campaign in the UK, we needed only three full work days for preparation and one workshop with five people in order to complete the task (over a period of two weeks).*

Now the question is how to determine the relative importance of the project and the corresponding effort. Again, the attempt to be very exact would be a waste of time, because there are too many influences. For that reason, we cannot give exact instructions. But we like to point out two major aspects.

In practice, we found that one of the main aspects to consider is *for how long we intend to use the data or the resulting information*. The longer the usage is planned, the more expensive is the project and the more the future business will be influenced. Naturally, we would put more time and resources into the selection as the expected impact increases.

Another important aspect is the *strategic relevance of the business goal*. Even if we use the data temporarily, the results may have long term effects if they are used for strategic decisions. That is why we prefer a more intensive selection of data in this case. In case of short-term operational goals, we would keep the selection process much simpler.

In section 4, we will mention what precise choices we have to adapt the process and what the impacts of these choices are. For now, we want to look at the process model. For the most part, the model can be used independently of the fact that it was developed for the selection of data for direct marketing. We will outline the point where this comes into account later on.

The selection of (external) data sources is part of the overall data mining process. That is why we see our process model as one block of activities within the data mining project plan. For the execution of data mining projects we refer to the CRISP-DM process model, which is an open industry standard [6]. Fig. 1 illustrates our model for the selection process of external data sources.

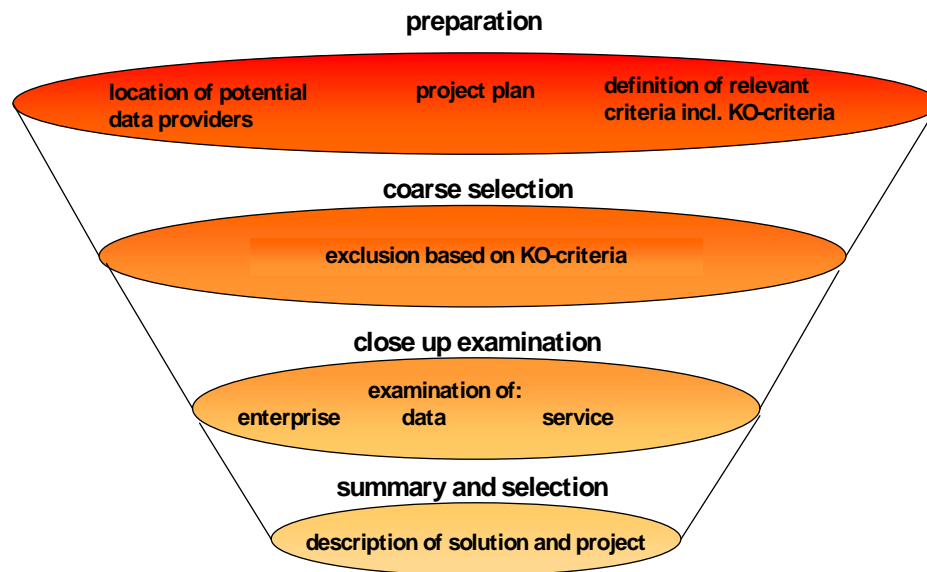


Fig. 1. Overall process model for data selection

Each selection starts with preparation. The most important outcome is the *initial project plan* for the data selection, which corresponds closely to the project plan of the respective data mining project [5, 6]. The plan is necessary because we need both an internal status quo (e.g., the timelines for the project) and basic knowledge of the possible external data providers (e.g., addresses and phone numbers), before we can contact the latter. At the beginning, we *consider all possible alternatives of potential data sources*. Hence the funnel in Fig. 1 has its widest diameter.

The next step is to contact all possible data providers. The aim is, first to gather information and, then if possible, to *reduce the number of providers* based on KO-criteria defined prior in the project plan. This is very important for saving time and money. This way we can exclude candidates, we would have excluded later anyway.

We call the third step close-up examination. Independent of the process adaptation, this is the *most time- and resources -consuming phase*. Here we evaluate the data as well as the data providers. To do so, we need an intensive dialog and data transfer with the vendors. We developed an evaluation approach based on three dimensions. The outcome of this step is an evaluation portfolio, illustrating the *position of all data providers* (except the ones excluded in step two), as we will explain later on.



During the last step, we make the decision and produce the *final report*. The report is mainly a summary of the selection process and its experiences. It helps to understand the decision process in future projects and to store the knowledge gained. So we overcame one of the complications mentioned in section 2.

## **4 Detailed Model Description**

### **4.1 Preparation**

The first task in preparation is the *determination of the business and data mining goals*. We can obtain the primary objectives from the data mining project plan and transfer them into sub-goals for our data selection. We recommend defining just one (or two corresponding) primary goal(s) and to submit all other goals strictly. This will help to avoid target conflicts as mentioned in section 2. If there are more key objectives we would handle them within separate projects. Note that this decision influences the expenditure we should spend for the whole selection (see section 3).

Now we have to execute a *situation assessment*. Therefore, we list all resources available to the project (e.g. personal, software, hardware, data). Again, we can use lots of information from the data mining project plan and add our specifics.

After defining the goal(s) and having assessed the situation, we *derive and weight the selection criteria*. We need these requirements in order to contact the potential providers properly, as we will explain in section 4.2. A general system of evaluation criteria is described in section 4.3, where the actual evaluation takes place. To derive and weight the criteria, we built a team of people from all relevant departments (e.g. Marketing, IT, Controlling, Management, etc.) and organize a workshop. Here we use common techniques like work groups, brainstorming, brown paper method or sensitivity analysis.

Within the criteria found, we must *name KO-criteria*. If one KO-criterion is positive for a specific data source (provider), the source (provider) will be excluded for good early in the selection process. Because of that, we must be very careful when picking the right KO-criteria. We also should take into account that we can apply the criteria easily. This is necessary because we want to sort out adequate data sources with low expenditure (see section 3).

*In practice we experienced that KO-criteria are found straightforwardly by means of brainstorming. One example of a good criterion we found that way is the image of the data provider. For DaimlerChryslers premium brand Mercedes Benz it is very important not to work with data providers who have a bad reputation in public. Especially, if we work with data of private persons for marketing purposes. The criteria are relatively easy to apply as well (e.g. we can search press articles for the providers name).*

Another task to fulfill during preparation is to *locate potential data providers* (gathering of information like names, phone numbers, addresses etc.). For that we can use public information sources like the world wide web, yellow pages or business address providers.

From all the tasks described before, we *develop the initial project plan* for the data selection. It represents the intended plan for achieving the defined goals and lists the precise activities to be executed, together with duration, resources required, inputs, outputs as well as dependencies.

All these tasks must be completed for every selection process. This means that there is no way to adapt the process here. The only difference is that we have varying intensity depending on the nature of the business goals and the intended time of data usage (see section 3).

## **4.2 Coarse Selection**

After preparation we start to contact the providers of potential data sources according to the project plan. We can accomplish this task through *oral or written interviews*. In any case, we suggest using an uniformed questionnaire. So it is less complicated to compare the results. The *questionnaires should include basic information like date, contact, phone, etc., all KO-criteria* as well as a first look at the *most important criteria*.

Most important are these criteria which were highly weighted during preparation. The early evaluation of these criteria is essential for three reasons. First, if we do not have the time or resources to check all criteria derived, we are able to *find the most promising ones (e.g. in terms of the degree of assessment, measurement, reliability, etc.)* near the beginning. Second, if we gather the information during coarse selection, we can cross-examine it during close-up examination and, hence, *increase reliability*. Third, we are capable of using the gathered information for a *first ranking of the data sources before entering close-up examination*. The latter can help to speed up the whole process or save costs later on.

The next step after making the first contact is the *exclusion of data sources or providers based on KO-criteria*. As mentioned before, we sort out a source or provider if one or several KO-criteria are positive (see section 4.1). But often, we can obtain only uncertain information. That is why we advise *rechecking the results* if we are about to exclude a presumed high potential source (provider). A source or provider is considered high potential, e.g., if there is a wide range of information offered, if it is a major company (e.g. in terms of market share, market experience, service offerings, etc.) or if we have good experiences from the past. We are not able to provide a certain and complete list of criteria, because again the criteria and the accesses to the corresponding information vary among different projects.

*Yet, to outline the importance of the recheck we want to give a short example from one of our projects. When we did the coarse selection for a long-term strategic marketing project, we were about to exclude one data provider (and therefore several data sources), because there was no service hotline offered. The whole coarse selection step lasted several weeks (because of internal difficulties by DaimlerChrysler). When we rechecked the criterion it came to our attention that a new service hotline was about to be established for free. The person who had given the information the first time did not know about this fact. Later in the process this provider was chosen exclusively.*

The outcome of this process step is a *list containing all data providers to be evaluated in close-up examination*. The list includes a first ranking and goes along with basic information about the most important criteria in best case.

In contrast to the first step, we have a variety of possibilities for process adaptation here. We can choose, at least for the type of interviews, the inclusion of most important criteria and the addition of the recheck task. Of course, there are several levels of intensity possible again.

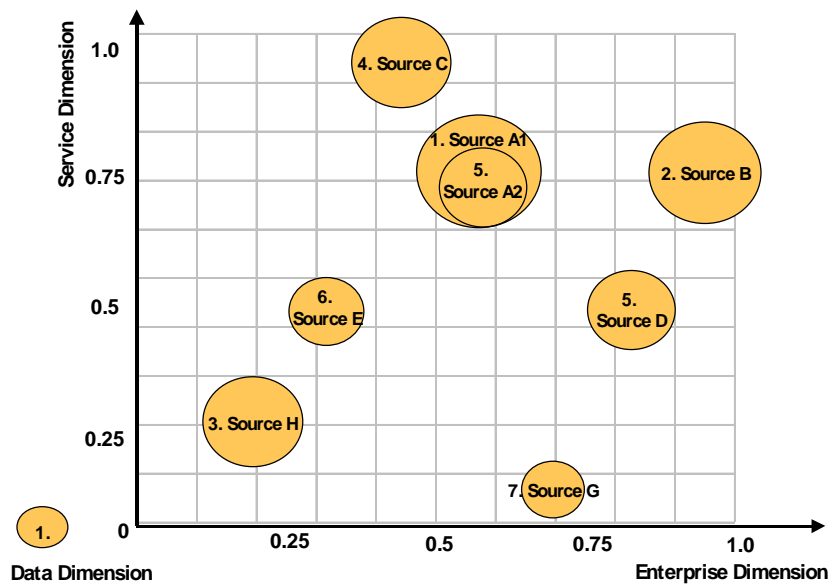
### **4.3 Close-Up Examination**

Entering the phase close-up examination we reach the core of our process model. In this section, we start with talking about the general tasks to fulfill, explaining the dimensions of the evaluation and discussing the problems of criteria measurement. Then, in sub-sections 4.3.1 through 4.3.3 we describe a framework for the arrangement of the criteria within the evaluation dimensions.

The aim of the close-up examination is to *evaluate and compare each data source (provider) with all others*. Therefore, we go back to the providers and have a closer look than we had in step two. But of course, we use the information obtained before as a starting point and for reference. What tasks we have to complete in detail will be mentioned in the appropriate sub-sections.

For the examination we suggest a *three dimensional evaluation space*. Since we talk about information quality and buying external data, naturally, the most important dimension is the *data dimension*. But in a business environment we have to consider other aspects as well. As our example in section 4.1 shows, there can be significant criteria concerning the enterprise which is offering the data. We found several such criteria and for that reason, we grouped them to yield our second dimension, the *enterprise dimension*. The last dimension we suggest is the *service dimension*. Here we combine all criteria regarding the service level of the data provider.

The evaluation dimensions as well as the corresponding criteria are arranged after our needs and experiences. Because of that, *the arrangement may be expanded, reduced or reorganized according to specific project demands* and represents a general suggestion only. Here is *much room for process adaptation*. In practice we found that most criteria can be sorted into the framework and that it is therefore a helpful tool for organizing the evaluation. The three dimensional evaluation space can be illustrated through the portfolio technique [12]. Fig. 2 shows an example.



**Fig. 2.** Final evaluation portfolio

The service and the enterprise dimensions are represented by the two axes. The data dimension is shown through the size of the circles and the corresponding numbers. This *final evaluation portfolio is the outcome of the close-up examination*. It illustrates the relative position of all data sources. If a certain provider offers more than one data source and if we want to view them separately, the circles will have the same center but probably different diameters (see Fig. 2).

If there are two or more sources close together and we are uncertain which one to prefer, we advise making another *recheck*, at least concerning the data sources in question. During the recheck we can verify the former results, use new measures for the information gathered before or collect additional information. The recheck is necessary because there are several *inaccuracies and uncertainties in the measuring and the combination of the criteria*. Before going into the sub-sections we want to talk about these difficulties in general.

The first challenge is to ask the right questions during the data (information) collection. That means we have to closely and correctly specify the wanted information in advance. Only this way we can be certain that we obtain the *intended information* and that it is *comparable* later.

*We want to explain the fact with an example from practice. If we ask a data provider for the turnover, he can state the turnover for the whole company. In the case of a diversified company like Bertelsmann (or GE) this would be a huge amount. But is this really the information we want to obtain and can we compare this number with the turnover of a much smaller data provider? The answer to both questions is no. Instead we should have asked for the turnover of the specific subdivision in question.*

The second challenge is to measure and aggregate real world information without distorting it too much. The data containing the information can be qualitative and quantitative in their nature

and thus, demand different types of measurements. Yet, they all have one feature in common: they are all made on some kind of scale. In detail, we distinguish the following main kinds of scales [11]:

- Nominal scale,
- Categorical scale,
- Ordinal scale,
- Interval scale,
- Ratio scale.

The list of scales above is ordered after the information content (amount of information) they carry and could be divided even further [13]. With the aim of producing an aggregated view at the data sources, we have to transfer information from one scale to another as well as to aggregate it. In order to make this task as simple as possible, we advise *thinking about the scale for each criteria carefully before starting the data (information) collection*. Again, there is no general approach for data collection or transformation. We have to *find practical solutions in each case*.

We would like to give an example for scale transformation and aggregation of information. Fig. 3 shows a table containing two criteria measured with different scales: number of available addresses and overall completeness of records. Four potential data sources (A, B, C, D) are evaluated. In the example, both criteria are weighted equal (with 0.5). First, we transfer the scales (transformation rows; the biggest number corresponds with the highest rank) and then we calculate the aggregated value (as shown in the last row). The aggregated value is generated through the calculation of the relative value for each criterion and the summarization of all relative values (e.g. the calculation for Source A is:  $1:3 * 0.5 + 1:4 * 0.5 = 0.29$ ).

Name of Criterion	Weight	Source A	Source B	Source C	Source D
<b>Number of addresses</b>	0.5	300,000	304,000	600,000	1,220,000
Transformation 1 (ordinal scale)		< 500,000	< 500,000	< 1,000,000	< 1,500,000
Transformation 2 (rank)		1	1	2	3
<b>Completeness of records</b>	0.5	87%	90%	95%	99%
Transformation 1 (rank)		1	2	3	4
<b>Aggregated Value</b>		0.29	0.41	0.7	1.0

**Fig. 3.** Example for scale transformation and aggregation

*This example shows that the transformation process is highly subjective and error-prone. In this case, e.g. we decided that the difference concerning the number of addresses in sources A and B*

is not large and that we treat them as equal. But one may find reasons not to do so. It gets even more complicated if we have to aggregate qualitative and quantitative attributes. The transformation into ranks might be a working solution for this problem as well.

### 4.3.1 The Enterprise Dimension

This dimension aims to evaluate general enterprise criteria of potential providers. As we will show, these criteria mainly refer to the characteristics of the data provider. Fig. 4 gives an example for possible evaluation criteria and how they can be arranged.

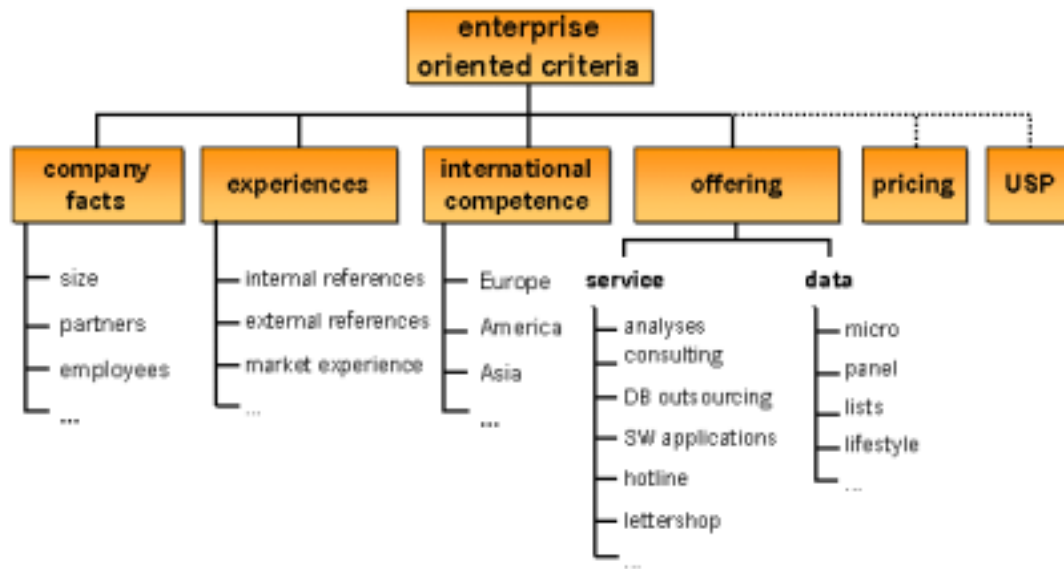


Fig. 4.

Examples for criteria in enterprise dimension

The cluster *company facts* summarizes information about the providers business. Here we consider criteria like business partners, turnover or number of employees. These help to estimate the available personnel and financial resources having an impact on the possibilities of collaboration. Furthermore, they give valuable hints if the provider has substantial power to develop innovative approaches or to react to our future demands (see section 2).

Within the second group, *experiences*, we look at all possible reputations the data provider can have. We look from two broad perspectives: the image and the real experiences. If the image is bad we can face serious complications within our own company (e.g. acceptance problems) and outside (see section 4.1). The most reliable information within this cluster is the internal recommendation. Especially if there has been no collaboration in the past, we must ask for external references as a second criterion. Market experiences is the number of years, for which the provider has offered this kind of information sources. Typically, market and external experience are correlated highly. Nevertheless, we can get hints on how much internal knowledge about the relevant topics the data provider has already collected.

*International competence* is especially important if we intend to use the data for direct marketing projects in various countries. But also if this is not actually planned, good international

competence could influence the providers ability to resolve domestic problems through the knowledge built elsewhere. In addition, we might do future business abroad and therefore check the possibilities.

*Although the collaboration with an international data provider seems promising at first, in practice we learned otherwise. Typically, there are remarkable differences concerning legal issues between the varying countries. Another problem is that even the same provider offers completely dissimilar data within different borders. This is, e.g., due to the data sources he can legally access, the various ways the basic data was collected or the differences in his own company development. For these reasons we cannot transfer marketing or data mining concepts easily. We experienced that differing data sources from different providers normally present the most appropriate solution for cross border projects.*

The *offering portfolio* of the provider is closely related to the business goals of our project and must be compared with the internal requirements. The examples of sub-criteria, as shown in the figure above, are linked to our direct marketing projects. Which ones are picked and how they are measured depends on the project's specifics. Here we have a high need for adaptation.

Now we leave the core enterprise criteria and take a broader view (dotted line). *Pricing* is often meant to be a very important criterion. But we learned that the price is only considered if two or more providers are very similar within other criteria. When including this criterion, not only the costs for data, but all process costs should be taken into account. These are, e.g., costs for preparing the data and in marketing for adding personnel addresses to the keys (often done by the provider).

Another criterion we suggest asking for is the *USP* of a provider. Most providers offer one or more services or data sources exclusively. We check how they fit into our project and gather know how that might be used in future projects or give hints for new marketing possibilities and approaches.

In case of short operational projects, offerings and prices must be checked especially. If a long-term partnership is planned, company facts as well as experiences and international competence play a bigger role. In case an enterprise just started to offer these products, it is uncertain whether it will still exist in two or three years. Then data from providers with a higher market experience are preferable.

We can say that enterprise related criteria (compared to the other dimensions) are usually quickly to obtain but difficult to measure. They also act as KO-criteria very often and are used during coarse selection (see 4.2).

Most of the information can be gathered through interviews with the provider (we recommend inviting them for presentations). Other valuable sources are companies called as references and public sources like journals, corporate reports and so on.

### 4.3.2 The Service Dimension

The service dimension reflects the quality of the collaboration with the data provider independent of its products and its company characteristics. We focus on the service level. The role of service for a successful relationship is often neglected. When the data are not delivered in time and in the quality agreed upon, this leads to additional costs, delays, and incorrect results. Fig. 5 illustrates criteria that may be used for catching the service ability of the data providers.

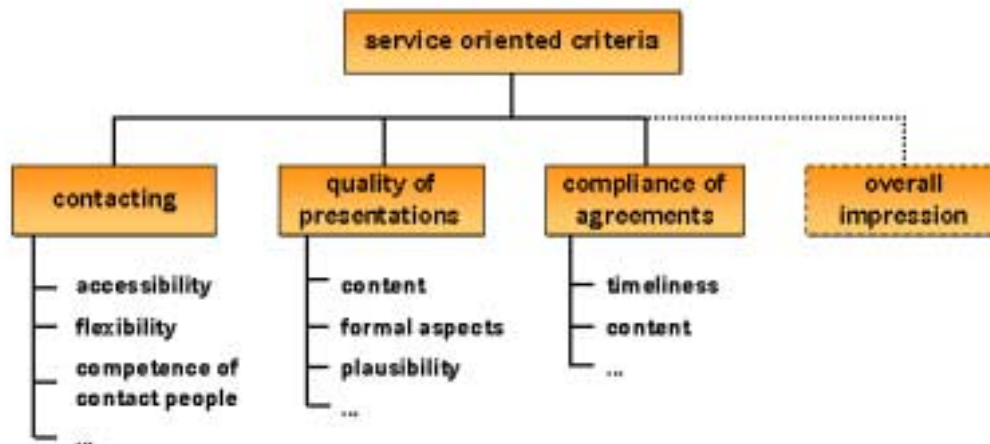


Fig. 5. Examples for criteria in service dimension

The cluster *contacting* contains criteria that are directly related to personal contacts with the provider. We suggest measuring *the accessibility, flexibility and competence* of the contact people. So we obtain good indicators on how the provider can react to special needs and time restrictions or if he understands our problem and really wants to resolve it. Sometimes providers just want to sell standardized products that are hardly possible for direct marketing applications. For a good evaluation we advise using uniform questionnaires with multiple choice questions (interval scales), which should be filled out by as many people as possible.

Another good idea to measure the service level of a potential provider is to invite him for a *presentation*. If we specify our expectations to him in advance, we can gather valuable information. Because even if the expected content has been clarified previously, the quality of the presentations often varies considerably. This group of criteria is strongly related to the contacting cluster but creates additional insights. Again we recommend using questionnaires and a team of interviewers.

We assume *compliance of agreements* to be a separate category because of its importance for achieving the project goals. Especially when test data are examined *timeliness* and *content* of the delivery can be tested. If no test data are used, this criterion cannot be gathered according to its importance.

The last category (criterion) we suggest gathering is the *overall impression* of the provider. Similar to customer satisfaction inquiries we can use this as separate overall criterion.



During our projects at DaimlerChrysler we learned that the service dimension is very important to practitioners and therefore gets high weights always. When we gather the information we have two main sources: information that we obtain directly from contacting the providers during the project and information we can collect from prior work (if there were any). For the first source we should make sure that the contacts (presentations) are within a short timeframe and that we have only one team of interviewers.

### 4.3.3 The Data Dimension

Since we are up to the selection of external data sources, the data dimension is the core of our evaluation system and for that reason most important. As mentioned before, it is also most difficult, complex, and time consuming. Here, we also talk about the dimension that is most closely related to the business issue itself.

For the measurement of this dimension we need to work with test data. We distinguish between specified and non specified test data. If we use specified data (in direct marketing), we send a sample of our own addresses to the providers and ask them to enrich the file with their data (attributes). If using unspecified test data, we just ask for a sample from the providers database(s). Of course, the latter method is less valuable.

Fig. 6 shows that there are two main categories for evaluating the data dimension. The first category – *data* - deals with the product itself. The second category summarizes the quality of the *documentation*. Both are decomposed on the first level in *substance* and *format*, each consisting of various criteria.

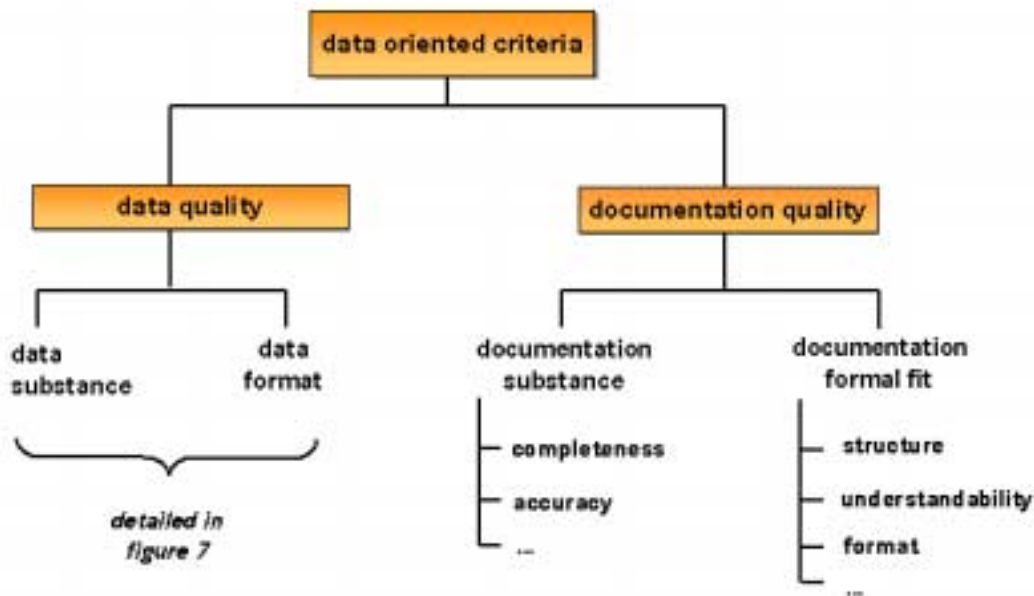


Fig. 6. Examples for criteria in data dimension

We first want to discuss the *documentation criteria*. In our projects we realized that an incomplete documentation may make good data worthless because their potential cannot be identified and exploited. Moreover, lacking documentation necessitates more communication with the provider and faults become more probable. Generally, we distinguish documentation substance and formal fit. Within the *substance* we have to check if all demanded information is delivered and exists in an appropriate quality. Within *formal fit*, we suggest investigating structure, format, and understandability.

Most documentation criteria are qualitative in their nature. But scales can be applied which enable us to capture whether the respective criterion is completely, mostly, partly or not fulfilled. This evaluation has to be done by the people working with the data.

*In practice we learned that there really is a huge difference in documentation quality. The differences can occur concerning all criteria listed above. We had to face things like wrong language, wrong descriptions or inaccessible documentation at all. These difficulties ended mostly in extended data understanding and preparation phases [2].*

Now we want to look at the second category, data quality. Fig. 7 shows the sub-categories and related criteria in detail.

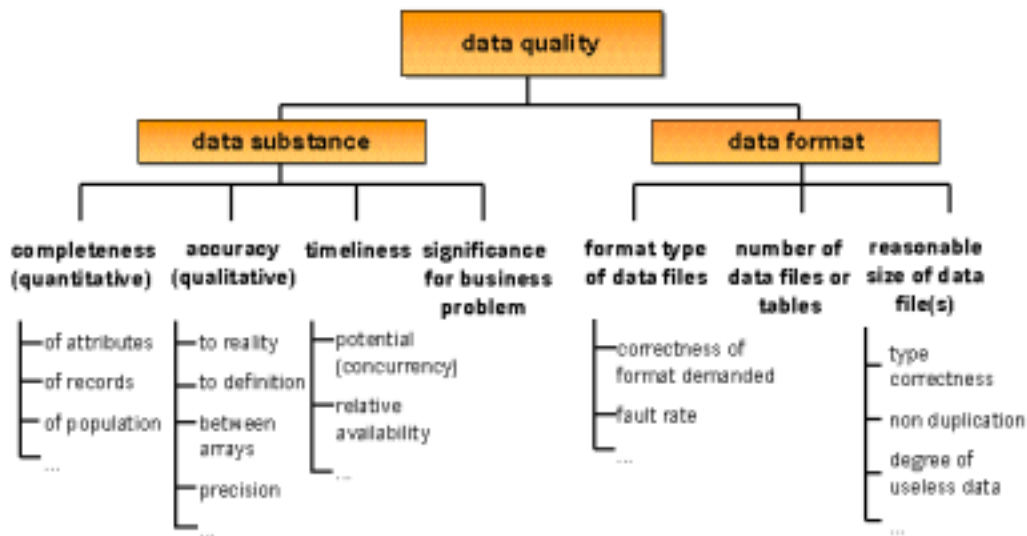


Fig. 7. Examples for data criteria

Completeness, accuracy, and timeliness are common criteria for *data substance* [10]. They can only be measured through analyzing the test data. Fig. 7 shows that *completeness* can be related e.g. to attributes, records and population. In the marketing context the latter means how the data source covers the target market (in percent of the overall population) [1].

*Accuracy* can be related to different referring points as well (e.g. to reality, to definition, etc.) [7]. Accuracy between fields measures how the data corresponds within the same record. We distinguish between hard and soft correspondence. Hard correspondence means a fixed

relationship (e.g. between zip code and area code). A soft relationship refers to information that is possible, but most likely to be untrue (e.g. age: 18, monthly income: 120,000 USD).

*Timeliness* is viewed from diverse perspectives as well [15]. We can measure when the delivered data has been collected the last time (*backward* measurement) and when the data will be updated the next time (*forward* measurement). We can also look at the turn in which it is usually collected. Note that there is not just one timeliness within a data source. Some fields (attributes) might be updated differently within the same source or there might be a time lag between distributed data sources (measured e.g. through concurrency) [7].

Completeness, accuracy and actuality can be collected directly. Completeness is available as proportion of 100%. Accuracy is more difficult to gather. One possibility consists in defining levels of accuracy from very accurate to hardly accurate. Typically, actuality is measured on a ratio scale. In case of the usual update turns we use an interval scale. For aggregating, all criteria scale transformations might be necessary.

*Before explaining the next criterion we would like to take the opportunity to talk about the usefulness of specified test data in order to measure the data oriented criteria mentioned above. First of all, we can improve the reliability of all criteria, since we have control over the records to be enriched (the provider cannot just send his “best” data). Second we can check correctness and actuality for addresses where the true value is known. And last but not least, we can check how the provider handles dirty data. E.g. for direct marketing purposes it is highly interesting to see how the provider works with incorrect addresses.*

*Significance for business relevance* is the most important criterion within data substance and even within the data dimension. Actual and accurate data are useless if they don't contain information meeting the business goals. Unfortunately, this criterion is very difficult to obtain. On the one hand, this is due to the fact that not all details of the project are known in advance (see section 2). On the other hand, the internal data situation and the project goals can change in a way that influences the business relevance of the data. The third problem is that even if we know the business and data mining goals in advance, we often do lack a proper evaluation system [8].

*In practice, we tried to simulate the real situation. E.g., we generated test models (using the real target variable) for predicting potential customers. But this procedure was very time consuming. Because of the variety of available data, it did not show satisfying results by now. When dealing with many and large data sources, we face the problem of attribute selection as mentioned in section 2.*

*Data format* is the second sub-category to examine within the data dimension and its importance is underestimated frequently. First, we check if the *format type* of the data files corresponds to our specified demands. This will help to process the data later.

The second criterion – the *number of data files or tables* – must be specified for each project individually. Usually, we like the provider to process the data as far as possible. Because, the more tables or files are delivered, the more work we must undertake to join them. Sometimes,

the different files even correspond to different aggregation levels, and matching the files is very laborious. But there may be reasons to prefer in-house processing.

The *reasonable file size* as third criterion is important for handling the data. The bigger the file, the slower is the data processing and the more resources are necessary. Examples for common mistakes concerning the file size are doubled information, wrong data type specifications or fields containing useless data (e.g. keys from former matching).

## **4.4 Summary and Decision**

The first step of this phase is to make the final decision about the data sources (providers). For that we use the outcome of the prior phase, the final evaluation portfolio. As we have seen in 4.3, the portfolio seldom shows an absolutely clear favorite. More often we have the choice between several alternatives. Therefore, it becomes clear, that often after close-up examination we *do not have a final solution but we reduce the possibilities*.

In case of very close positioned data sources (even after the recheck in close-up examination), we can attempt to *weight the relative importance of the evaluation dimensions*. Another possibility is to *combine the data sources* or to *use them successively*. For making the final decision we recommend organizing a last workshop with all (ore most) persons which have been involved. This way we can overcome the subjectivity mentioned in section 2.

The project ends with the production of the final report where all the threads are brought together. For that, we carefully review the project. Besides the results obtained, the report should also describe the process, define the deviations from the original plan and note the assumptions and uncertainties. Consequently, it is a summary of all experiences, but should also make recommendations for future work.

*Within our department at DaimlerChrysler we gathered a collection of 5 such reports. They are all structured similarly (since we improve the structure continuously). We use this document to collect different experiences and to spread them widely within the company. We also see promising developments in cost and time efficiency.*

## **5 Conclusion**

### **5.1 Summary**

We presented a practical approach for the selection of external data sources. It was not possible to generate a detailed process model but to provide a framework with manageable tools. However, the approach needs correct adaptation for each project. Hence, the amount of time and resources we have to put in for the selection varies extremely.

Clearly we can not overcome all problems mentioned in section 2 but we have shown feasible solutions and ways to alleviate the effects. E.g. we cannot generate objective decisions but replace subjectivity by “inter subjectivity”.

Moreover, we showed that we also have to look at enterprise and service criteria not at data quality alone. Additionally, we learned that there is no real data (information) quality without proper documentation. If we cannot understand and access the data rightly there is little information to gain.

Finally, we state that in most cases there is no one and only final solution. Because of the problems pertaining to collecting, measuring and comparing the selection criteria we just reduce the possibilities. But from experience we know that the outcome is normally worth the effort.

## **5.2 Further Research/ Open Issues**

As we tried to illustrate in this paper, it is hard to standardize the selection process. The suggestions made by us help to make the process repeatable and give hints on how to overcome handling issues. Further development could broaden the framework for organizing the selection criteria, and advance the process model or the standardization of tools like, e.g., questionnaires, check lists or rating scales.

For selection processes which last over a long time (e.g. 10 months) we must consider how we handle the issue that the information obtained about data sources and providers can change dramatically during this time.

From our point of view, the area of information processing leaves the most room for improvement. We need more advanced approaches properly integrating all steps (collection, measurement, aggregation and comparison of data (information)) in order to aggregate the information to a high level without losing too much detail.

## **6 References**

1. Arndt, D.; Gersten, W.: Data Management in Analytical Customer Relationship Management. In: Workshop Data Mining for Marketing Applications, In: Proceedings of the ECML/PKDD 2001. Springer, Heidelberg (2001) (to appear)
2. Arndt, D., Gersten, W., Wirth, R.: Kundenprofile zur Prognose der Markenaffinität im Automobilsektor. In: Hippner, H., Küsters, U., Meyer, M., Wilde, K. (eds.): Handbuch Data Mining im Marketing. Vieweg, Braunschweig Wiesbaden (2001) 591-606
3. Berry, M.J.A., Linoff, G.S.: Mastering Data Mining. Wiley, New York (2000)
4. Berthold, M.; Hand, D.J.: Intelligent Data Analysis. Springer, Heidelberg (1999)
5. Chapman, P. et al.: CRISP-DM 1.0. SPSS Inc., München (2000)
6. CRISP-DM: Cross-Industry Standard Process Modell for Data Mining. In: <http://www.crisp-dm.org/home.html> (2001)
7. English, L.P.: Improving Data Warehouse and Business Information Quality. Wiley, New York (1999)

8. Gersten, W., Wirth, R., Arndt, D.: Predictive Modeling in Automotive Direct Marketing: Tools, Experiences and Open Issues. In: Proceedings of the 6<sup>th</sup> International Conference on Knowledge Discovery and Data Mining. ACM, New York (2000) 398-406
9. Heinrich, L.P.: Informationsmanagement: Planung, Überwachung und Steuerung der Informationsinfrastruktur. München, Wien (1999)
10. Hipp, J., Günzer, U., Grimmer, U.: Data Quality Mining – Making a Virtue of Necessity. In: Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD 2001), In: Proceedings of the 6<sup>th</sup> ACM SIGMOD. Santa Babara, CA (2001) 52-57
11. Pyle, D.: Data preparation for data mining. Morgan Kaufmann Publishers, San Francisco (1999)
12. Schul, F.: Neue Konzepte des strategischen Portfolio-Managements im diversifizierten Unternehmen. Stuttgart (1981)
13. Schwarze, J.: Grundlagen der Statistik I. Verlag Neue Wirtschafts-Briefe, Berlin (1998)
14. Shepard, D. et al.: New direct marketing. McGraw-Hill, New York (1991)
15. Strong, D., Lee, Y., Wang, R.: Data Quality in Context, Communications of the ACM, Vol. 40 (5), 1997

# A Strategy for Managing Data Quality in Data Warehouse Systems

— Work in Progress —

Markus Helfert, Eitel von Maur  
Institute of Information Management, University of St. Gallen  
Mueller-Friedberg-Strasse 8, St. Gallen, CH-9000, Switzerland  
phone: +41-71-224-33 82 fax: +41-71-224-21 89  
<http://datawarehouse.iwi.unisg.ch>  
{Markus.Helfert | Eitel.vonMaur}@unisg.ch

## ABSTRACT

High level data quality and the management of ensuring data quality is one of the key success factors for Data Warehousing projects. The following article describes an approach for Data Quality Management, which is based on theories as well as practical experiences. Starting from effects of insufficient data quality in practice, a definition for information, data and data quality will be worked out. Based on the concept of total data quality management the Data Quality Management (DQM) for Data-Warehouse-System will be described. As key part DQM an approach for operative DQM (planing and measuring data quality) will be illustrated and explained. Finally, based on the research results further conclusions are summarised.

**Keywords:** Data Warehousing, Data Quality, Data Quality Management, Data Quality Measuring, Information Systems

## 1. INTRODUCTION

Data warehousing has captured the attention of practitioners and researchers for a long time, whereas aspects of data quality is one of the crucial issues in data warehousing. (English 1999; Helfert 2000a) In general, ensuring high level data quality is one of the most expensive and time-consuming tasks to perform in data warehousing projects (Mueller 2000; Haeussler 1998). Because of insufficient data quality, data warehouse projects are frequently discontinued (Helfert 2000b).

In several publications (Jarke et al. 2000; Helfert 2000b; Huang et al. 1999; English 1999; Tayi/Ballou 1998) approaches for managing data quality have been suggested, but the question of how to ensure high level data quality in data warehouse systems still remains. A key question in data quality management is the operative tasks of quality planing, specifying processes and measuring them on an integrated and most objective base. An evaluation of major approaches shows that there is still a lack of gathering data quality requirements, quality planing and transforming these requirements into a specification and controlling them. One approach, which was developed at the MIT, models quality requirements based on Entity-Relationship-Models.(Wang et al. 1993) Another approach, which was developed within an European research project, considers technical aspects of integrating quality requirements into meta data management.(Jarke et

al. 2000) Most approaches are based on data quality criteria lists and are not linked to measurement systems with quality indicators. They also lack guidelines and methods for applying them to company specific requirements. However, there is still no adequate model for integrating operative data quality management into data warehouse systems.

## **2. DATA-WAREHOUSE-SYSTEMS**

In today's competitive and global business environment, understanding and managing information is crucial for companies in order for them to make timely decisions and to respond to changing business conditions. Data processing applications have proliferated across a wide variety of systems over the last two decades, complicating the task of locating and integrating data within the enterprise wide information system. They are often developed separately resulting in different data models, data descriptions and interpretations. As a result, to manage and use the data, many organisations today are building data warehouse systems. A data warehouse system supports information processes by creating an integrated database of consistent, enterprise wide and historical data. It integrates data from multiple, incompatible systems into one consolidated database.

Central component of a data warehouse system is the data warehouse data base, which is simply a single, complete, and consistent store of data obtained from a variety of sources. (Devlin 1997) The data base is used by a number of tools and applications which form the data warehouse system. (Winter 2000) Figure 1 gives an overview of the main components of the data warehouse system, which are to be described in the following briefly.

Starting points are the operational systems and external information systems, which act as data suppliers. With the help of a transformation component the data is extracted, transformed and transferred into the central data warehouse data base. Partly, for subject oriented data supply, smaller, redundant views from the enterprise database are stored in so-called data marts. The databases are accessed by the users with a number of end-user tools. These tools reach from creation of reports over ad-hoc-queries up to multi-layered, multidimensional analyses and data mining.

These components form the basis-system of the data warehouse system by providing the data flow from the data source up to the data use. Alongside this, a co-ordination system, generally named meta data management system, exists. (Holthuis 1999) It consists of a meta data base and tools for storage and administrating the system components and data flows.

## **3. DATA QUALITY IN DATA-WAREHOUSE-SYSTEMS**

Discussion in literature about information, data and data quality shows that these terms are complex and still no widely accepted definition exists. There are numerous approaches for defining information (Bode 1997), quality (Juran 1998) and data quality (respectively information quality) (Huang et al 1999; Tayi/Ballou 1998; Wand/Wang 1996) and therefore it is necessary to clarify these terms. Many approaches do not distinguish between data and information and define data quality and information quality equally. Before defining data quality, in the following a suitable



definition for knowledge, information and data will be introduced and the different views on quality are described.(Bode 1997; Helfert 2001)

### **3.1. Knowledge, Information and Data**

**Knowledge** is any form of representation of parts of the real or conceptual world in a material media.(Bode 1997) Characteristic of this definition is the representation of real world objects. Knowledge is an image and is not identical with the real world. But however it is related to the real world and thus has some meaning (semantics). Based on this definition, **information** can be understood as a subset of knowledge, which can be expressed in some form of human language.(Bode 1997) Human language is limited to languages for communicating between humans. Following this definition, **data** can be defined as a subset of information, which is oriented to be processed by machines (e. g. applications and data base systems) .(Bode 1997)

### **3.2. Quality**

The term quality is as complex as the term information.(Juran 1998) As a consequence of the discussion on this term, a variety of definition and interpretation approaches exists. The aim of the definition is to reduce the complexity of the quality phenomenon and to attain operational statements.

According to the classification of (Garvin 84) quality approaches can be differentiated into five quality definitions. The **transcendent view** defines quality as a synonym with “innate excellence” or superlative, as a synonym for high standards and requirements. This, rather abstractly philosophical understanding that quality cannot be precisely defined is insufficient for further work in the context of this thesis. **Product-based** definitions are quite different; they view quality as a precise and measurable variable. Quality is so precisely measurable through inherent characteristic of the product. **User-based** approaches start from the opposite premise that quality is stated by the user. Individual consumers are assumed to have different wants, and those products that best satisfy their preferences are those that they regard as having the highest quality. This is a idiosyncratic and personal view of quality, and one that is highly subjective. In contrast to this subjective view, **manufacturing-based** definitions focus on the supply side and are primarily concerned with the production processes. All manufacturing-based definitions virtually identify quality as conformance to requirements. Once a design or a specification has been defined, any deviation implies a reduction in quality. **Value-based** definitions consider terms of costs and prices. According to this view, a quality product is one that provides performance at an acceptable price or conformance at an acceptable cost.

It is important to note, that all these different approaches (apart from the transcendent view) are important on different levels of the design process. It is not possible to focus only on one perspective. The different approaches represent the levels of requirement analysis, product and process design and the actual manufacturing process.

### **3.3. Data Quality**

Like the terms above, data and information quality are described in literature through many different views, whereas the user-oriented view dominates. There are many different definitions with a vast quantity of quality criteria. (Wang et al. 2001; Mueller 2000; Naumann/Rolker 2000; Wolf 1999; English 1999; Tayi/Ballou 1998; Jarke/Vassiliou 1997; Wang/Strong 1996; Redman 1996) Result of this work is a multiplicity of criterion lists and classification frameworks for different areas. In the context of the thesis and as basis for developing a data quality model, these approaches are to be examined and classified.

Although the terms data and information quality are used without uniformity, the analysis of the approaches shows a conformance that data quality is determined according to a user-oriented view. This view is concretised through quality criteria, which depend in its meaning and intensity on the application. The approaches do not show any conformance regarding the quality criteria lists, their definitions and systematic. Generally the quality criteria are created intuitively on the basis of experiences (Ballou/Pazer 1985; Laudon 1986; Morey 1982), literature (Naumann/Rolker 2000) or by empirical research (Wang/Strong 1996).

Basis of further work is often the quality criteria list from (Wang/Strong 1996). On the basis of these criteria Jarke and Vassiliou suggest a model for quality factors in data warehouse systems. (Jarke/Vassiliou 1997) Main differences to the initial model lie in the greater emphasis on historical as well as aggregated data. Figure 2 illustrates the hierarchy of quality factors used. After describing the term data quality and list and some relevant quality criteria, the concept of data quality management will be described in the next section.

## **4. DATA QUALITY MANAGEMENT**

Quality management includes concepts of quality policy, quality strategy, quality planning, quality control and quality assurance as well as quality improvement. (Juran 1979; Deming 1982; Seghezzi 1996) One widely accepted concept for quality management is the concept of total quality management (TQM). The concept states the current research in quality management and has already been successfully implemented in the manufacturing sector. Currently the concept of TQM is applied to sectors like the service industry and data quality. (English 1999; Redman 1996; Wolf 1999) Typical for TQM is the orientation on customer requirements, the participation of people, continuous improvement and the comprehensive management approach. All enterprise wide activities are integrated into an enterprise wide structure aiming continuously to improve products, services and process quality and therefore satisfying customer requirements. (Seghezzi 1996) Following the total quality management approach a concept for data quality management for data warehouse systems can be proposed. Three aspects are fundamental to this concept (see also Figure 3) (Helfert/Radon 2000; Wolf 1999; Huang et al. 1999; English 1999):

- Management has to commit to accept the philosophy of high level data quality and show this commitment in each activity. On the basis of data quality principles and goals a data quality policy and strategy have to be deduced.
- A quality management system with organisational structure, process organisation, standards and specifications, guidelines and rules as its basic elements founding the structure for the

management concept. Frequently inspections ensure continuous improvement of the organisational structure, processes and standards.

- Employees are supported to fulfill the quality processes by adequate methods, techniques and tools.

The operative level of data quality management deals with four main tasks: *Quality planing* gathers requirements and expectations of data consumers, then transfers these requirements into data delivery processes and specifications.( Juran/Gryna 1993; Seghezzi 1996) Therefore quality criteria have to be selected, classified and prioritised.(Wallmueller 1990) *Quality control* controls the data delivery processes and complies the stated specification. Therefore adequate steps have to be identified and implemented. To reach this, product and process quality must be measured and expressed in quantitative indices. Most important techniques of quality control are quality examinations.(Juran/Gryna 1993; Wallmueller 1990) *Quality assurance* aims to detect systematic risks in order to avoid them. *Quality improvement* as the forth task, supports continuous and dynamic quality improvement.

Before analysing and improving data quality it is essential to plan, define and assess quality goals and measure current quality levels (Quality planing and Quality control). Data quality planning and controlling are therefore key success factors of the data quality management concept. This gives the possibility to state current quality levels and compare the results in time. It is possible to identify quality trends and evaluate the effects of quality improvements, which provide the foundation for cost benefit analysis.

#### **4.1. Operative Data Quality Management: Quality planing and measuring**

In light of the importance of data quality planning and quality control the thesis will be focused on these two quality management areas. Part of this is a quality model to define quality goals and measure current data quality levels. Therefore, the major goal of the thesis is to develop a suitable data quality model for specification and measurement of data quality in data warehouse systems. This enables data quality planing and data quality control. Figure 4 shows the main structure and focus.

Goal of the quality model is to provide a way to specify quality requirements, create a system for evaluating quality specifications and to measure the resulting data quality. On the basis of quality requirements, which depend on user groups as well as the tasks of the data warehouse system, a framework for a specification have to be developed. A substantial aspect of the data quality model is the decomposition of quality criteria. The general quality term, which is characterised by quality criteria, is decomposed into process and product characteristics. These characteristics are measured by quality indices (defined as quality indicators). Measuring techniques as well as suitable measuring points and times have to be determined.

Starting point for the identification of suitable quality indices is the data warehouse basis-system. As a first step the data delivery processes have to be identified from the data occurrence through the operative system to the data usage. Secondly, appropriate data quality indicators have to be assigned to each data delivery process and data set. Task is to identify typical relations in the data sets and typical characteristics of transformation processes within the basis-system. With the

help of statistics dynamic modifications and inconsistencies in data sets and transformation processes can be detected. The following paragraph shows a simple example of this approach, where research is still to be done.

#### **4.2. A case for the Data Quality Model**

To develop a realistic scenario for the proposed data quality model, I worked together with eight large enterprises and we selected an example from an insurance company.

**Analytic question: “Number of contracts per region and per sales representative from the perspective of the controlling department. “**

This — on a first glance — seemingly simple question turned out to be highly sophisticated and complex to provide through a data warehouse system. First of all, we discussed problems and common data quality requirements, which lead us to data quality measuring approaches.

The main data quality difficulties are different interpretations and multiple applications for this information in different contexts. For example different departments define the term “contract” and the relevant transaction date differently. Problems such as movement of sales representative from one region to another during an accounting period, causes associating difficulties. Through the discussion it turned out that data quality levels are highly dependent on user groups and their intended tasks. Expressed data quality requirements and their related data quality criteria are summarised in the following table (see Table 1):

In a second step we worked out a typical data flow, the data transfer and the transformation processes, which are all shown in Figure 5. First of all, during sales conversations, sales representatives gather contract data and customer information. Regularly this data is synchronised with operative systems (e.g. Mainframe systems). Data typists manually enter non electronic data into the system (e.g. Contracts which could not be entered into the sales representatives’ system for some reason). In a further step the final contract is sent to the customer. The customer can accept, change or even cancel the contract (within a given time). Frequently some ETL-Component (Extraction, Transforming and Loading) extracts the new or updated data (“delta data”) from the operative systems and transfers it into a staging area. In this temporary data base, quality verifications and improvements are performed. Inconsistencies between new data and data, which is already stored in the central data warehouse data base, could be identified through a link between contract data and sales representatives data already in the data base. A separate data cleansing process handles these quality deficiencies and may possibly improve them. Data which has passed the quality verifications and improvements is then loaded into the data warehouse data base (mostly without any further integrity checks). In a last step, the data is then provided through front-end-tools and subject oriented data marts to data users.

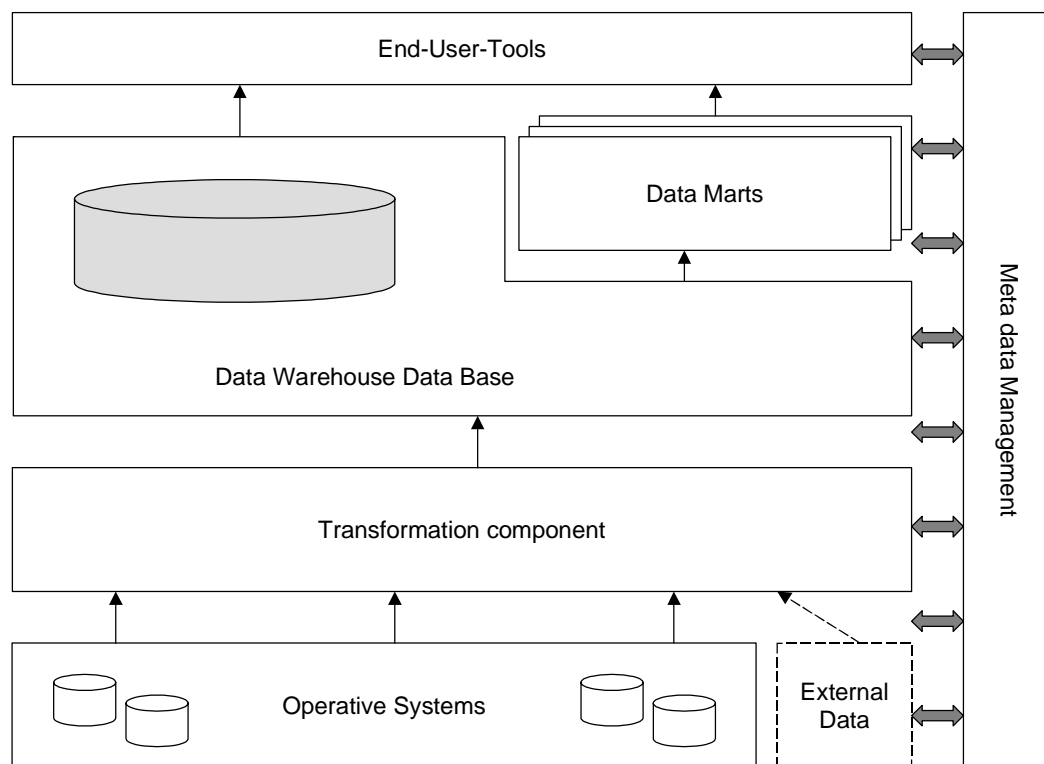
In a next step, we structured the data flow and expressed it in a formal way according to Figure 6, which represents the corresponding model. At the top the relations between the conceptual model, the data model and the physical data store is presented. The left part represents operations on the data store, like insert, update, delete, selection and projection. Usually these operations are performed by SQL and handled by some data base system. Transformations are functions performed on data values, which are elements of data sets. A transformation takes one or more data

sets. The output of transformations are one or more data sets. For example relevant data transformations are gathering data, conversion data values, enrichment and aggregation. (Devlin 1997) One particular transformation is data validation, which takes data values and separates these into two data sets (accepted and not accepted data values). The main control element of the data flow is the session, which runs queries and transformations.

Formalised the data flow in Table 2, we then identified quality indicators (process and product) for measuring data quality. These are shown in Table 3.

## 5. CONCLUSION

The research so far shows that data quality in data warehouse systems is a crucial issue but also a highly sophisticated task to fulfil. On base of the proposed data quality model it is assumed that there are user group and task dependent quality requirements as well as quality indicators along the data delivery process. These requirements are usually expressed in natural language by end users. To structure and express these user requirements a conceptual modelling language has to be developed and integrated in the conventional data modelling process for data warehouse systems. In further research the quality requirements have to be linked to data quality criteria. These quality indices and their target values have then be linked to the data delivery processes. One more technical aspect is to be integrated into the quality planing and controlling into the conventional meta data management. The data warehouse basis-system could so be controlled and would lead to a controlled and higher data quality in data warehouse systems.



**Figure 1: Data Warehouse System (Mueller 2000)**

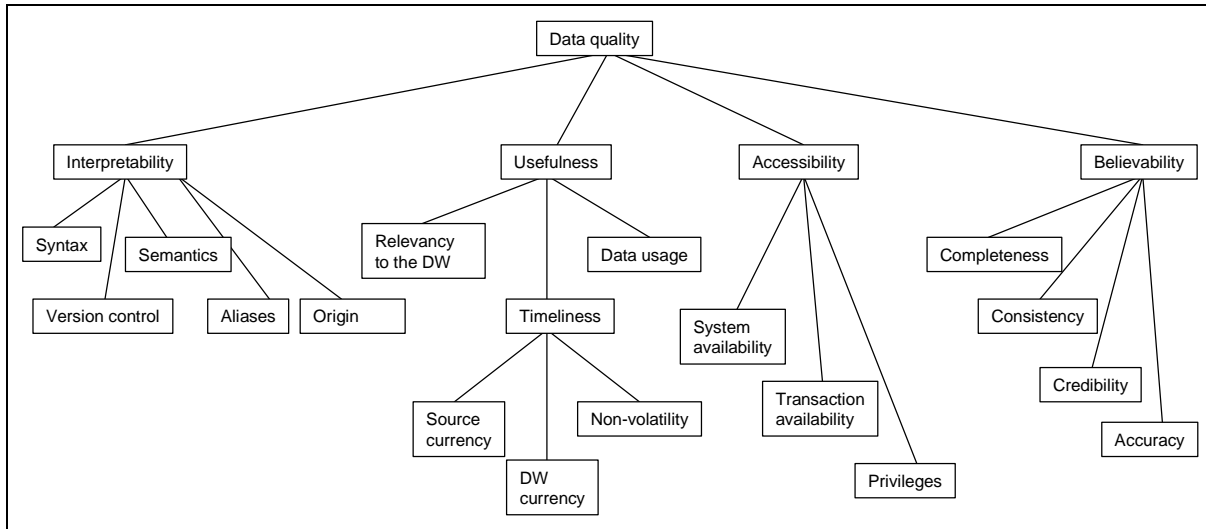


Figure 2: Quality factors for data warehouse systems (Jarke/Vassiliou 1997)

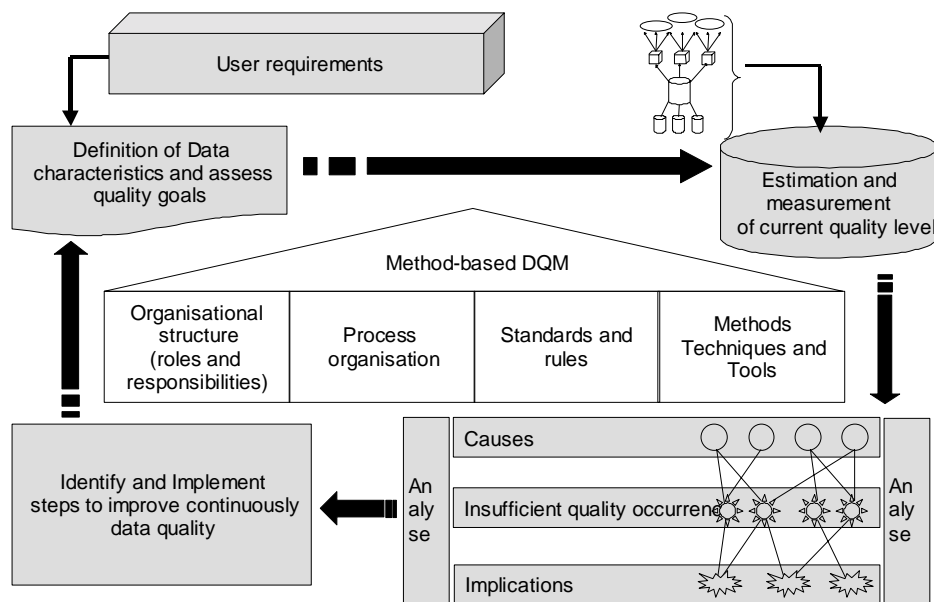


Figure 3: Data Quality Management (Helfert/Radon 2000)

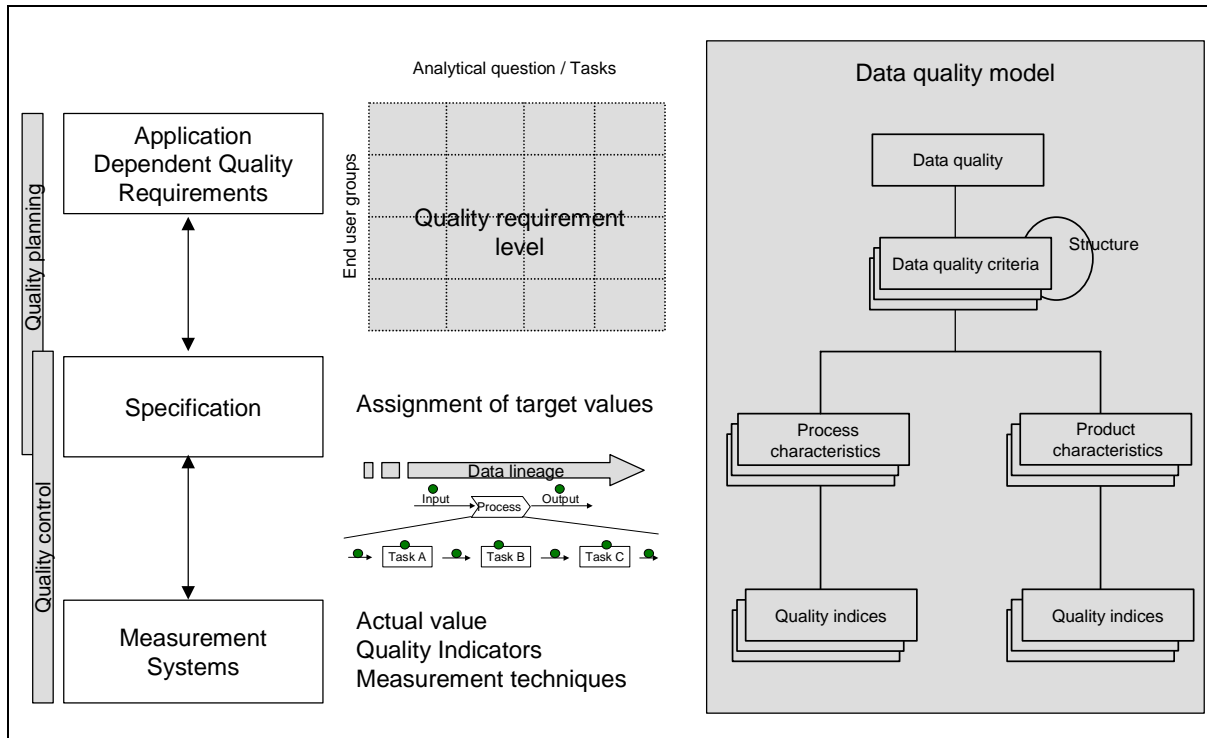


Figure 4: Structure and Focus of the Data Quality Framework

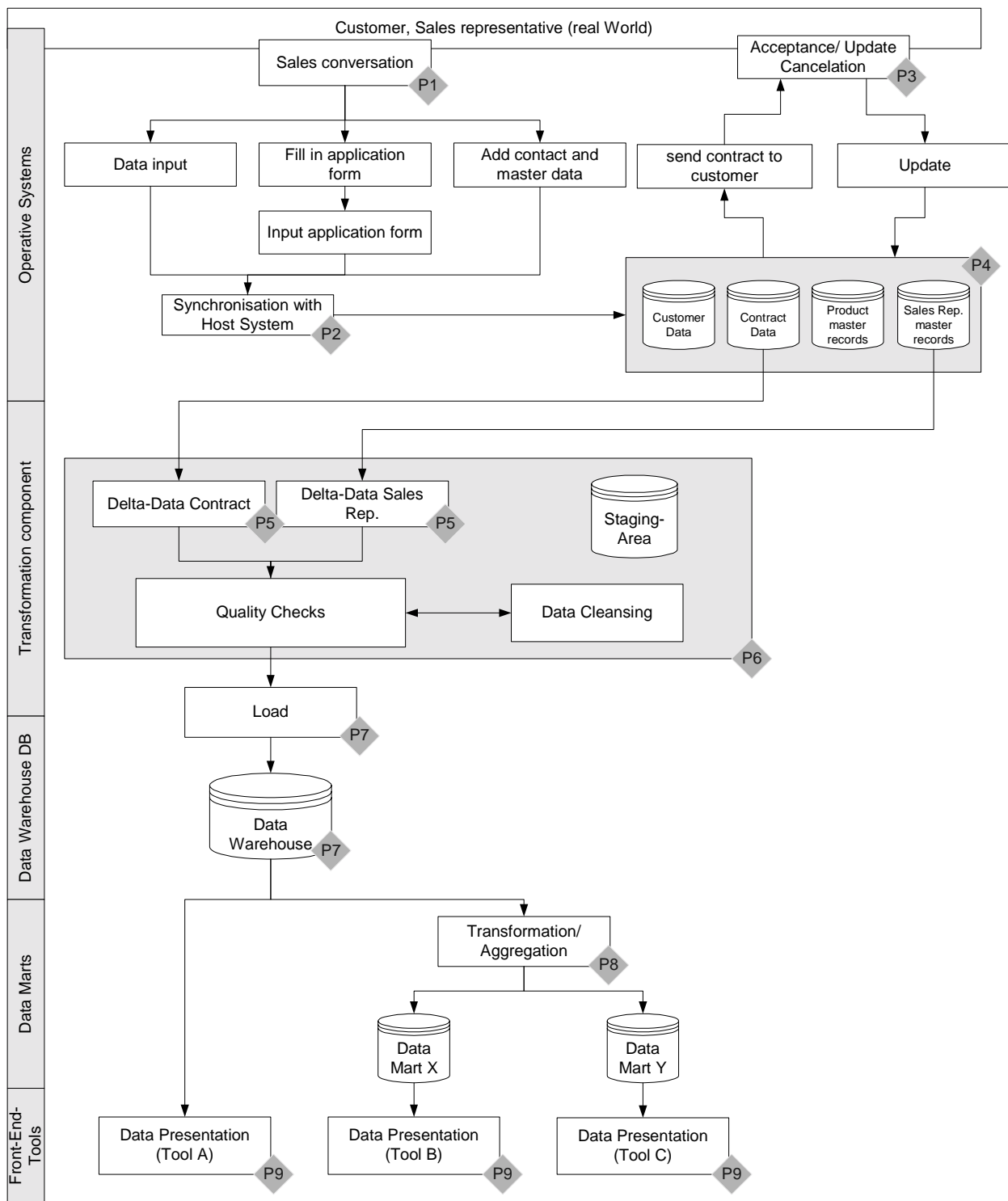
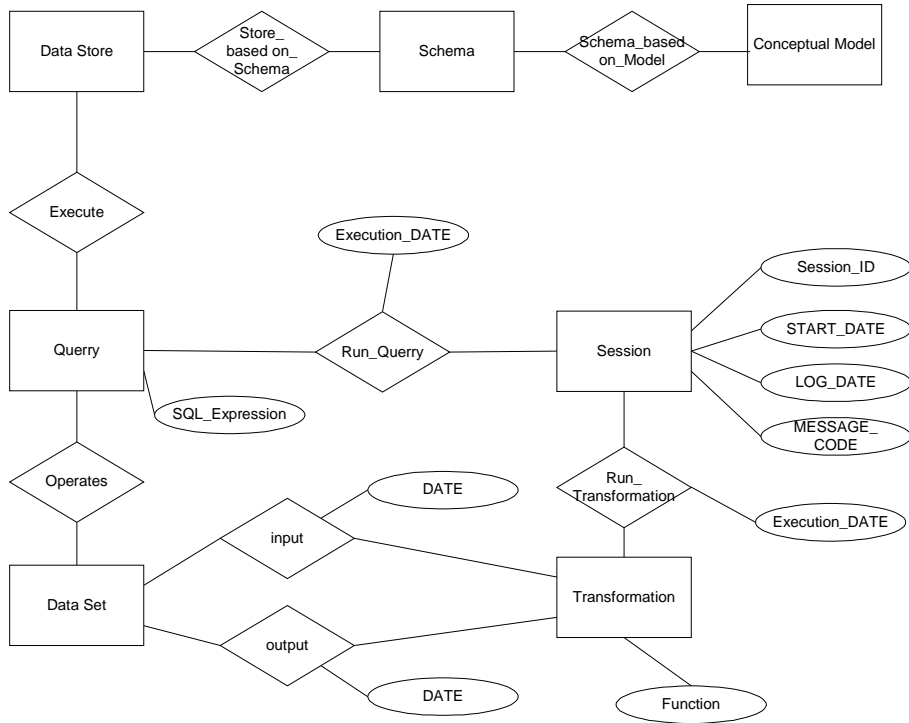


Figure 5: Typical data delivery processes and quality indicators





**Figure 6: Meta model for Data Flow**

Expressed Data Quality Requirements	Data Quality Criteria
Maximum allowed divergence to real number of contracts is +/- 2 %.	Accuracy
All sales representatives and regions have to be listed.	Completeness
Format for date has to be “dd/mm/yy”.	Interpretability, Format, Syntax
Information is updated monthly	Timeliness
On the second working day of each month a trend has to be identified (Accuracy +/- 5 of real value).	Accuracy, Timeliness
On the fifth working day of each month the final information has to be provided (In case of later changes reasons have to be given).	Accuracy, Timeliness
Information of the ten best and worst sales representatives have to be accurate.	Accuracy
Number of contracts from new products should be accurate.	Accuracy
Sum of contracts per region and per sales representative have to correspond with the total number of contracts sold.	Accuracy, Consistency
Responds time should be less than three minutes	Timeliness

Table 1: Data Quality Requirements

Data Flow processes	Labels in Figure 6
Data Gathering through sales conversation	P1
Synchronisation with operative System	P2
Validation of contract information by customer	P3
Data update and storage in operative Data Bases	P4
Data extraction (delta data)	P5
Data validation and transformation (Cleansing)	P6
Data load and storage in Data Warehouse	P7
Data aggregation and transformation in multi-dimensional Data models	P8
Data presentation	P9

Table 2: Data Flow processes

<b>Data Quality criteria</b>	<b>Data Quality indicators and measuring points</b>
Timeliness (currency)	Data gathering date [P1] Last execution time / Scheduled time vs. Execution time / Version control through time stamps (protocol evaluation) [P2, P6, P7,P8]
Completeness	Completeness of optional data values [P1, P4, P6] Numbers of default-values compared to average [P1, P4, P6]
Consistency	Plausibility verifications [P1, P4, P6]
Accuracy	Frequency of changes [P4] Customers' feedback [P3] Data user valuation [P9]
Interpretability	Data user valuation [P9] Domain violation [P1, P4, P6]

**Table 3: Data quality indicators**

## 6. REFERENCES

(Ballou/Pazer 1985) Ballou, D. P.; Pazer, H. L.: Modeling Data Process Quality in Multi-input, Multi-output Information Systems. In: Management Science 31 (1985) 4, pp. 150-162.

(Bode 1997) Bode, J.: Der Informationsbegriff in der Betriebswirtschaftslehre. In: Schmalenbachs Zeitschrift für betriebswirtschaftliche Forschung, 49 (1997) 5, pp. 449-468.

(Deming 1982) Deming, W. E.: Quality, Productivity and Competitive Position. Cambridge, 1982.

(Devlin 1007) Devlin, B.: Data Warehouse: From architecture to implementation. Addison-Wesley Longman, Reading, MA et al. 1997.

(English 1999) English, L.: Improving Data Warehouse and Business Information Quality. Wiley, New York 1999.

(Garvin 1984) Garvin, D. A.: What does "Product Quality" really mean?. In: Sloan Management Review 26 (1984) 1, pp. 25-43.

(Haeussler 1998) Haeussler, C.: Datenqualitaet. In: Martin, W. (ed.): Data Warehousing. ITP, Bonn 1998, pp. 75-89.

(Helfert 2000a) Helfert, M.: Eine empirische Untersuchung von Forschungsfragen beim Data Warehousing aus Sicht der Unternehmenspraxis. Working Paper BE HSG/CC DWS/05, Institute of Information Management, University of St. Gallen 2000.

(Helfert 2000b) Helfert, M.: Massnahmen und Konzepte zur Sicherung der Datenqualitaet. In: Jung, R.; Winter, R. (ed.): *Data Warehousing Strategie – Erfahrungen, Methoden, Visionen –* Springer, Berlin et al. 2000.

(Helfert 2001) Helfert, M.: *Managing and Measuring Data Quality in Data Warehousing*. In: *Proceedings of the World Multiconference on Systemics, Cybernetics and Informatics, Orlando, FL 2001*, pp. 55-65.

(Helfert/Radon 2000) Helfert, M.; Radon, R.: *An Approach for Information Quality measurement in Data Warehousing*. In Klein, B. D., Rossin, D. F. (ed.): *Proceedings of the 2000 Conference on Information Quality*. Massachusetts Institute of Technology, Cambridge, MA 2000, pp. 109-125.

(Holthuis 1999) Holthuis, J.: *Der Aufbau von Data Warehouse-Systemen: Konzeption – Datenmodellierung – Vorgehen*. Dt. Univ.-Verlag / Gabler, Wiesbaden 1999.

(Huang et al. 1999) Huang, J.; Lee Y. W.; Wang R. Y.: *Quality Information and Knowledge*. Prentice Hall, Upper Saddle River, NJ 1999.

(Jarke et al. 2000) Jarke, M.; Lenzerini, M.; Vassiliou, Y.; Vassiliadis, P.: *Fundamentals of data warehouses*. Springer, Berlin et al. 2000.

(Jarke/Vassiliou 1997) Jarke, M.; Vassiliou, Y.: *Foundations of Data Warehouse Quality – A Review of the DWQ-Project*. In: Strong, D. M., Kahn, B. K. (ed.): *Proceedings of the 2<sup>nd</sup> International Conference on Information Quality, Cambridge, MA 1997*, pp. 299-313.

(Juran 1979) Juran, J. M.: *Quality Control Handbood*. 3<sup>rd</sup> ed. New York 1979.

(Juran 1998) Juran, J. M.: *How to think about Quality*. In: Juran, J. M., Godfrey A. B. (ed.): *Juran's quality handbook*, 5<sup>th</sup> ed., McGraw-Hill, New York 1998, pp. 2.1-2.18.

(Juran/Gryna 1993) Juran, J. M.; Gryna, F. M.: *Quality Planing and analysis: from product development through use*, McGraw-Hill, New York 1993.

(Laudon 1986) Laudon, K. C.: *Data quality and due process in large interorganizational record systems*. In: *Communication of the ACM 29 (1986) 1*, pp. 4-11.

(Morey 1982) Morey, R. C.: *Estimating and improving the quality of information in the MIS*. In: *Communication of the ACM 25 (1982) 5*, pp. 337-342.

(Mueller 2000) Mueller, J.: *Transformation operativer Daten zur Nutzung im Data Warehouse*. Dt. Univ.-Verlag / Gabler, Wiesbaden 2000.

(Naumann/Rolker 2000) Naumann, F.; Rolker, C.: *Assessment Methods for Information Quality Criteria*. In: *Proceedings of the 2000 Conference on Information Quality, Cambridge, MA 1999*, pp. 148-162.

(Redman 1996) Redman, T. C.: Data quality for the information age. Artech House, Norwood 1996.

(Seghezzi 1996) Seghezzi, H. D.: Integriertes Qualitätsmanagement: das St. Galler Konzept. Hanser, Munich et al. 1996.

(Tayi/Ballou 1998) Tayi, G. K.; Ballou, D.: Examining Data Quality. In: Communication of the ACM 41 (1998) 2, pp. 54-57.

(Wallmueller 1990) Wallmueller, E.: Software-Qualitaetssicherung in der Praxis. Hanser, Munich et al. 1990.

(Wand/Wang 1996) Wand, Y.; Wang R.: Anchoring Data Quality Dimensions in Ontological Foundations. In: Communications of the ACM 39 (1996) 11, pp. 86-95.

(Wang et al. 1993) Wang, R. Y.; Kon, H. B.; Madnick, S. E.: Data Quality requirements analysis and modeling. In: Proceedings of the 9<sup>th</sup> international conference on data engineering (ICDE), IEEE Computer Society, Vienna 1993, pp. 670-677.

(Wang et al. 2001) Wang, R. Y.; Ziad, M.; Lee, Y. W.: Data Quality. Kluwer Academic Publishers, Boston et al. 2001.

(Wang/Strong 1996) Wang, R. Y.; Strong, D. M.: Beyond Accuracy: What Data Quality Means to Data Consumers. In: Journal of Management Information Systems, 12 (1996) 4, pp. 5-33.

(Winter 2000) Winter, R.: Zur Positionierung und Weiterentwicklung des Data Warehousing in der betrieblichen Applikationsarchitektur. In: Jung, R.; Winter, R. (ed.): Data Warehousing Strategie: Erfahrungen, Methoden, Visionen. Springer, Berlin et al. 2000, pp. 127-139.

(Wolf 1999) Wolf, P.: Konzept eines TQM-basierten Regelkreismodells fuer ein „Information Quality Management“ (IQM). Praxiswissen, Dortmund 1999.

## **Information Products for Remanufacturing: Tracing the Repair of an Aircraft Fuel-Pump**

Yang Lee  
Northeastern University  
y.lee@neu.edu

Thomas Allen  
MIT Sloan School of Management  
Tallen@mit.edu

Richard Wang  
Boston University  
rwang@bu.edu

(Research-in-Progress)

**Abstract:** This paper reports our initial study of the remanufacturing process of an organization, with a special focus on the role and nature of the information involved. As a first step, we trace a key physical aircraft component, an aircraft fuel pump, throughout the process of remanufacture. We apply an information product perspective to guide the tracing of information relevant to all physical parts, products, and work activities involved. Remanufacturing is a complex process, which involves repairing and refurbishing parts and products in the carcass. It has been a common practice for aircraft, railway locomotives, and heavy construction equipment. As landfill becomes a scarce resource, remanufacturing will undoubtedly be extended to other products and industries.

The quality of information is determined by characteristics of how the information integrates physical artifacts and activity process. Conversely, the quality of information can determine the performance quality of remanufacturing. An example of one aspect of this performance is the capacity to predict supply material effectively. The remanufacture process lends itself as a useful and comprehensive setting for studying the complex intricacies involved in role and nature of information. This study can be used to determine what should be required for information products for remanufacturing.

Upon completion of our research, we aim to show a clear picture of the entire remanufacture process of an aircraft fuel pump and the nature and various roles of information involved in the process. We will then be able to determine the required information products for this process. We believe that tracing one specific part through remanufacturing in one organization will pave the way for understanding the information needs in manufacture and remanufacture process.

### **1. Introduction**

The process that we label “remanufacturing” has been a common practice for certain high investment capital products, such as aircraft, railway locomotives, and heavy construction equipment. Remanufacturing is a complex process, which involves repairing and refurbishing parts in the carcass of the product. The process also involves planning activities to supply these parts and products, and coordinating the necessary personnel involved in the process. Remanufacturing is also practiced in the automobile aftermarket, where reconditioned parts and re-machined engines are readily available. We will undoubtedly see remanufacturing extended to other products and industries as the availability of landfill space becomes scarcer.

This initial study attempts to understand the remanufacturing process with a special focus on the role and nature of the information involved. As a first step, we simultaneously trace a key physical aircraft component and the related information throughout the process of remanufacture.

We chose to trace a fuel pump, an essential part of every modern aircraft. The aircraft is not viable unless it has the pump; the pump is not viable unless it has the stator installed. A stator is a system of stationary airfoils in the compressor of an aircraft fuel pump. We studied a particular type of pump that is installed on many military aircraft. We also chose a specific organization, the United States Air Force, in which to trace the process. We refer to the information product perspective [15, 16] to guide tracing information relevant to all physical products, parts, and work activities involved.

Research on remanufacturing has been conducted to find solutions and strategies for operational problems at hand. The research conducted, particularly under the rubric of the Lean Sustainment Initiative at MIT, offers useful perspectives on understanding the complex processes involved and the various operational problems and solutions [4, 9, 12]. We have not found, however, any research that focuses on fundamental treaties of remanufacture, and in particular research that covers the importance of information.

Our particular purpose in mapping out the remanufacture process is to understand the characteristics of any discrepancies between the available information and information needs. We will then be able to determine the required characteristics of information that should be embedded in the information for remanufacturing. We focus in this study on the central importance of information. Far too often the information process necessitated in manufacture and remanufacture is treated in a secondary or subsidiary manner relative to the processing of the physical product itself. We believe this to be a serious mistake. The processing of information is central to the process of manufacture and is possibly even more critical to remanufacture. We show the importance of information in remanufacture by tracking and conceptualizing the hidden or missing links that information provides. Information, when designed and used properly, can link the movement of physical products and work activities involved in remanufacture.

Our research is at an initial stage and will require completing data collection and verifying conflicting data from the field interviews. Upon completion of our research, we aim to show a clear picture of the entire remanufacture process and the nature and various roles of information involved in the process. We will then be able to determine the required information products for the remanufacture process for an aircraft fuel-pump. We believe that tracing one specific part through remanufacturing in one organization will pave the way for understanding the information needs in manufacture and remanufacture process.

## **2. The View from the Field: Sources of “Dirty” Data**

Based on our initial observations from field interviews, two related areas need further investigation in order to identify the sources of poor quality data. One is the area of obtaining quality data for effectively predicting the need for parts; and the other is the area of effectively providing and recording work activities performed on the parts and in other remanufacture processes.

Unlike initial manufacture where all of the parts needed to assemble a product are known well in advance, remanufacture has far less predictability. Instead, remanufacture involves many unscheduled, variable, and evolving activities. Much of the uncertainty in the process stems from the fact that there are two possible supply chains. One is similar to that found in the initial manufacture, in which new parts are fabricated and delivered by suppliers internal or external to the organization. Unlike initial manufacture, however, this is not the only source of parts. A second “supply line” delivers the parts that are contained in the “carcass,” or product that is to be repaired. Particularly, the unpredictable quality of the parts contained in the carcass that is the

major source of uncertainty. Not knowing whether the parts delivered from the carcass are workable or not makes the need for additional parts through the normal supply chain unpredictable. It follows that a way to reduce this uncertainty is to find better ways to predict the state of the parts contained in the carcass. The predictive capability of these models is, of course, highly dependent upon the quality of the stored information. It is the interaction of the two supply chains that makes remanufacture the complex setting. This is where understanding the information process becomes critical.

Our interviews with those involved in the overhaul process lead us to believe that there is much that can be done to improve the prediction of parts that will be needed. The principal complaint voiced in interviews had to do with data quality. We were told that the current models were unable to accurately predict parts needs because the data on which they are based is faulty and questionable. To describe this, the term “dirty data” is used. Some data that is the input to the predictable model contains serious errors. The question then becomes, where and how do these errors enter into the information process? Some have suspicions about how past demands and future predictions are calculated, to state a few. These suspicions have never been tested to check their validity.

In our research, we chose to trace the fuel booster pump. The pump, as shown in Figure 1, plays a key role in the maintenance and improvement of mission capability and aircraft flying hours. In this research, we trace in detail the repair of the pump, in relation to the aircraft remanufacture managed by the airline.

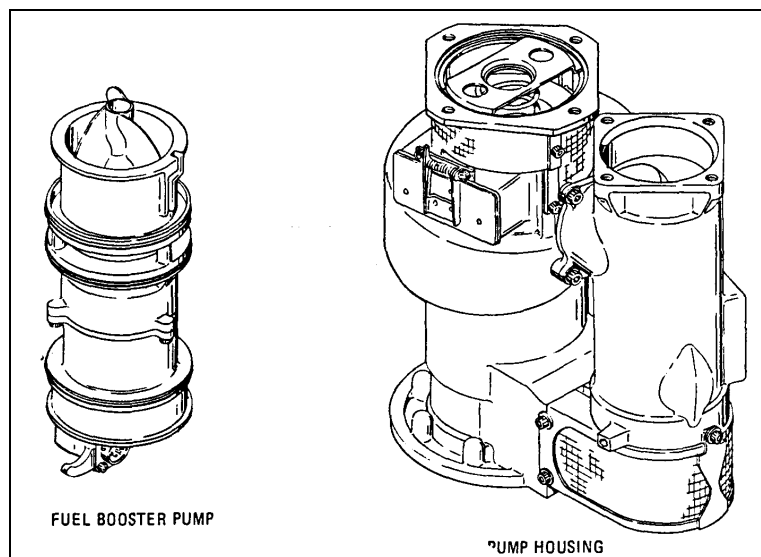


Figure 1: Fuel Booster Pump and Pump Housing

Based on our initial field interviews and document review, we hypothesized that the root cause of problems with the fuel pump is the stator. Therefore, we also traced the stator flow in detail. In so doing, we documented the work roles related to the pump and stator. An exploded view of the stator position in the fuel booster pump is shown in Figure 2.



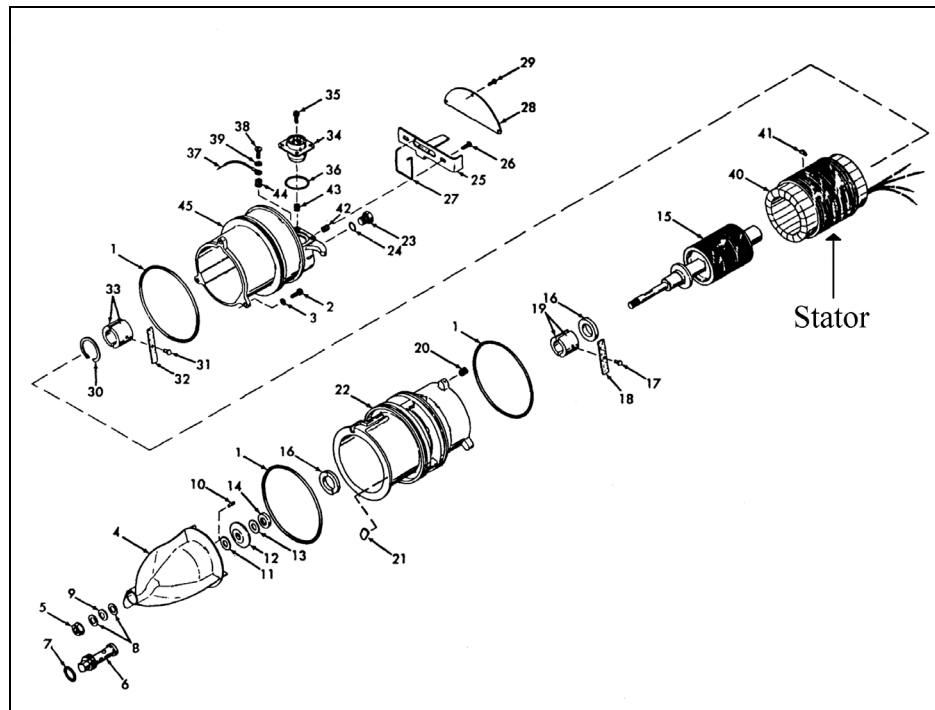


Figure 2: An Illustrative Stator position in the Fuel Booster Pump, Exploded View

Pump repair is performed in two geographically dispersed places: depots and fields. Depots conduct regular scheduled overhaul, whereas fields handle surprises, the immediate problems at hand. Throughout the repair process, information about the work process and the physical products or parts are isolated from each other by operational procedures. The direct impact is that it becomes difficult to connect the two kinds of information. One needs the ability to retrieve and understand the physical parts as well as the work information. For example, when a repairperson sees a pump needing repair, the repair history is not easily accessible. As such, the repair history, the supply information, and the conformance-testing information is stored and used separately. Most of the relevant information is stored and categorized meticulously, but without consideration of cross-area retrieval and access. The connection between physical parts and process information part is missing. Currently, one has to contact multiple agents and places over phone and email to track down the information needed to make this connection. We view that this observation can be an input for designing certain information products for remanufacture.

We observe yet another area for improvement. A supply vendor initially produced the engineering specification of the stator, for the pump. We encountered some opinions that the engineering specifications were not consistent with the stators manufactured and delivered to the Air Force. After going through revisions, the updated specification document and drawings were not stored most effectively by AF and the vendor. In the process, different people could develop different understandings of what the official engineering specifications for a stator should be. Meanwhile, the aircraft has to fly. Some work-arounds might have been performed to meet the demand using parts of questionable quality. In short, the lack of management of data products (in this case, the blue prints of the stator) led to a possible lack of quality in physical products. We hypothesize that changes in engineering specifications over time have been poorly communicated between the Air Force and vendors in terms of specific design problems and resolutions.

### **3. Information Products for Aircraft Fuel-Pump Remanufacturing**

Viewing information as a product implies two essential information management requirements. For historical and future use requirements, information must be stored and protected against undesired change. For current use, information must be kept as current as possible. Information stored in databases is typically safeguarded to preserve these two aspects of quality among others.

As a first step toward solving the problem of dirty data, we have traced the process of remanufacturing fuel pumps from the time that they are removed from an aircraft, from inspection, repair and remanufacture through to re-installation. We fully recognize that every part is unique, as is every organization in the remanufacture business. Different parts in different organizations will not undergo the same process. Nevertheless, we believe that there will be some common elements across both parts and organizations. A thorough understanding of how one part in one organization is handled will enable us to ask the right questions as we extend the study to other parts and other organizations.

Based on our preliminary work. We identify four types of information managed in the remanufacture process of the fuel pump: 1) Blueprints for manufacturing parts and products involved, 2) Conformance lists for testing performance of new and reconditioned parts and products, 3) Plans for supply schedules, and 4) Work records such as repair activities on parts and components. All of these types of information have various life cycles and transfer routes. This variety of information is represented in different forms, processed by different agents, and interfaced with different computer systems. In short, remanufacture demands managing information that resembles what archeologists wish to have when they investigate an archeological site: perfect visibility with all historical integrity attached.

### **4. Discussion and Conclusion**

We believe that information product integrates necessary physical products and work processes. Other research shaped our views on information products for remanufacture. The research and our interpretations are summarized below.

Mead's [8] classical premise of a disjunction between human actions and human grasp of actions raises the question of quality of data that can be the representation and manifestation of reconstructed human actions. Von Hippel [13] conceptualized a reason for costly transfer of certain information, useful for innovation, as "sticky" information. Allen [1] demonstrated that "gatekeepers" can play an important role in the transfer of technical information and thus impacting the technology acquisition and dissemination. Rein and Schon [10] also emphasized the criticality of problem framing that can make a considerable consequences for the nature and type of searching for solution information. Wand and Wang [14] explained the difference between real-world situation and stored information as data quality problems, using an ontological perspective. Madnick [5-7] identified a stream of data quality problems that arise from transferring data from one context and using it in another (different) context. He suggested their reconciliation with "Context Interchange" technology. Strong, Lee, and Wang [11] explored how specific characteristics of data quality problems are changed as data are transferred from one locale to another. Huang, Lee, and Wang [3] suggested examples of information products such as eye-glass prescriptions. Davidson [2] reported a preliminary methods for mapping information products.

Characteristics of how information integrates physical artifacts and activity process determine the kinds and quality of information product. The quality of information reveals how

effectively and efficiently planning and implementation of such integration is performed. We believe that remanufacture process lends itself as useful setting for studying the complex and detailed intricacies involved in information product and its performance.

## 5. References

- [1] Allen, T. J., *Managing the Flow of Technology*. The MIT Press, Cambridge, MA, 1977.
- [2] Davidson, B. and A. T. Chun. Developing a Data Production Map to Identify Data Quality Problems. in *Proceedings of 5th International Conference on Information Quality*. Cambridge, MA: pp. in the CD Proceedings only, 2000.
- [3] Huang, K., Y. Lee and R. Wang, *Quality Information and Knowledge*. Prentice Hall, Upper Saddle River: N.J., 1999.
- [4] Welcome to the Lean Sustainment Initiative Website. 2001. <http://www.leansustainment.org/>.
- [5] Madnick, S. Metadata Jones and the Tower of Babel: The Challenge of Large-Scale Semantic Heterogeneity. in *Proceedings of 1999 IEEE Meta-Data Conference* 1999.
- [6] Madnick, S., X. Chen, J. Funk and R. Wang. Corporate Household Data: Research Directions. in *Proceedings of AMCIS 2001*. Boston, Massachusetts: pp. 2001.
- [7] Madnick, S. E., Database in the Internet Age. *Database Programming and Design*, 1997, pp. 28-33.
- [8] Mead, G. H., *Mind, Self, and society*. University of Chicago Press, Chicago, 1934.
- [9] Millard, F. and M. Lavoie. Developing Data Product Maps for TDQM: The Case of Georgia Vital Records. in *Proceedings of Conference on Information Quality*. Massachusetts Institute of Technology: pp. 17-27, 2000.
- [10] Schön, D. and M. Rein, *Frame Reflection*. Basic Books, New York, NY, 1994.
- [11] Strong, D. M., Y. W. Lee and R. Y. Wang, Data Quality in Context. *Communications of the ACM*, 40(5) 1997, pp. 103-110.
- [12] Tsuji, L. C. (1999). *Tradeoffs in Air Force Maintenance: Squadron Size, Inventory Policy, and Cannibalization*. Master Thesis, Massachusetts Institute of Technology.
- [13] von Hippel, E., "Sticky Information" and the Locus of Problem Solving: Implications for Innovation. *Management Science*, 40(4) 1994, pp. 429-439.
- [14] Wand, Y. and R. Y. Wang, Anchoring Data Quality Dimensions in Ontological Foundations. *Communications of the ACM*, 39(11) 1996, pp. 86-95.
- [15] Wang, R. Y., A Product Perspective on Total Data Quality Management. *Communications of the ACM*, 41(2) 1998, pp. 58-65.
- [16] Wang, R. Y., Y. L. Lee, L. Pipino and D. M. Strong, Manage Your Information as a Product. *Sloan Management Review*, 39(4) 1998, pp. 95-105.

# Data Quality Challenges in Enabling eBusiness Transformation

(Research in Progress)

Arie Segev, Professor & Director\*

Fisher Center for Information Technology and Marketplace Transformation  
University of California, Berkeley  
<http://haas.berkeley.edu/citm>

Richard Wang

Associate Professor Boston University	Research Affiliate Fisher CITM, U. C. Berkeley	Co-Director for MIT TDQM Program
--	---	-------------------------------------

**Abstract:** This paper discusses data quality challenges in the context of eBusiness Transformation. It presents the major differences between traditional and eBusiness as they relate to business models, organizations, processes and technologies, and then outlines the differences with respect to data quality approaches. The scenarios described pose significant data quality (and other) challenges, and the paper discusses work in progress to construct a data quality strategy and implementation methodology.

## 1. INTRODUCTION

The field of data quality has witnessed significant advances over the last decade. Today, researchers and practitioners have moved beyond establishing data quality as a field to resolving data quality problems, which range from data quality definition, measurement, analysis, and improvement to tools, methods, and processes [1, 3, 5, 6, 11-19]. With many of the theoretical foundations developed, researchers have begun to go beyond the fundamental data quality research to solving critical business problems. For example, research has been initiated to investigate how to develop *data production maps* for information supply chain management and remanufacture [8]. Another area of active research is the conceptualization and software implementation for *corporate household* [10]. One research area that has not been actively pursued, however, is data quality in the context of eBusiness.

The Internet and eBusiness added new complexities to data quality primarily due the increase in a company's interaction with its environment – *externalization*; and new levels of *data integration* resulting from new business models. That business-to-business (B2B) integration calls for the augmentation of data manufacturing models with *data logistics* concepts. Furthermore, it is imperative that organizations establish data quality strategies and implementation methodologies combined with their eBusiness transformation approaches. In this paper we focus on B2B eBusiness, but there are obvious links to B2C eBusiness, for example, product and inventory information, which is used for B2C purposes, would inherit quality problems that were introduced in the data manufacturing process.

---

\* The work of this author was supported by the External Acquisition Research Program (EARP) under contract N00244-99-C-0034

eBusiness Transformation entails business, organizational and technological aspects. It should be based on a comprehensive top-down view of the enterprise and its environment and incorporates proven principles when applicable. Basic principles of conventional information systems methodologies that have been developed in the last ten or more years still apply, but the scope and context have changed significantly. Section 2 discusses the eBusiness transformation process and elaborates on the business integration aspect. Section 3 then elaborates on the inter-company aspect and discusses four different scenarios; examples from the domain of B2B eProcurement are presented.

## 2. eBUSINESS TRANSFORMATION

eBusiness Transformation entails business, organizational and technological aspects. It should be based on a comprehensive top-down view of the enterprise and its environment and incorporates proven principles when applicable. Basic principles of conventional information systems methodologies that have been developed in the last ten or more years still apply, but the scope and context have changed significantly. The new context is characterized by:

- New business models, applications and related requirements
- The externalization level of companies
- The degree of required interconnectivity and integration
- The rate of change (technology and business models).

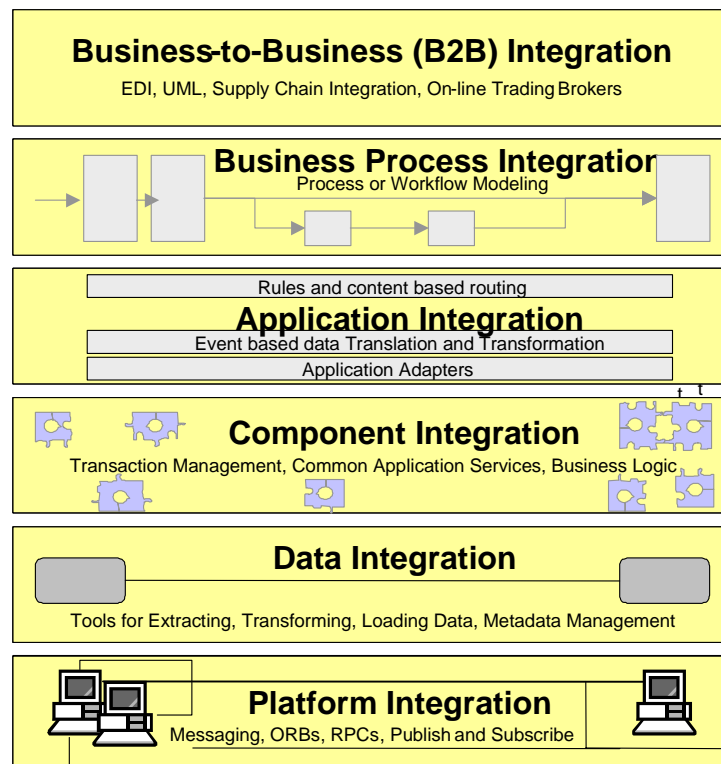
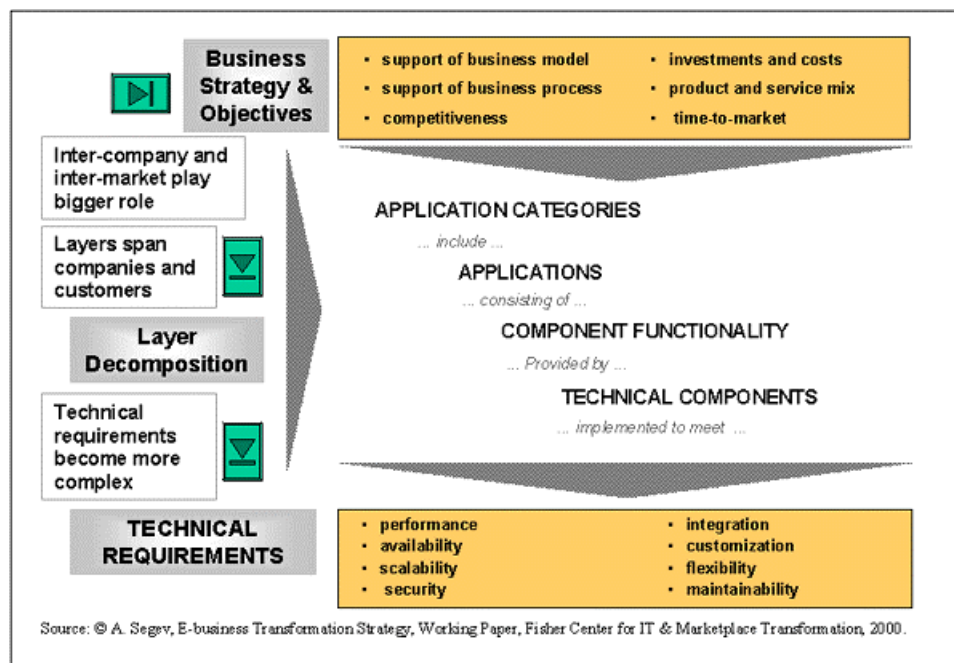


Figure 1: EAI Market Segmentation [Gold-Bernstein 1999]

The second bullet point indicates that increasingly, company's processes are shifted outwards as part of new business models involving interactions with customers, suppliers and partners. This, in turn, has led to an exponential increase of the company's interfaces (i.e., the level of business connectivity). From a process and data perspective a new level of Business-to-Business integration need emerged. A typical methodology used in addressing this need has been to expand the Enterprise Application Integration (EAI) technology beyond the corporate walls and delivers the full promise of eBusiness by integrating customers, suppliers and partners (see Figure 1). The basic principle is to create a decomposition-based application and technical infrastructure to support the business objectives and satisfy various performance constraints.

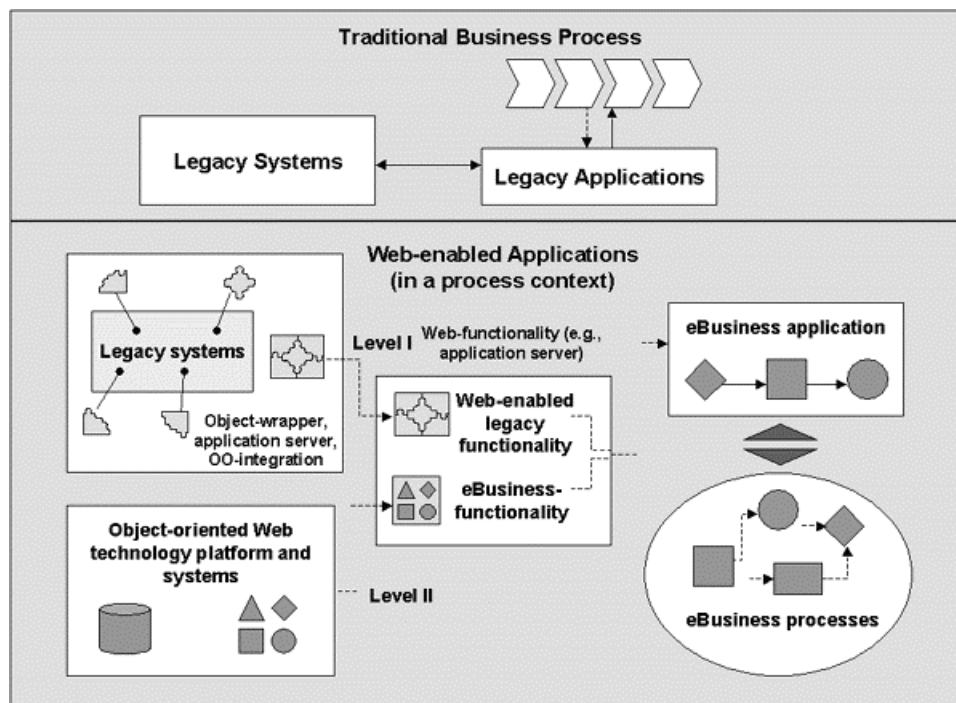
The previous figure represents sound decomposition principles, but it has the disadvantage of conveying a single company's perspective. We prefer to use the diagram of Figure 2, below to emphasize the new eBusiness requirements. It is important to note that while the requirements at the bottom of the figure are referred to as "technical," they of course have significant business and cost ramifications. The figure does not represent a decision model itself, but rather includes the scope, the elements, and the overall ideal order of the decision types. Frequently, one has to deal with a subset of the issues in a narrower and less systematic fashion, but whenever possible, this general framework should be followed or related to. The emphasize here is on the scope of the inter-organizational processes, the required infrastructure, as well as new organizational and skill dimensions.



**Figure 2: A Framework for eBusiness Integration**

The new type of eBusiness applications involves a business and technology change in delivering products and services. An immediate requirement that companies face is to **Web-enable** legacy systems. Web servers, and the application server in particular, have become the foundation of business service delivery and, consequently, the Web service model must be central to

modernizing or moving out of legacy systems. It is important to understand that this model is as much a business model as it is a technology model. The range of Web-enabling possibilities is wide, but two general approaches are used when legacy systems are present. We refer to them as Level I and Level II. In Level I solution there is no significant change in functionality and it is based on creating an interface between the legacy system and the Web server. The concern here is the presentation and user interface, and it is similar to “GUI wrapping” that became a popular approach in the early days of client server. In the case of Level II solution, additional process functionality is introduced. The advantage of an application server is that it can be used for various degrees of Level I and Level II integration as shown in Figure 3. It also allows various degrees of inter-process integration and data quality enhancement. As an example, the application server enabled legacy application in the figure can provide data to a new object-oriented Web-based application, and the integration of the two at the application server provides the unified added value service that underlies the new eBusiness process. Furthermore, simple, but important, data quality enhancements can easily be introduced at Level II, e.g., performing validity checks on data attributes that were not implemented in the original legacy system. Cleaning the data at this junction is more effective and cheaper than data cleaning procedures downstream. In addition to the accuracy enhancement of Level I, Level II enhancements can be not only functional but also data quality. Relevant dimensions are completeness - enhanced through capturing more data and possibly relating it to other data (semantic completeness); timeliness – enhanced by capturing real-time data instead or in addition to other channels. In the next section we analyze in more details the data logistics as it moves across companies; the web-enablement approach described above is also applicable to many of those cases.



Source: © A. Segev, eBusiness Transformation Strategy, Working Paper, Fisher Center for IT & Marketplace Transformation, 2000.

Figure 3: Web-Enabling Legacy Systems

### **3. INTER-COMPANY eBUSINESS INTEGRATION AND QUALITY ENHANCEMENT**

There are four primary cases of inter-company eBusiness scenarios discussed below with respect to data quality strategies. These cases are discussed in the context of coordination and negotiation in [2], [9], [4], [7]. While not capturing all possible scenarios, we believe that these cases are the most important and represent the majority of real-life scenarios.

#### **Case I: 1C**

The case of a single company corresponds to the traditional intra-company data quality scenario. As discussed earlier in this paper, this case has received intensive attention in the last ten years both in academia and industry. The application server example in the preceding section is applicable to this case.

#### **Case II: 2C**

The case of two companies corresponds to dedicated systems between two trading partners, ranging from faxed papers and telephone calls, to traditional EDI and Web-EDI, to contemporary XML-based connectivity. Data quality issues in traditional systems (many of them are legacy systems) were identified and addressed long time ago both in research and in industry. In the case of fax, telephone errors are introduced due to “misunderstanding” and more errors through retyping. There is often a “semantic reduction” as a result of translations to other systems. For electronic transmissions the following are common cases.

**EDI:** in addition to cost (setup and operational) many recipients print and re-input; in particular small companies. Translators and mappers improved situation somewhat. Further semantic problems arise in matching the received data with other data - from the same partner but from other systems, frequently arising because of the complexity, cost and time to modify existing EDI systems.

**Web-EDI:** primary objective was to reduce transport cost and possibly by-pass expensive VANs. One quality dimension improved is the timeliness when periodical downloads from VAN is replaced by more “real-time” web-based connectivity.

**XML-based:** These are contemporary systems, most implemented in the context of Case IV below. One should distinguish between two primary types of applications:

Transactional applications: including XML-EDI and new “pure” XML connectivity such as in Desktop Procurement Systems (DPS), e.g., DPS connectivity to inventory systems of the supplier. In many cases the 1-to-1 connectivity was changed to 3C2L by using the services as a content intermediary.

Collaborative applications: e.g. design, customer support; added connectivity and timeliness. More complete information. Workflow technology plays a major role.

Common problems are similar to those encountered twenty years ago when companies moved from file-based systems to databases by emulating the former on the latter, resulting in more efficient GIGO process. Unless this process is accompanied by a methodology-based process



and data quality improvement, the results will be similar, but with much more serious (and perhaps catastrophic) results to the business. The lower portion of Figure 4 illustrates the case of direct connectivity between the buyer and the seller in the context of eProcurement. It typically involves a significant business relationship that justified the cost of setting the one-to-one business integration. A main obstacle to data quality enhancement is the legacy EDI conduit which makes it difficult to add to the functional business integration, leading to parallel systems that don't integrate well relative to the end-to-end process. There is also typically ambiguity about the responsibility of each company for the data quality.

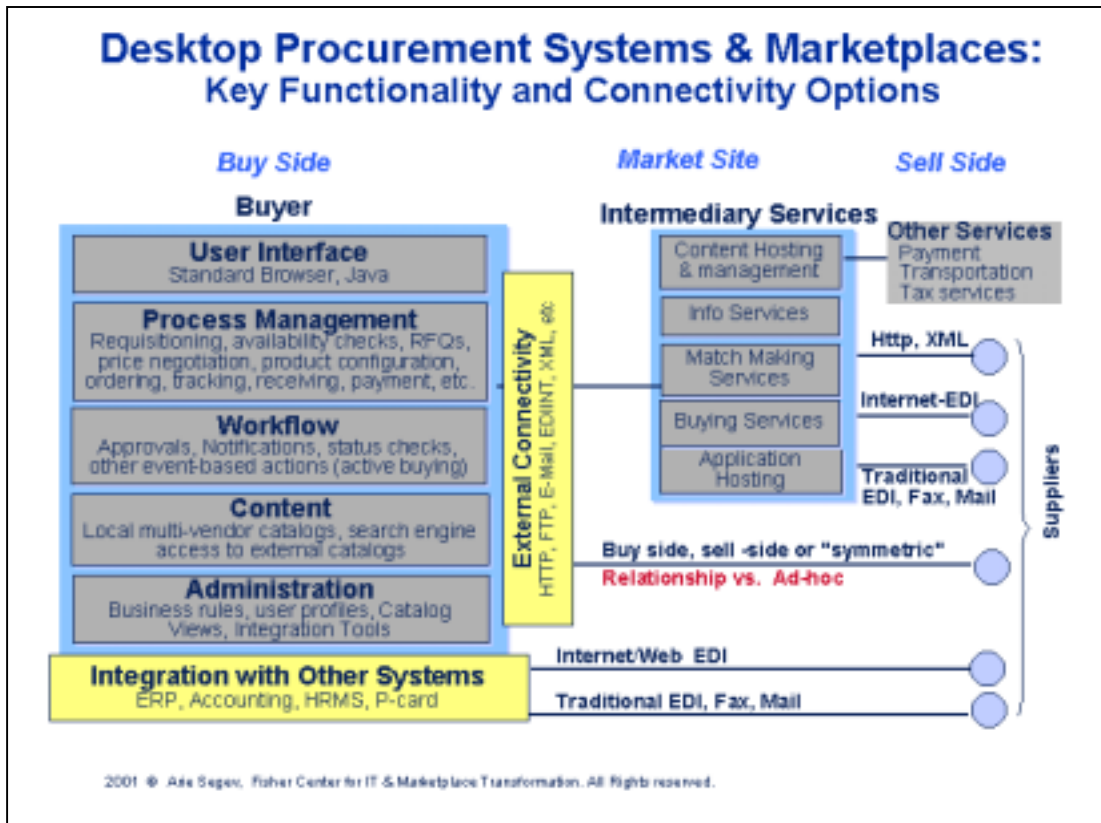


Figure 4: eProcurement Marketplace

### Case III: nC1L

This case corresponds to providing a solution from multiple complementary vendors. Turn-key solution providers, e.g., restaurant kitchens, and travel packages are examples of traditional processes. Basic data quality problems in such an environment are not new and result from lack of collaboration among the suppliers, which are further complicated by severe data segregation constraints, e.g., a tour operator's data warehouse must maintain separate customer lists received from the various suppliers (airlines, hotels, etc). This last example was an information product example; for physical goods, there are various business models, including buying into inventory and then providing the components or serving as a broker and interacting with the multiple vendors.

In a dynamic eBusiness environment, major data quality problems have adverse effects on coordination, product compatibility, etc. The data quality problems are in integrating the data from the multiple vendors.

#### Case IV: nCmL

The case of n Companies and m Levels represents a supply web environment that includes elements of the previous cases. A particular sub-case, 3C2L, is the most common form of marketplace intermediation. The upper section of Figure 4, illustrates it in the context of eProcurement. Some of the most difficult data logistics problems arise here, since many suppliers are not enabled to move high quality XML-based data from their transactional systems to the e-catalogs of the intermediaries. The intermediaries add value by creating an integrated e-catalog after cleaning the data from the various sources, parts or whole of which are downloaded by the buyer. While the data semantics and completeness is generally better than the non-intermediated case, it can actually be less timely than the EDI solutions discussed in Case III.

The most general case involves multiple levels (supply chain) as well as multiple parties at a given level. It generally requires industry-wide standards for data and processes and often pursued in the interoperability context. Figure 5 represents such a case in the context of a joint project between the department of Architecture and CITM. It involves architectural and interior design processes and technologies combined with multi-level supply web environment, both B2C and B2B. Key functional components are listed in the figure, all present data quality issues. Document workflow systems are the basis of collaborative design and negotiations and they need to be integrated with e-catalogs and other pieces of information in many real-life situations. This introduces the most difficult interoperability issues, but without their solutions, the particular business model can't be implemented.

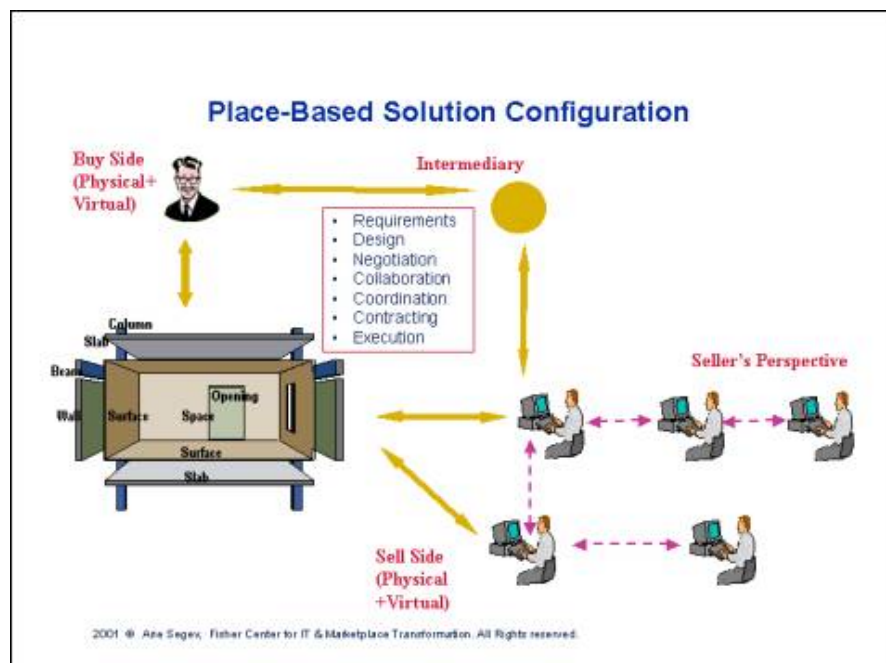


Figure 5: Place-based Solution Configuration

#### **4. SUMMARY**

This paper discussed data quality challenges in the context of eBusiness Transformation. The increased levels of company externalization make B2B integration a difficult proposition. That integration is of business models, processes, and technologies. The paper focused on eBusiness transformation for the B2B case and data quality implications. The eBusiness scenarios described pose significant data quality (and other) challenges; taxonomy of the scenarios and understanding the various data quality pitfalls are part of a data quality strategy designed to effectively deal with multiple points of data quality enhancement. The work presented in this paper is work in progress to construct a data quality strategy and implementation methodology.

## REFERENCES

- [1] Ballou, D. P., R. Y. Wang, H. Pazer and G. K. Tayi, Modeling Information Manufacturing Systems to Determine Information Product Quality. *Management Science*, 44(4) 1998, pp. 462-484.
- [2] Bichler, M. and S. A., Methodologies for the design of negotiation protocols for E-markets. *Computer Networks*, 2001.
- [3] Fedorowicz, J. and Y. Lee, Accounting Information Quality. *Journal of Accounting Information Review*, 3(1) 1999, pp. 1-7.
- [4] Gold-Bernstein, B., EAI Market Segmentation. *EAI Journal*, 1999.
- [5] Huang, K., Y. Lee and R. Wang, *Quality Information and Knowledge*. Prentice Hall, Upper Saddle River: N.J., 1999.
- [6] Kahn, B. K., D. M. Strong and R. Y. Wang (1999). Information Quality Benchmarks: Product and Service Performance. *Communications of the ACM*, Forthcoming, 2002.
- [7] Kim, S. A. a. J., Frictionless Market and Automated Coordination. *CITM Working Paper*, 2001.
- [8] Lee, Y., T. Allen and R. Wang. Information Products for Remanufacturing: Tracing the Repair of an Aircraft Fuel-Pump. in *Proceedings of Sixth International Conference on Information Quality*. Cambridge, MA: pp. 77-82, 2001.
- [9] M., G., G. J. and S. A. Multi-Vendor Electronic Catalogs to Support Procurement: Current Practice and Future Directions. in *Proceedings of Bled Conference on Electronic Commerce*. Slovenia: pp. 1999.
- [10] Madnick, S., R. Wang, F. Dravis and X. Chen. Improving the Quality of Corporate Household Data: Current Practices and Research Directions. in *Proceedings of Sixth International Conference on Information Quality*. Cambridge, MA: pp. 2001.
- [11] Redman, T. C., ed. *Data Quality for the Information Age*. 1996, Artech House: Boston, MA. 303 pages.
- [12] Storey, V. C. and R. Y. Wang. An Analysis of Quality Requirements in Database Design. in *Proceedings of the 1998 Conference on Information Quality*. Massachusetts Institute of Technology: pp. 64-87, 1998.
- [13] Strong, D. M., Y. W. Lee and R. Y. Wang, Data Quality in Context. *Communications of the ACM*, 40(5) 1997, pp. 103-110.
- [14] Wand, Y. and R. Y. Wang, Anchoring Data Quality Dimensions in Ontological Foundations. *Communications of the ACM*, 39(11) 1996, pp. 86-95.
- [15] Wang, R., J. Funk, Y. Lee and L. Pipino, *Journey to Data Quality*. MIT Press, Cambridge, Massachusetts, Forthcoming.
- [16] Wang, R., M. Ziad and Y. Lee, *Data Quality*. Advances in Database Systems, ed. A. K. Elmagarmid. Kluwer Academic Publishers, Norwell, Massachusetts, 2001.
- [17] Wang, R. Y., A Product Perspective on Total Data Quality Management. *Communications of the ACM*, 41(2) 1998, pp. 58-65.
- [18] Wang, R. Y., Y. L. Lee, L. Pipino and D. M. Strong, Manage Your Information as Product: The Keystone to Quality Information. *Sloan Management Review*, forthcoming, 1997.
- [19] Wang, R. Y. and D. M. Strong, Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, 12(4) 1996, pp. 5-34.

## Improving the Quality of Corporate Household Data: Current Practices and Research Directions

**Stuart Madnick**  
M.I.T.  
smadnick@mit.edu

**Richard Wang**  
Boston University  
rwang@bu.edu

**Frank Dravis**  
Firstlogic Inc.  
frankd@firstlogic.com

**Xinping Chen**  
Bose Corporation  
xinping\_chen@bose.com

**ABSTRACT:** Corporate household data not only refers to the strict hierarchical structure about and within the corporation, but also the variety of inter-organizational relationships. It is becoming increasingly important for many purposes ranging from CRM and ERP applications, to risk management, supply chain management, and marketing. We propose conceptual definitions for corporate household, corporate household knowledge, and corporate household knowledge processor. After describing research challenges and conceptual definitions, we summarize current practices and approaches. We then present a two-part plan: (1) continue our qualitative research to describe the various different sources, views, and purposes for corporate household data, including the rules used in each case; (2) apply the context interchange theory to represent the corporate household data and underlying knowledge and enable the context mediation technology to correctly understand and reason about both the context of the sources and the context of the user's query about corporate household data.

### 1. Introduction

How corporate structure and corporate relationships are interpreted and used depends on the context. Additionally, they evolve dynamically in the rapidly changing business environment. The ability to capture, manage, and use knowledge of corporate structure and relationships are fundamentally critical tasks underpinning many important activities, such as marketing promotion, financial risk analysis, and supply chain management. The problem goes beyond existing data quality research such as those found in [2, 5, 9-11, 13-16]. A recent literature search of ABI/INFORM from 1986 to 2001 [7] reveals that there is no corresponding concept for *corporate household data* to that of *individual household data*, and the terms **corporate household** and **corporate householding** have not been mentioned. Although there are many papers on "corporate structure" [1, 8, 12], that term does not capture the rich phenomena we are exploring. By **corporate household** we not only mean the strict hierarchical structure within the corporation, but also the variety of inter-organizational relationships.

To put this issue in perspective, consider a traditional household. As family structure evolves, such as the increasing number of single families, families with no children, or husband and wife with different last name, it becomes more difficult to define and identify "household", [6]. For example, are grandparents or visiting cousins living at same address to be considered part of the same household? Are two unmarried people living together a household? There is no single "right" answer; the answer depends upon the intended purpose of the question.

Similarly, a corporate household must be conceptualized within certain scope, content, and context, with the relationships identified within a corporation, between suppliers and the corporation, and between the business customers and the corporation. Conceptually, the corporate structure would also be different depending on different contexts such as a financial

perspective, legal perspective, and the reporting structure. Identifying those contexts and representing the right structure for the right task can provide competitive advantage.

Our primary research methods are semi-structured interviews via telephone, face-to-face, case studies, and surveys. From these results, we explore possible solution approaches and new technologies. Our long-term research goal is to define the concepts of *corporate household* and *corporate householding* and develop mechanisms to capture the metadata and business rules representing the semantics of *corporate household data*, and the *corporate household knowledge processor* that would produce the appropriate corporate household that would be fit for use depending on the context. We begin with expository example problems.

## 2. Corporate Household Knowledge Challenges

Corporate household knowledge serves various purposes, which in turn necessitate different interpretations of the information. Consider the list of organization names in Figure 1. What is the relationship among these names? As it turns out, these are all names that are in some way related to each other and International Business Machines Corporation (the name at the top of the list). These names include abbreviations (such as IBM), divisions (such as IBM Microelectronics division), wholly owned subsidiaries (such as IBM Global Financing), partially owned subsidiaries (such as IBM de Colombia, S.A.), companies that were acquired by IBM (such as Lotus Development Corporation), companies that were acquired and then later sold by IBM (such as SoftwareArtistry, Inc.), companies in which IBM has a minority joint venture interest (such as the Dominion Semiconductor company), and companies that IBM has a majority joint venture interest in (MiCRUS). It even includes IBM's original name, Computing-Tabulating-Recording Company.

What is the significance? Consider a rather simple question: "How many employees does IBM have?" In a recent study of a major insurance company, this was an important question asked in setting premium rates for business owner protection insurance [17]. Which entities listed above for IBM should be included in this count? How to avoid double counting? The answer depends upon the purpose of the question. The important and subtle issue is: "When is one entity to be considered part of the another entity?" Such corporate household knowledge is used for many different purposes: (1) financial risk, (2) account consolidation, (3) marketing (multiple divisions & subsidiaries), (4) customers & supplier consolidation, (5) customer relationship management, (6) regional and/or product separations, (7) legal liability in insurance, (8) conflict of interest & competition, and (9) ad hoc/temporary structures. In some cases the two entities should be combined and in other cases the two entities should not be combined.

International Business Machines Corp
IBM
IBM Microelectronics Division
IBM Global Services
IBM Global Financing
IBM Global Network
IBM de Colombia, S.A
Lotus Development Corporation
Software Artistry, Inc.
Dominion Semiconductor Company
MiCRUS
Computing-Tabulating-Recording Co.

**Figure 1. List of organization names**

As noted before, corporate household changes over time; thus, the context also changes over time. For example, at one point Lotus Development Corporation was a separate corporation

from IBM. When doing a historical comparison of growth or decline in “number of employees” of IBM, should current Lotus employees be counted in a total as of today? Should the Lotus employees in 1980, when it was a separate corporation, be added with the IBM employees of 1980 to make a meaningful comparison?

Corporate household knowledge could be applied in many other areas. An executive of a global manufacturing company had concerns, in the global sourcing context, in identifying a manufacturing site that could produce a particular product with the lowest costs. A big part of manufacturing cost is raw material cost, and therefore identifying and maintaining relationship with material vendors are critical in reducing costs. However, due to localized systems, different manufacturing sites may have different, independent relationships/contracts with the same vendor for the same material. Inconsistencies between systems make it difficult to understand a vendor globally or know how much of a raw material is used on the global basis. As a result, the company could neither take advantage of nor negotiate low prices across all of its manufacturing sites [7].

### 3. Conceptual Definitions

*Random House Webster's Unabridged Dictionary* defines *corporation* and *household* as:

**Corporation:** (1) an association of individuals, created by law or under authority of law, having a continuous existence independent of the existences of its members, and powers and liabilities distinct from those of its members, (2) any group of persons united or regarded as united in one body.

**Household:** (1) the people of a house collectively; a family including its servants, (2) of or pertaining to a household: household furniture, (3) for use in maintaining a home, esp. for use in cooking, cleaning, laundering, repairing, etc., in the home: a household bleach, and (4) common or usual; ordinary.

The term “householding,” although not in most dictionaries, has been used in an increasing number of contexts. For example, it was recently used in notices, such as the one below, sent to hundreds of thousands of people and organizations as a result of a recent SEC rule.

#### HOUSEHOLDING ELECTION

This notice has been placed in this mailing on behalf of your Broker or Bank. In December 2000, the Securities and Exchange Commission enacted a new rule that allows multiple shareowners residing at the same address the convenience of receiving a single copy of proxy and information statements, annual reports and prospectuses if they consent to do so. This is known as "Householding." Please note that if you do not respond, Householding will start 60 days after the mailing of this notice. We will allow Householding only upon certain conditions. Some of those conditions are:

- The Issuer agrees to have its documents Household,
- You agree to or do not object to the Householding of your materials,
- You have the same last name and exact address as another shareowner(s),
- Consistency with your Broker's or Bank's practices.

If all of these conditions are met, and Securities and Exchange Commission regulations allow, your household will receive a single copy of proxy and information statements, annual reports and prospectuses. . . . Your affirmative or implied consent to Household will remain in effect until you revoke it by calling the telephone number listed in the HOUSEHOLDING ELECTION paragraph.

In terms of the scope and content of a corporate household, we found at least three types of important entities: the *corporation*, *suppliers/vendors*, and *customers*. A corporation includes relationships, functions, and people within the entity of the corporate, such as the one represented by organizational chart. We therefore propose the following conceptual definitions:

**In the dynamic and rapidly changing e-business environment, any group of persons united or regarded as united with the corporation, such as suppliers and customers whose relationships with the corporation must be captured, managed, and applied for the purpose of activities such as marketing promotion, financial risk analysis, and supply chain management in their entirety forms a corporate household. Note that there might be multiple overlapping but distinct corporate households, depending upon the precise set of relationships that are important for the task at hand.**

**The knowledge developed for such purposes are termed corporate household knowledge.**

**The algorithms and corresponding software system that produces the appropriate corporate household knowledge fit for use for the task at hand is called a corporate household knowledge processor.**

With example problems and conceptual definitions described, we next explore solution approaches.

#### **4. Inter-entity Relationships: D&B Family Tree**

Dun & Bradstreet (D&B) has developed a representation of corporate structure. D&B's Data Universal Numbering System [3], D-U-N-S Number, is a unique nine-digit non-indicative identification number assigned to every business entity in D&B's databases. It widely used for keeping track of millions of corporate family structures and their relationships worldwide. The D&B Family Tree is comprised of linkages and business relationships. Linkage, in general terms, is the relationship between different companies or specific sites within a corporate family. Linkage occurs in D&B WorldBase when one business location has financial & legal responsibility for another business location. Other types of family relationships may occur but are not linked in the D&B file because the affiliated company has no legal obligation for the debts of the other company, such as businesses affiliated through common officers or situations where one corporation owns a part or minority interest in another (50% or less).

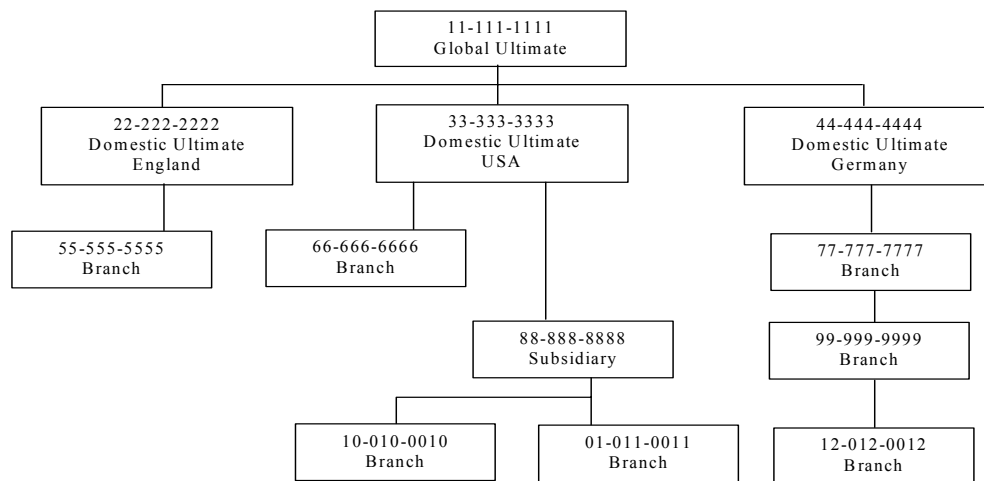
D&B's corporate family tree is structured with eight types of entities (single location subsidiary, headquarters, branch, division, subsidiary, parent, domestic ultimate, and global ultimate) and two types of relationships (branch to headquarter linkage, and subsidiary to parent linkage). Each entity is uniquely identifies by a D-U-N-S number. Finally, there is one more business relationship, the Organization Parent D-U-N-S, which is the top most subsidiary, which reports to the Global Ultimate; it has been identified so that customers who prefer can link multi-level family trees, subsidiaries & branches to the direct reporting parent of the global ultimate.

For the purposes of linking these relationships to define corporate responsibility, each



family member carries up to four D-U-N-S Numbers: (1) its own Case D-U-N-S Number, (2) the next highest level in the family: parent or headquarter D-U-N-S, (3) the highest level within its country: its domestic ultimate D-U-N-S, and (4) its top global ultimate; global ultimate D-U-N-S.

Each record carries a set of linkage elements which help to identify the type of record it is, as well as its relationship to other records in the family tree. They are the Status Code, Subsidiary Code, Hierarchy Code, Dias Code and Global Ultimate D-U-N-S Number. The status code is a one-digit field which identifies a record as Single Location, Headquarters, or Branch. The subsidiary code is a one-digit field which identifies the record as Subsidiary, or Non-Subsidiary. The hierarchy code is a two-digit field which determines the record's relative position in a family tree by indicating its relationship to other records. Global Ultimates have a hierarchy codes of "01", while subsidiaries have a hierarchy code of one greater than their parents', and branches have a hierarchy code equal to their headquarters'. The dias code is a nine-digit field which sorts a corporate family tree into family sequence. The dias code changes each time the linkage file is updated. In general terms, all branches will be listed directly below their headquarters while subsidiaries will be listed directly under their parents. In a situation where a parent/headquarters has both branches and subsidiaries reporting to it, the branches will be listed first, followed by the subsidiaries. Branches are sorted alphabetically by country, while subsidiaries are sorted alphabetically by company name. Figure 2 presents a typical corporate family tree structure.



**Figure 2: An example of corporate family tree**

Although the D&B corporate family structure is useful for financial and legal purposes within the corporate boundary, there are many other types of corporate relationships, such as those in the global manufacturing context. Those corporate relationships extend beyond the corporate boundary, and are fundamental in facilitating business decisions such as how to provide a single global sourcing capability or consolidate manufacturing plants globally.

## 5. Entity Identification: Knowledge Structuring using Subject Matter Experts

Another important aspect of corporate household data is "entity identification." As a simple example, it would involve recognizing that "MIT," "M.I.T.," "Mass Inst of Tech," and

“Massachusetts Institute of Technology” are all names referring to the exact same entity. In this section, we will review an approach that was developed by Firstlogic, Inc. The approach is used when an organization must structure its data to represent the desired abstract view, be it a risk aggregation context, supplier context, global customer context, etc. Often a CRM or ERM system is being used to host the data. It is often the initial load/migration operation into the CRM system that provides the impetus.

No matter the type of householding, be it residential-marketing or corporate-financial, a Subject Matter Expert (SME) approach can be used to help identify and build hierarchical structures to represent relationships between two families. The “family” can be either a two-person residential household or a bank with eight hundred legal entities. In any case the organization engaged in householding is seeking to either: (a) identify the entities in their own family structure, the **internal view**, or (b) identify the entities in the family structure of their business target, the **external view**.

The SME approach is applied when an organization attempting to build the external view has data in its computer systems representing the business relationship, has SME’s that are knowledgeable of both the data and corporate goals, but has no architecture in place that contains and represents the data according to desired abstract view of the organization.

What follows is an example corporate householding project in which a Firstlogic consultant assisted a global financial software (GFS) vendor to:

- a. Allow GFS a single view of its customers
- b. Allow GFS to identify the relationships, at the entity level, between its own corporate tree and those of its top corporate customers.
- c. Allow GFS to plan strategies and marketing campaigns to more effectively leverage the relationships between it and it’s top customers.

We will use the example of GFS to draw out and highlight the step-by-step corporate householding entity identification approach employed. The approach is used to first extract the true business goals, then desired views, rules, and then superimpose those rules on the data. Ultimately, the desired representation is achieved by designing a schema that fits the targeted information repository which, in the case of GFS, was a CRM system.

**Step 1 – Establish project goals.** In any project it is crucial to first establish the goals. In the case of GFS the goal was to achieve a single view of their current corporate customers. Unlike some corporate householding operations, GFS was not interested in consolidating data up and down the organizational tree of their customers, but was instead solely interested in consolidating customer accounts across each branch in the hierarchical structure.

While GFS had data and SME knowledge concerning the parent-child relationships of their customers, their customer purchasing decisions were not driven from the top corporate parent, but mostly from the divisional headquarters. For example, GFS knew that IBM was the parent of Lotus and Informix, but Lotus and Informix from GFS’s perspective retained their own purchasing decisions, at least at the price and volume levels at which GFS sold their product. Consequently even though GFS had the data to vertically rollup the corporate structure of IBM, for example, the purchasing patterns of their customers dictated that they instead consolidate

across each level or branch of the IBM corporate tree. Thus their business rules, represented by the match and consolidation rules, focused on obtaining a single view of the purchasing patterns of Informix or Lotus, etc. But, it was still important, and a challenge, to recognize when two instances, such as “Lotus” and “Lotus Development Corp,” were, in fact, Lotus, for example.

**Step 2 - Define applicable terms and gain cross-functional agreement.** For GFS this meant defining a customer, a contact, a confidence level, etc. The terms had different meanings to different people within GFS. In order for the project to move forward under a common understanding, everyone had to at least agree on the project lexicon. Standardization of meta data definitions allows for uniform queries and reporting of information across the enterprise, in addition to identifying data anomalies. It was often found that when people disagreed on a definition, the opposing parties are really saying the same thing just in a different way, and are reluctant to give up the semantic tug of war. Having an objective third party can help steer the participants back to the corporate goals.

**Step 3 - Define the business rules that attain the goals.** In the case of GFS this meant writing the rules out in English and confirming with a cross functional team that the rules supported the organizational goals. The written business rules included:

- Rules identifying duplicates
- Rules identifying duplicate record confidence levels
- Examples of acceptable duplicates
- Examples of confidence levels
- Rules governing consolidation logic
- Examples of acceptable consolidated records
- Process flow. Sequential matching steps needed for multi-level matching
- Special field level consolidation logic. In the case where two company records are to be consolidated, but individual fields contain different or opposing contents, such as phone numbers or account numbers, what should be done.

When the rules were written out significant debate often occurred concerning the language, especially syntax, of the rule. Another point of debate regarded identifying duplicate or redundant rules. For example, GFS submitted the following three rules that identified a duplicate contact person at a corporate client.

1. Duplicate contact = same name at same address
2. Duplicate contact = same name, address, and e-mail
3. Duplicate contact = same name and e-mail

In the three rules above, rule #1 will identify as duplicate all records with the same contact name and address. For those contact records that have the same name, possibly *different* address, but same e-mail, rule #3 will identify duplicates. “Associating” the duplicates found by rule #1 and #3 via contact name will create a complete set of duplicate contacts across the data set. Rule #2 is redundant. While it is the more specific rule, it was too specific for the contact record patterns found in GFS’s data. Rule #2 would identify duplicates, but always less than Rule #1 or #3. As show below:

<u>Name</u>	<u>Address</u>	<u>E-mail</u>
1. Jon Smith	100 2 <sup>nd</sup> St	Smith@abc.com
2. John Smith	100 2 <sup>nd</sup> St	Smith@Abacuscomputers.com
3. J Smith	100 Front St	Smith@abc.com

Within the matching rules defined by GFS, “Jon”, “John”, and “J.” are allowed as the same first name as long as the contacts had the same exact last name. The “2<sup>nd</sup> Street” and “Front Street” addresses are actually the same physical address with the Front St address being a “prestige” address. The “Abacus Computers” e-mail address was derived from the original company name before it was changed to “ABC”. Matching rule #2 would not have identified any of the three records as duplicates. Only the combination of the two more general rules #1 and #3, and the joining (associating) of results via name could duplicates be identified with confidence at or above the thresholds set by GFS. It took a bit of explaining by the consultant to convince the GFS SMEs that rule #2 was not needed, and that it would slow the matching process with no benefit.

The process of writing the rules in English forces the SME’s to think in concrete terms what they want to achieve and how. When the rules are placed on paper the text will often fail to match what the SMEs thought they meant. The crafting of written rules is an important test. If a written rule can not be agreed upon by the parties involved, this is an important indication that either the rule is unfinished or there is a fundamental misunderstanding of the goal driving the rule. In either case, the conflict must be resolved if the desired single view of the customer, according to the applicable context, is to be obtained.

**Step 4 – Create Rule Matrix.** Once the written business rules are agreed upon and signed-off by all of the applicable personnel the match and consolidation rules are stored in a rules matrix. The purpose of a rules matrix is to gather into one table all the householding rules. The business rules will initially be recorded in a project plan, a requirements document, a statement of work, or some other document provided by the client, and can be embedded in many pages of text. Extracting the match and consolidation rules into one matrix makes it easier to view the entire body in context and further evaluate the rules for redundancy, or the existence of flaws in the householding logic.

**Step 5 – Verify Rules Matrix.** The match and consolidation matrix must be verified and accepted by the SME’s. Redundant rules are marked for deletion and new rules to be added are highlighted.

**Step 6 – Create Application Parameters.** The match and consolidation rules in the matrix must be converted to application parameters. In the case of GFS, Firstlogic’s Information Quality Suite was used to perform the match and consolidation. The consultant loaded the rules matrix into the IQ Suite via the application’s graphical user interface. Each row in the matrix represented a set of match or consolidation criteria that the IQ Suite accepted as job control parameters. The complete translation process from business goals to executable match and consolidation criteria of GFS’s corporate householding project is depicted in Figure 3 below.

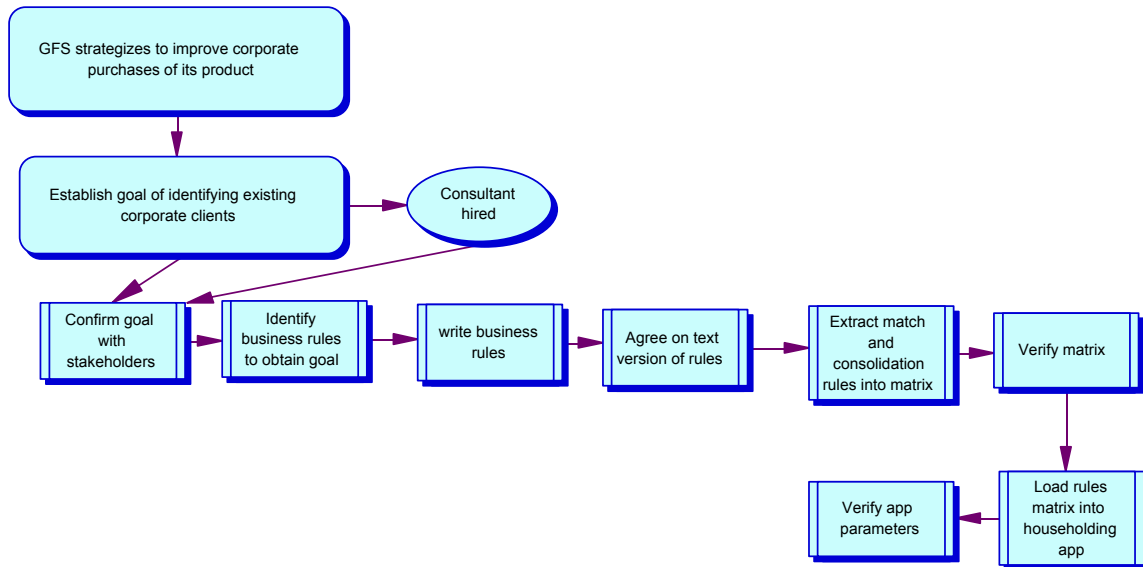


Figure 3. The pathway from corporate goals to application parameters

**Step 7 – Check rules for cohesion.** Using the job file verification feature of Firstlogic’s IQ Suite, the rules are checked for cohesion and operability. Once all warning and error messages were cleared each individual match rule was tested independently of the other rules and the results of the match run approved by GFS. It was during this phase that GFS tweaked their business rules via editing the match parameters as the results of the individual runs highlighted singularities in GFS’s data they were unaware of. Specifically, GFS adjusted the match criteria of firm names, and added *custom* firm names to the IQ Suite firm name parser to allow for the wide distribution of firm name data entry. For example, “Global Scientific” and “Global Sciences Accounting Dept” were entered as valid firm names for the same company. Only after *each* match criteria was verified for expected behavior against live (albeit duplicated) data, would the criteria be signed-off and the next criteria tested.

**Step 8 – Verify entire Match Sets.** Upon complete testing of each match criteria, the entire match sets (combinations of match criteria) were run and the results verified.

**Step 9 – Verify consolidation criteria.** In the process of consolidating duplicate records the opportunity arose to adjust the high-level business rules and even consider modifications to data structures. While match criteria identified records as duplicates, crucial differences existed in field-level data. For example the root corporate records had two child address records one for shipping and one for billing. The marketing department of GFS had been using the billing address record to store the location of the data center where the software was to be used. In reconciling the issue, a third child address record was created, so that there would be one record for each address context: billing, shipping, and data center.

**Step 10 – Run Household Process.** After all consolidation criteria were tested and confirmed for proper operation, the entire householding process was run.

**Step 11 – Completion.** In the case of GFS the data set resulting from the match and consolidation project was used as the initial load into a new CRM system. The investment of conducting essentially a “pre-cleansing” operation prior to loading the CRM package paid

dividends immediately as there were fewer records to load, the records schema had been modified and verified, and the atomic data fields (address, phone number, names) had been cleansed. The pre-cleansing heightened the quality reputation of the information in the CRM system which encourage and accelerated broad use of the system.

In completing the householding project, GFS gained a single view of the purchasing patterns of each self-contained corporate entity at the granularity dictated by their business rules. The approach employed by the consultant in conjunction with the goals and business rules of GFS accomplished the earlier assertion that a good household structure should be able to analyze individual level data within the household environment, and have a structure capable of supporting the various individual-level data as demanded by the context. The three address contexts (shipping, billing, data center) being an example. While the over arching context of GFS's householding project was the vendor-corporate client relationship, the methods employed by GFS would work for any context be it vendor – supplier, risk aggregation, residential marketing, etc.

## **6. Reasoning with Context Knowledge: Context Interchange**

The importance of context in interpreting information has been considered in other research. As a simple example, one source of information might provide length information measured in “meters”, yet the user might require or expect length information in “feet.” How can the “contexts” of these different parties be reconciled? The COntext INterchange (COIN) project [4] has addressed these needs through a mediation approach for semantic integration of disparate information sources. The set of Context Mediation Services comprises a Context Mediator, a Query Optimizer, and a Query Executioner. The Context Mediator is in charge of the identification and resolution of potential semantic conflicts induced by a query. This automatic detection and reconciliation of conflicts present in different information sources is made possible by general knowledge of the underlying application domain, as well as informational content and implicit assumptions associated to the receivers and sources. These bodies of declarative knowledge are represented in the form of a domain model, a set of elevation axioms, and a set of context theories respectively.

The COIN approach allows queries to the sources to be mediated, i.e., semantic conflicts to be identified and solved by a context mediator through comparison of contexts associated with the sources and receivers concerned by the queries. It only requires the minimum adoption of a common Domain Model, which defines the domain of discourse of the application. The result of the mediation is a mediated query. To retrieve the data from the disparate information sources, the mediated query is then transformed into a query execution plan, which is optimized, taking into account the topology of the network of sources and their capabilities. The plan is then executed to retrieve the data from the various sources; results are composed as a message, and sent to the receiver.

The knowledge needed for integration is formally modeled in a COIN framework as depicted in Figure 4. The COIN framework is a mathematical structure offering a sound foundation for the realization of the Context Interchange strategy. The COIN framework comprises a data model and a language, called COINL, and is used to define the different elements needed to implement the strategy in a given application:

- The Domain Model is a collection of rich types (semantic types) defining the domain of discourse for the integration strategy (e.g., “Length”);
- Elevation Axioms for each source identify the semantic objects (instances of semantic types) corresponding to source data elements and define integrity constraints specifying general properties of the sources;
- Context Definitions define the different interpretations of the semantic objects in the different sources or from a receiver's point of view (e.g., “Length” might be expressed in “Feet” or “Meters”).

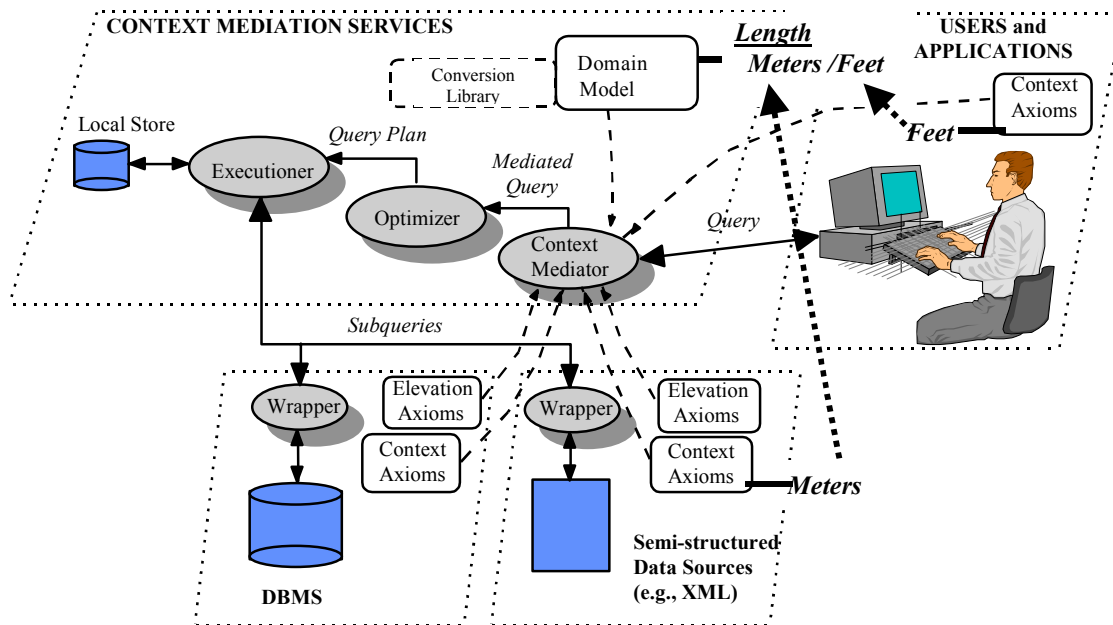


Figure 4. The Architecture of the Context Interchange System

Finally, there is a conversion library which provides conversion functions for each modifier to define the resolution of potential conflicts. The relevant conversion functions are gathered and composed during mediation to resolve the conflicts. No global or exhaustive pairwise definition of the conflict resolution procedures is needed. Both the query to be mediated and the COINL program are combined into a definite logic program (a set of Horn clauses) where the translation of the query is a goal. The mediation is performed by an abductive procedure, which infers from the query and the COINL programs a reformulation of the initial query in the terms of the component sources. The abductive procedure makes use of the integrity constraints in a constraint propagation phase, which has the effect of a semantic query optimization. For instance, logically inconsistent rewritten queries are rejected, rewritten queries containing redundant information are simplified, and rewritten queries are augmented with auxiliary information.

## **7. Research Plan**

Our plan is twofold: (1) Continue our qualitative research to document the various different sources (such as the D&B example), views, and purposes for corporate household knowledge, including the rules used in each case (such as the Firstlogic example); (2) Extend the context interchange framework to be able to represent the corporate household knowledge and rules and enable the context mediation technology to be able to correctly understand and reason about both the context of the sources and the context of the user's query.

Thus, when questions, such as "How many employees does IBM have," are asked, the answer will be the one appropriate to the questioner.

## **8. Conclusion**

Corporate structures and the corporate relationships are changing constantly. The corporate household structure is different under different contexts. Our inquiry in this research has been the understanding of what constitutes a corporate household, how do organizations utilize the concept of corporate household in their business activities, and how they adapt the concept in various tasks. Currently, we are investigating how the concepts and problems that we have identified can be matched with solutions developed in practice, and how research on an innovative technical solution can be developed through extending existing solutions, such as the Context Interchange framework.

**Acknowledgments:** Work reported herein has been supported, in part, by Firstlogic, Inc. and other sponsors of the MIT Total Data Quality Management (TDQM) Research Program.



## References

- [1] Arnold, S., Risk Managers. *Strategic Finance*, 81(12) 2000, pp. 60-64.
- [2] Ballou, D. P., R. Y. Wang, H. Pazer and G. K. Tayi, Modeling Information Manufacturing Systems to Determine Information Product Quality. *Management Science*, 44(4) 1998, pp. 462-484.
- [3] D&B (2001). *A Guide to Working with Dun & Bradstreet Family Trees*. In
- [4] Goh, C. H., Bressan, N., Madnick, S.E., and Siegel, M.D., Context Interchange: New Features and Formalisms for the Intelligent Integration of Information. *ACM Transactions on Office Information Systems*, 1999.
- [5] Huang, K., Y. Lee and R. Wang, *Quality Information and Knowledge*. Prentice Hall, Upper Saddle River: N.J., 1999.
- [6] Kotler, P., *Marketing Management: Analysis, Planning, Implementation, and Control*. 9th Edition ed. Prentice Hall, 1997.
- [7] Madnick, S., X. Chen, J. Funk and R. Wang. Corporate Household Data: Research Directions. in *Proceedings of AMCIS 2001*. Boston, Massachusetts, 2001.
- [8] Putnam, L., The American Keiretsu: America's New Competitive Advantage. *American Business Review*, 16(1) 1998, pp. 113-120.
- [9] Redman, T., The Impact of poor Data Quality on the typical enterprise. *Communications of the ACM*, 41(2) 1998, pp. 79-82.
- [10] Redman, T. C., ed. *Data Quality for the Information Age*. 1996, Artech House: Boston, MA. 303 pages.
- [11] Shankaranarayan, G., R. Y. Wang and M. Ziad. Modeling the Manufacture of an Information Product with IP-MAP. in *Proceedings of Conference on Information Quality*. Massachusetts Institute of Technology: pp. 1-16, 2000.
- [12] Shull, B., The Right Corporate Structure for Expanded Bank Activities. *The Banking Law Journal*, 115(4) 1998, pp. 65-96.
- [13] Strong, D. M., Y. W. Lee and R. Y. Wang, Data Quality in Context. *Communications of the ACM*, 40(5) 1997, pp. 103-110.
- [14] Wand, Y. and R. Y. Wang, Anchoring Data Quality Dimensions in Ontological Foundations. *Communications of the ACM*, 39(11) 1996, pp. 86-95.
- [15] Wang, R., M. Ziad and Y. Lee, *Data Quality*. Advances in Database Systems, ed. A. K. Elmagarmid. Kluwer Academic Publishers, Norwell, Massachusetts, 2001.
- [16] Wang, R. Y., V. C. Storey and C. P. Firth, A Framework for Analysis of Data Quality Research. *IEEE Transactions on Knowledge and Data Engineering*, 7(4) 1995, pp. 623-640.
- [17] Zeng, D. (2001). *Analysis of XML and COIN as Solutions for Data Heterogeneity in Insurance System Integration*. MIT.

## **Accessing the Quality of Online Classified Websites: An Empirical Study of the 100 largest US Newspapers**

Adenekan Dedeke and Beverly Kahn

Sawyer School of Management, Suffolk University, Boston, MA

**ABSTRACT:** The use of websites and web applications to advertise, inform and to execute business transactions has become a standard practice in our modern economy. Customers expect every form of organization that they interact with to develop a web presence. Newspaper organizations cannot afford to overlook this phenomenon for two reasons. First, these organizations compete in the marketplace by being the first to publish information. Since the Internet is especially fitting for speedy publication of information, news organizations of all types must to use this media. Second, which might be a more compelling reason, is that consumers are using the Internet as an information resource. It is therefore important for Newspaper publishers to integrate this media into their business.

This study focuses on how Newspaper organizations are using or not using the Internet to publish rental ads. The study of this niche is interesting because it is a profitable area of the firms. This study evaluated the web sites of the 100 largest US daily newspapers to document the level of quality of these sites. The quality criteria of intrinsic data quality (DQ), contextual DQ, accessibility DQ and representation DQ were used to evaluate these sites. The results show that while the larger organizations have better sites than smaller ones, that quality improvement of these sites is needed across the board.

### **INTRODUCTION**

This study of the 100 largest US Newspapers occurred partly because of the need to use online classifieds to assess rental properties in Boston area. It became apparent that some web sites were just too poorly designed to be useful. Due to our data quality experience, the decision to study classified ads was rather natural. It was natural to apply data quality expertise to classified ads. This study appears to be the only one focused exclusively on the quality of online classified rental properties. This study is important both from the customer's, the provider's and the web-programmer's viewpoint. It is hoped that this study would create an impetus for the discussion of programming improvements for electronic classified ads. The Newspaper Association of America (NAA) estimates that the market for classifieds ads is \$16.8 billion. Almost all of this revenue is from print classifieds. Typically classified ads make up between 35-40% of the total revenues of a daily newspaper.<sup>1</sup> The NAA also estimates that there are more than 1,000 employment, 900 real estate and 500 automotive web sites offering classifieds-type information.<sup>2</sup> From a competitive standpoint every provider of electronic classifieds on the web is competing with other providers of print-based classifieds. It is reasonable to expect that the quality of online classifieds ads could give a producer competitive advantage over other competitors. High levels of quality would also benefit users of such services since they would find their information in efficient and cost-effective ways. The results of this empirical study would describe the quality levels of online rental classifieds is a first step towards their future improvements.

## **APPROACHES FOR ENSURING AND INVESTIGATING SOFTWARE QUALITY**

The focus on software quality is becoming more and more important in all organizations that depend on software products. In the past, only the larger software firms had the resources and the interest in investing in quality improvement processes. Economic factors sometimes hinder quality assurance efforts and lead some software firms to release their products before standard quality inspection processes have been completed. This often results in costly updates, returns and bad public image. To rectify the weakness of quality inspection, many software firms have embraced quality assurance and implementation processes that rely more on the approach of building quality into all software design, development and prototyping. These tools include system development life cycle (SDLC), prototyping, and computer-based methods (Kendall and Kendall,<sup>3</sup> Merlyn and Parkinson,<sup>4</sup> Zahedi<sup>5</sup>). The Capability Maturity Model (CMM), developed by the Software Engineering Institute of Carnegie Mellon University, is another tool that firms use to integrate quality into each process of software development (Parzinger and Nath<sup>6</sup>).

While the approach of building quality into the development process has a lot of advantages, one must still measure or evaluate the quality of a software product before one could ascertain the value of a product. The task that we face in this study was to decide on the dimensions to use for the software evaluation process. There are different criteria that are in literature in regard to the topic of data, data quality, information quality and information systems quality. Levitin and Redman<sup>7,8</sup> identified several properties of data without linking the properties to software systems. Miller<sup>9</sup> describes the ten dimensions of information quality from the perspective of organizations dealing in information products. The criteria he gives therefore focuses on the more general features of information. Some researchers in literature investigated the quality of information systems from the service quality perspective (Pitt, Watson and Kavan<sup>10,11</sup>). Many of the authors that conduct service quality based views in their studies use the SERVQUAL dimensions (Parasuraman, Zeithaml, and Berry<sup>12</sup>). DeLone and McLean's<sup>13</sup> works on the quality of information systems are probably the most visible in the areas of evaluating the quality of information systems. These authors recommend two dimensions of system and information quality and several criteria for the evaluation of the quality of such systems.

A final field of research that is beginning to grow in literature is the attempt to evaluate the quality of websites. Examples of work in this area include tools such as WEBQUAL, with its two dimensions of attractiveness and informativeness (Barnes and Vidgen<sup>14</sup>) and the work of Katerattanakul, P. and Siau, K.<sup>15</sup> These focus on the quality of general web pages rather than on specific web application. The present research will therefore contribute to research work that seek to describe the quality of web application, in particular web applications that Newspapers use for their rental services. Though WEBQUAL was developed for a web application for bookstores, the present work would be developed based on the quality dimensions that were developed by Wang and Strong (1996)<sup>16</sup>, rather than on WEBQUAL dimensions. There are two reasons for this choice, first the former dimensions were developed using rigorous investigation of IS managers. It also includes a finer differentiation of web application quality dimensions than the two dimensions that WEBQUAL instrument offers.

## **RESEARCH APPROACH**

This exploratory research is seen as the first segment of a two-part study. Initially, this research targeted rental classifieds. The choice of the rental ads was made so that we could exploit our experiences with such web applications even though several other categories of classifieds exist, such as automotive, real estate sales, employment classifieds. It is assumed that

availability of multiple "standard" criteria for rental ads would make it easier to examine first. A laboratory research approach was selected because it would allow us to test and rate all classified ads (the research objects) in the same way. Additionally, a laboratory study approach could be completed faster than a mail-survey and would therefore yield time and cost-savings.

The determination of the size of the demographics of the research sample was also an important research decision. Based on the data of the NAA, in 1999 there were 736 morning newspapers and 760 evening newspapers in the US (a total of 1,483 after compensation is made for those producing both morning and evening papers).<sup>17</sup> In April 2000, it was estimated that about 1,200 US daily newspapers had established online services.<sup>18</sup> However, many of these daily news had no online classifieds in the true sense of the word. NAA found that only about 70% of its members have online classified advertising.<sup>19</sup> In this study, the focus would be on morning newspapers since they have the higher level or scope of classifieds. Based on the estimation that there are 1,200 websites, we estimate that about 50% of these would be for morning newspapers. We estimate that our potential research population is about 600 classifieds ads sites. After the available data was analyzed, it was decided to focus on the largest 100 US newspapers instead of a statistical random sample. Two facts weighed against the use of a random sample. First, about 75% of all daily newspapers in the US have a daily circulation of less than 50,000. This implies that a representative random sample would have a high proportion of newspaper organizations with small circulation. However, evidence in literature suggests that these newspapers are less likely to have online classifieds due to limited resources and skills.<sup>20</sup> Since this study is designed to document the practices of existing web sites, and newspapers with larger circulation could be expected to have such web sites, the decision was made to focus on the 100 largest daily newspapers in the US.

These 100 largest newspapers studied cover 38 out of the 50 US states. The largest newspaper in the study had daily circulation in 1999 of 1,671,530 and the smallest newspaper had a circulation of 101,948 (1999). Table 1 shows a breakdown of the visited web sites. Only 96 of the web sites visited had valid or research relevant data. Two web sites were reserved only for members. One site had no classified ads while another had no data at all. The research population was therefore reduced to 96 classifieds websites. Table 2 shows the structure of the parent population from which the research sample was drawn. It shows that 92% of the US market has a circulation of 100,000 or less. . Our research sample focuses on the high end (top 7.07%) of the US market.

Table 1: Break-down of the web sites visited

Restricted web sites	2 (2%)
Web sites without classifieds	1 (1%)
Incomplete web sites	1 (1%)
Final research sample	96 (96%)

There are at least two approaches for a laboratory test of a web site. First, one could compare each web site to an ideal standard. This approach was not used since we do not have an ideal standard. Second, one could subject all the tested web sites to a form of relative benchmarks. The reference standards used in this study do not represent an ideal, rather they constitute functions that leading providers offer and have been accepted by an evaluator as part

of a comparative tool. This latter approach is used in this study. The content of the instrument (comparative tool) is described in the next section.

Table 2: Description of parent population: Structure of the US daily newspapers market (1999)<sup>21</sup>

Circulation	Under 50,000	50,001-100,000	100,001-250,000	Over 250,001
Number of newspapers in category	1,244 (83.88 %)	134 (9.03%)	65 (4.38%)	40 (2.69%)

### DEFINING QUALITY DIMENSIONS AND DESIGNING THE SURVEY INSTRUMENT

The issue of defining data quality has been prominent in literature in the past decade. As could be expected each information media has its own peculiarities in regard to how quality is defined. In this section of the report, relevant data quality literature, which develops a quality framework, is summarized in this section. Definitions from previous work were adapted for this study. First, data quality is defined as the set of characteristics that describes the degree to which information meets or exceeds the expectations of users.<sup>22</sup> A data quality dimension is defined as a group of quality characteristics whose components manifest associative properties. The association may be due to the similar manner in which users respond to the characteristics in a group<sup>23</sup> or based on the relationship of the characteristics to a third global factor. In defining a data quality framework for this study, reference is made to previous work by Wang and Strong (1996). These authors identified four quality dimensions in their research: intrinsic data quality (DQ), contextual DQ, accessibility DQ and representation DQ. These dimensions focus on four key issues and questions. The interpretation of these issues and questions for this work is stated as follows:

*Intrinsic DQ:* What degree of care was taken in the creation and preparation of information?

If a high degree of care was taken in the preparation of information it could be expected to have accuracy, objectivity, reputation and believability.

*Representation DQ:* What degree of care was taken in the presentation and organization of information for users?

If a high degree of care was taken in the presentation and organization of data, it would be structured, concise, consistent, easy to read, interpret and understand.

*Accessibility DQ:* What degree of freedom do users have to use data, define and/or refine the manner in which information is inputted, processed or presented to them? Information with a high level of accessibility DQ would permit users to easily access modify and refine data. Users would also be able to select the most appropriate input, processing and representation approaches for their needs.

*Contextual DQ:* To what degree does the information provided meet the needs of the users? Information with a high level of contextual DQ would add value, be relevant, complete, timely and appropriate for decision-making.

While the same data quality dimensions presented above was adapted for our research, the criteria used under each dimension were specifically developed for classified ads web sites. The development of the instrument has two parts. First, a model of the process of creating online classifieds was prepared. Table 3 shows each of the stages of the process as well as the possible data quality issues.

Table 3: Description of the process used by web-based classified advertising firms

Procedural Steps	Relevant Data Quality Issues
<i>Step 1.</i> Agents and owners having properties to rent visit an online site and enter the content of their classified ad in an electronic form that is submitted to a newspaper (alternately, the data could be mailed in an envelope).	This is the step that significantly defines the <i>intrinsic</i> data quality of classified ads that are put online. Errors in spelling, formatting, abbreviations and in the scope of data collected deter the intrinsic DQ of the information that is provided.
<i>Step 2.</i> The submission is processed and the ad is published if there is no credit-verification problem.	This step is part of the preparation phase. A newspaper organization could inspect and correct errors made in step 1 here.
<i>Step 3.</i> Individuals seeking rental properties visit the classifieds section of an online newspaper to preview the available ads.	This phase reveals the accessibility of a web site. If the classified section is malfunctioning most of the time or not well positioned on a web site, or is located on an inaccessible server, the accessibility quality may suffer. A special case of “low” <i>accessibility</i> DQ occurred in this study in the case of certain newspapers that have their online classifieds under password-protection.
<i>Step 4.</i> Visitors enter information about the property that they want to preview (if the web-site permits such flexibility otherwise the process continues with step 7).	This step exposes the degree of contextual quality of an online classifieds section. High <i>contextual</i> DQ would mean that there are several different kinds of search fields and criteria that could be used to <i>define the ads</i> that the user is seeking. It would also imply that the user has some flexibility in determining the <i>method</i> used to search and the <i>scope of areas</i> to look for the defined matching ads.
<i>Step 5.</i> The results of the search is presented to the user and tools are made available to ease the browsing process	This step influences the level of <i>representation</i> DQ that an online classified section offers. The provision of ads with high level of precision in their contents, e.g. no mixing of 2 bedrooms with 3 bedrooms, ads with high uniformity and graphing tools and pictures all enhance the representation DQ.

<p><i>Step 6.</i> The visitor is offered tools that would permit him/her to manipulate, refine or modify the kind of information, the format and/or the scope of matching results that the application presents.</p>	<p>This step reveals the <i>accessibility</i> DQ of the search results. Static and inflexible data results have low accessibility DQ, while those that can be sorted, refined and easily manipulated by users, have high accessibility DQ.</p>
<p><i>Step 7.</i> User views the presented matching results (pre-selected or user-defined) and decides on the next step of action.</p>	<p>This step shows another side of the accessibility DQ. The provision of tools for printing, saving and e-mailing results or contacting the individual that paid for an ad, would enhance value and the accessibility DQ of results to users.</p>

This study focuses on the data quality issues occurring in Steps 3-6. Steps 1 and 2 were omitted since a different approach that concentrates on the identification of user errors and the study of the data entry process would be required. Step 7 was also excluded for the obvious reason that we expect that most Internet users would have the means and the know-how to email, save and print the results of their search. The survey instrument for this study was therefore designed to cover the dimensions of contextual, representation and accessibility DQ.

Table 4 summarizes the variables that were included in the data collection instrument. The variables are self-explanatory for the most part. Under contextual DQ, the instrument was used to gather information about the manner in which users define rental ads, the means by which users define their preferred location and the methods by which the computer searches for the matching ads and the scope of searches.

The accessibility DQ dimension of the instrument measures the degree of convenience with which results could be manipulated, refined as well as the ease with which one could interact with the search results.

Table 4: Description of the research instrument items/variables

Data Quality Dimension	Variables included in instrument
Contextual DQ	<p>Search methods available:</p> <ul style="list-style-type: none"> <li>- Distances</li> <li>- Places</li> <li>- Zip Codes</li> <li>- Newspapers</li> <li>- Other</li> </ul> <p>Search variables (defining property ad):</p> <ul style="list-style-type: none"> <li>- Bedrooms</li> <li>- Baths</li> <li>- Property type</li> <li>- Size</li> <li>- Number of rooms</li> <li>- Rent per month</li> <li>- Criteria priorities can be set</li> <li>- Date available</li> <li>- Realtor fees</li> <li>- Type of lease</li> <li>- Keywords</li> <li>- Date of ad listing</li> </ul>

	Search variables (defining property location): <ul style="list-style-type: none"><li>- Mixed and fixed counties and neighborhoods</li><li>- Multiple counties can be selected per search</li><li>- One county or all counties at once can be selected per search</li><li>- Multiple cities can be selected per county and search</li><li>- Multiple cities in multiple neighborhoods/counties can be selected per search</li></ul>
Accessibility DQ	Accessibility of data results: <ul style="list-style-type: none"><li>- Availability of sorting</li><li>- Number of variables that can be used for sorting results</li><li>- Highlighting and collection of ads for selective processing</li><li>- Searching through results</li><li>- Hypertext based skip-browsing (nonlinear) possibilities.</li><li>- Visibility of the total sum of matches</li><li>- Visibility of the total number of pages of printable results</li></ul>
Representation DQ	Representation quality of results: <ul style="list-style-type: none"><li>- Uniformity of listings</li><li>- Precision of listings</li><li>- Presentation of amenities on a separate page or section</li><li>- Map-based features for describing transportation issues</li><li>- Floor plans</li><li>- Availability of neighborhood information</li></ul>

---

The variables included in the representation DQ section of the survey measure the scope of data, and of the variety of the functions that the classifieds offer the user. These quality criteria and research instrument were developed based on a thorough pre-study of a few online classifieds. As a result of the pre-study, the data collection instrument was modified to include a few additional variables. The items included during the test process under the section of search methods (Q1) are: Newspapers and Other items. For Q2 the following items were added: keywords and priorities. Under Q3: the as possibility of the searching through search results was added. The updated version of the survey instrument is included in the appendix of this paper.

## **DESCRIBING THE TEST PROCESS**

The research was carried out early 2001. The following testing procedure was used for all sites:

1. Visit the web site of the newspaper and look for any information on the neighborhoods
2. Visit the classifieds section and look for the rental properties category
3. Determine if the newspaper owns the rental classifieds section or if it is a partnership web site run by a professional classifieds agency, e.g. Apartments.com
4. Create a search that would yield a multi-page result
5. Inspect results for accessibility and representation DQ
6. Finish the recording of results and move to the next web site

These basic steps were used for each site and the results were recorded on the survey instrument and later, the data was transferred into spreadsheets for further analysis.



## RESULTS OF THE STUDY

The following section provides a summary of the results of the survey. For each of the investigated functions, we would give the overall summary of the results and then control for the effects of circulation of the newspapers on the results.

### 1. *The use of partnerships (with professional firms) to provide online classifieds*

A partnership was assumed to exist only when an organization solely offers a third party web application on its site. The results showed that 54% of the web sites use the services of third parties to provide online classifieds for their users and 46% of the web sites studied owned their classified web sites. Few newspapers provided (hypertext) links to other classifieds web sites as an additional feature. This practice was not interpreted as partnership. . Table 5 provides a detailed analysis of the results with regard to daily circulation. A definite trend is obvious. The larger a newspaper is, the greater the likelihood that they would partner with a third party in creating their rental classifieds web site. The converse is also true for newspapers with smaller circulation.

Table 5: Tendency to work with third parties or partners

Circulation	Percentages
500,000 and higher	62.5%
Greater than 400,000 and less than 500,000	71.4%
Greater than 300,000 and less than 400,000	50%
Greater than 200,000 and less than 300,000	41.7%
Greater than 100,000 and less than 200,000	41.9%

### 2. *The availability of search functions on classifieds web sites*

Not all classifieds web sites allow the use of search functions. The most primitive sites use the web solely as an electronic billboard. These sites are referred to as "listing classifieds web sites" because they do not allow users interact with search results in a meaningful manner. In many cases, the user does not even have the tools to conduct a search at all. Other sites allow users to select a category or type of rental property they need, and results are then listed with no sorting possibilities. Listings are presented to users much in the same way as the hard-copy version of a newspaper listing. Overall, 66% of the sites permitted some form of search tools to the user, while 34% permitted little or no search capabilities. It seems that the sites operated by newspapers with a larger circulation tend to offer search functions than those being operated by newspapers with a smaller circulation (see Figure 6). In the next segment, we would present the kinds of search functions offered by the newspapers studied.

Table 6: Sites offering adequate search methods based on newspaper circulation

Circulation	Percentages
500,000 and higher	100%
Greater than 400,000 and less than 500,000	100%
Greater than 300,000 and less than 400,000	78.6%
Greater than 200,000 and less than 300,000	45.8%
Greater than 100,000 and less than 200,000	46.5%

3. *The scope of search methods provided for the definition of geographical locations*  
 The scope of search methods that are possible for online searches was investigated. In other words, how does a user define the geographical location of a desired rental property. The five categories investigated are: the use of distances, the use of names of places, the use of zip codes, the use of newspaper names and the “other” category. The overall results are as follows.

This study also included a detailed investigation of how the "names of places" fields are used to define geographical locations. Many websites use different levels of aggregation, such as counties, neighborhoods and city names, for defining geographical places. We define neighborhoods as an aggregation of cities which, by virtue of their proximity to each other, are recognizable as a sub-segment of a geographical region, such as North, North-East, South and so on. Counties are geographical areas that are each made up of multiple neighborhoods. This aspect of defining geographical places is described later on in this paper.

Table 7: Scope of search methods offered for defining locations (multiple responses possible)

Method	Percentages
Distances (combined with places)	4%
Use of the name of places	70%
Zip codes	5%
Newspaper names	9%
Other (e.g. street names)	2%

The results suggest that the name of places which could be cities, neighborhoods, counties or mixture. Searching based on name of place is the most popular method used by online classifieds. The use of this kind of method is effective, especially when a map of the area is provided for the user. All the other items such as newspaper names, zip-codes, distances and all other items listed are less common on web sites. The results provide some evidence that very few sites combine the use of name of places with other methods such as distances.

One advantage of using distances in combination with name of places is obvious. Users who view their commute as part of their decision criteria can make a more informed decision than would be possible with the use of name of places alone. The use of newspaper names and zip codes would primarily benefit users who are knowledgeable about the region in question. The analysis of the results with respect to size did not reveal any interesting differences between small and large organizations. The next section presents information about the methods that are provided for defining rental objects.

4. *The scope of fields provided for the description of rental ads or objects*

An important feature of every online classified section is the extent to which a user can define the rental ads to be listed. The results of this research reveal that most sites concentrate on a few variables (see Table 8). The "property type" variable (85%) occurs most often on all web sites. The uses of fields for “number of bedrooms” (63%), “rent per month” (60%) follow in second and third places respectively. Only 54% of the web sites permit users to search based on the number of bathrooms. 45% of the sites offer the use

of "keyword fields" and 43% permit the "date of ad posting" as a defining criterion. All the other possible fields included in the instrument were seldom used.

Table 8: Scope of search variables offered (multiple responses possible)

Search fields	Percentages
Property type	85%
Number of bedrooms	63%
Rent per month	60%
Number of bathrooms	54%
Keywords fields	45%
Date of posting of an ad	43%
Set priority	9%
Size of property (sq. ft.)	1%
Date available	1%
Number of rooms	0%
Realtor fees	0%
Type of lease	0%

Further analysis of the results revealed that the "keyword fields" and the "date of ad posting" often occur together. Furthermore, we discovered that the newspapers with the smaller circulation used these two features more frequently. 60% of the newspapers with a circulation of less than 200,000 used keyword search fields and 46.5% of them used the "date of ad posting" as a listing criterion. Altogether, we conclude that the scope of search fields provided in 40% of the sites investigated is of lower contextual quality.

5. *The scope of tools provided for the limiting of search neighborhood of rental property*  
 Table 9 describes the degree of flexibility that is possible in defining/limiting the scope of geographical search. The ideal scenario here would be a classified site that permits a user to select a state first and then allows them to select specific neighborhood(s) they want from a list of all possible areas. (If this is the first time neighborhood is used, it needs to be specifically defined in the section that describes geographical searching.) Finally, the system would permit a user to select the particular cities in each neighborhood that should be included in the search (i.e., from a list of all cities in the neighborhood chosen). Overall, most online classifieds do not offer their users the highest possible flexibility that is the possibility of selecting multiple cities in multiple neighborhoods per search. Table 9 shows the scope of tools provided for limiting the definition of neighborhoods. For this section, multiple responses per web site were possible. About one half of the sites do not offer their users the opportunity to search by cities or counties. Those that do, often create their own "self-defined" regions, which may even mix cities and counties together or combine two or more neighborhoods/counties into a category. Fifty four percent of the sites permit users to select multiple neighborhoods counties per search. Only nine percent of the web sites offer the highest flexibility of multiple cities and multiple counties per search. Though the quality seems to be low across the board, larger firms (with a daily circulation of more than 300,000) offer better flexibility than do the smaller ones in searching by geographical area.

Table 9: Scope of search fields used for describing neighborhoods (multiple responses possible)

Search fields	Percentages
Mixed and fixed counties and neighborhoods	52%
Multiple counties can be selected per search	54%
Only one county/neighborhood or all neighborhoods can be used	7%
Multiple cities per county can be used	0%
Multiple cities in multiple counties can be used	9%

6. *The degree of accessibility of search results*

When online classified ads produces search results, the output does vary with respect to the degree of accessibility of the results to users. Accessibility, in this context, describes the scope of the tools that are provided to enable a user to interact more meaningfully with search results that are generated after a rental property search. Tools such as sorting functions, page skip-browsing and selective processing tools were investigated. Tables 10 and 11 summarize the accessibility of the websites visited. The results in Table 10 show that about 38.5% of all sites do not sort results that are presented to users. According to the Table only 13% sort with 3 or more variables. Sixty three percent of the sites have sorting functions out of which forty eight percent of the websites tested only allow automatic sorting of the results using one variable. The fields that can be used for sorting range from very useful ones, such as rent and number of bedrooms, to less helpful ones such as alphabetical sorting of results. Alphabetical sorting offers a low level of accessibility but it is better than cases in which results were not sorted at all. In summary, about 61.5% of the sites investigated offered little accessibility in regard to sorting functions. Twelve and a half percent of the sites belong to the best class, which provides customizable sorting processes with three or more variables. When we factored in the circulation size, there were very little discernable differences between the smaller and the larger Newspapers.

Table 10: Sites offering sorting functions for results

Degree of sorting provided	Percentages
No sorting is done	38.5%
Sorting with 1 variable possible	48%
Sorting with 2 variables possible	1%
Sorting with 3 or more variables possible	12.5%

There are additional features that make it easier for users to interact with search results. Functionality such as the visibility of the number of matches or hits and of the number of the results pages was documented. Furthermore, we looked for the following: the feature that permits a user to skip through pages, the feature of allows users to select and compile ads and the feature that allows users to search through their search results. The availability of these enhanced accessibility features is summarized in the next section.

The most common feature of the sites is the estimation and presentation of the number of hits (92%). Seventy three percent of the sites allow the selection and collection of listings or records for further processing. Fifty seven percent of the sites show the number of

printable pages of search results. The majority of the sites do not present their results in a manner that permits hypertext-based jumping from page to page. Only 9% of the sites permit users to search through search results. On the whole there is room for the improvement of the accessibility DQ of the classified sites investigated.

7. *The level of representation data quality of web sites*

Table 12 summarizes the items that were evaluated in regard to quality of representing search results. These criteria include the uniformity of all listings and also the precision or degree of specificity of each listing. A listing with a high degree of precision includes only one rental property and exact rental costs, for example \$ 2000 is precise while \$2000-\$3500 is not precise. This criterion is important if one plans to have a useful sorting process based on rental cost. Representation quality also includes information tools that permit users to view a map of the location of the rental property. Additionally, representation quality includes the convenience of viewing the floor plan and layouts, neighborhood information and detailed list of the available amenities of the rental property.

Table 12: Scope of representation quality features on web sites (multiple listings possible)

Criteria	Percentages
Inclusion of neighborhood information	80%
Inclusion of floor plans	30%
Mapping function	30%
Separate section/feature for viewing amenities	30%
Precise listings	4%
Uniformity of content of listings	2%

As shown in Table 10, most web sites include general information about the different communities covered by their web sites. Such information includes crime rates, school performance and demographics. This seems to be the only area that is covered by most (80%) sites. All the other components of representation quality were neglected by at least two-thirds of the research sample. On 70% of the web sites, a user would not be able to generate an electronic map of the location of the rental property. Furthermore, users are not able to see floor plans nor are they provided with a distinct feature that permits them to view the amenities provided in the rental property in 70% of the sites. Based on the very low degree of features and options provided, it seems safe to say that the representation quality area could be improved.

**SUMMARY**

This laboratory-based study investigated the quality of rental advertising section of the online classifieds of the 100 largest US daily Newspapers. Ninety-six of the one hundred classifieds web sites visited were evaluated. The study investigated the contextual, accessibility and representation quality of the sites. The results of the study are summarized below.

Newspapers with a larger circulation seem to be more likely to form partnerships for the development of the online classifieds applications. Web sites that are owned by smaller newspaper organizations seem to be more likely to offer less search functionality to their users. In other words, most of the sites investigated do not offer the combination of quantitative and

qualitative measures for the definition of search areas. These web classifieds permitted users to select a state, and then multiple neighborhoods from the state and lastly multiple cities from each of the selected neighborhoods. About half of the population investigated did not permit the use of defined neighborhoods for geography-based searches. There are significant opportunities for data quality improvement in this area.

The study also gathered data with respect to the number and kind of fields that are used to describe a rental property in a classified ad. Property type number of bedrooms and rent per month are the most frequent fields provided. The fields most sites provided for the searching of rental ads could be described as sufficient. All items described till this point define the contextual quality of classifieds sites. On a scale of 1 (excellent) to 5 (bad), we rated the contextual quality of most of the sites tested as 3.0 (average).

The level of accessibility of search results to users was also investigated. Only 12.5% of the sites offered very good sorting function for search results. Many of the sites permitted sorting using 1 variable while about 38.5% permitted no sorting at all. Navigating through the search results was neither user friendly or easy. A high degree of accessibility is demonstrated by almost all of the sites providing the number of hits per search and most of them permitted the feature of highlighting and separating interesting ads for further use. In this area of accessibility quality, we would rate most of the sites as average (3.0).

The study also tested the representation quality offered by the sites. In this area two thirds or more sites lack the expected features. Most ads did not include any section for describing amenities, for generating maps of streets, and for floor plans. On the positive side, most of the sites had some form of neighborhood information for users. The representation quality seems to be in need of urgent improvement. We rate the level of representation quality of most sites as below average/expectation (4.0).

## CITATIONS

- 
- <sup>1</sup> Zollman, Peter (1998), Newspapers find success with online classifieds, [http://www.digitaledge.org/oci\\_report/business/overview/index\\_mid.html](http://www.digitaledge.org/oci_report/business/overview/index_mid.html)
- <sup>2</sup> McCourt, Kevin (1998), Competitive Landscape, <http://www.naa.org/classified/onlclass/tye.html>
- <sup>3</sup> Kendall, K. and Kendall, J. (1995). *Systems analysis and design*, Upper Saddle River, NJ: Prentice Hall.
- <sup>4</sup> Meryln, V. and Parkinson, J. (1994). *Development effectiveness: Strategies for IS organizational transition*, NY: John Wiley.
- <sup>5</sup> Zahedi, F. (1995). *Quality Information Systems*, Danvers, MA: Boyd and Fraser.
- <sup>6</sup> Parzinger, M. and Nath, R. (2000). A study of the relationships between total quality management implementation factors and software quality. *Total Quality Management*, (11)3, 353-371
- <sup>7</sup> Levitin, A. and Redman, T. (1998). *Data as a resource: Properties, implications, and prescriptions*. Sloan Management Review, Fall, (40)1.
- <sup>8</sup> Levitin, A. and Redman, T. (1998). *A model of data life cycles with applications to quality*. Information and Software Technology, 35, 217-224.
- <sup>9</sup> Miller, H. (1996). The multiple dimensions of information quality. *Information Systems Management*, Spring, (13)2.
- <sup>10</sup> Pitt, L., Watson, R. and Kavan, C. (1995). Service quality: a measure of information system effectiveness. *MIS Quarterly* (19)2, 173-187.
- <sup>11</sup> Pitt, L., Watson, R. and Kavan, C. (1997). Measuring information system service quality: concerns for a complete canvas. *MIS Quarterly* (21)2, 209-221.
- <sup>12</sup> Parasuraman, A., Zeithaml, V. and Berry, L. (1988). SERVQUAL: a multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing* (64)1, pp. 12-40.
- <sup>13</sup> DeLone, W. and McLean, E. (1992). Information system success: the quest for the dependent variable. *Information Systems Research* (3)1, 60-95.
- <sup>14</sup> Barnes, S. J. and Vidgen, R. (2000). Information and interaction quality: Evaluating Internet bookshop web sites with WebQual. *13th International Bled Electronic Commerce Conference*, Bled, Slovenia, Jun 19-21.
- <sup>15</sup> Katerattanakul, P. and Siau, K. (1999). Measuring information quality of web sites: Development of an instrument. *Proceedings of the 20th International Conference on Information Systems*, 279-285.
- <sup>16</sup> Wang, R. and Strong, D. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers, *Journal of Management Information Systems*, Spring, 12(4), 5-34.
- <sup>17</sup> NAA Study (2000), <http://www.naa.org/info/facts00/11.html>
- <sup>18</sup> NAA Study (2000), <http://www.naa.org/info/facts00/18.html>
- <sup>19</sup> NAA Study (2000), <http://www.naa.org/info/facts00/18.html>
- <sup>20</sup> Zollman, Peter (1998), Newspapers find success with online classifieds, [http://www.digitaledge.org/oci\\_report/business/overview/index\\_mid.html](http://www.digitaledge.org/oci_report/business/overview/index_mid.html)
- <sup>21</sup> NAA Study (2000), <http://www.naa.org/info/facts00/13.html>
- <sup>22</sup> Kahn, B. K., Strong, D. M. (1998). Product and Service performance Model for Information Quality: An Update, In: Chengalur-Smith, I. and Pipino, L. (1998) *Proceedings of the 1998 Conference on Information Quality*, Cambridge, MA: Massachusetts Institute of technology.

# **Managing Information Quality in Virtual Communities of Practice Lessons learned from a decade's experience with exploding Internet communication**

Andreas Neus

IBM Unternehmensberatung GmbH  
Hanseatic Trade Center  
Am Sandtorkai 73, 20457 Hamburg, Germany  
andreas.neus@de.ibm.com

Practice-oriented paper

**Abstract:** This practitioner paper examines why the rise of computer-mediated communication, driven by dramatically lowered cost, creates new structural problems from an information quality standpoint. The paper reviews how the new economics of information enable a new paradigm of collaboration and presents an evolutionary approach to collaborative content creation as a way to address information quality in virtual communities of practice. Based on experience gained in community projects, a few simple steps toward improving the quality of information in virtual communities of practice are presented and illustrated.

## **1. Introduction**

*Virtual Communities of Practice* are becoming more important as a means of sharing information within and between organizations.<sup>1</sup> While there is widespread agreement that controlling the quality of information exchanged is critical to the community's success (triggering either a vicious or a virtuous cycle) much of the published work has been focused around advances in technological or theoretical concepts for semi-anonymous<sup>2</sup> mega-sites such as amazon.com or ebay.com<sup>3</sup>, with users potentially in the millions. However, those may well be the wrong kind of approach to address the issue of information quality for the increasingly important, more closely knit Communities of Practice that companies are initiating to speed up cross-functional knowledge sharing. As a result, we have observed that for smaller-scale virtual communities, the issue of information quality is addressed in one of three typical ways:

1. *It is ignored altogether:* everyone posts and there is little structure or quality control, making it difficult to re-use knowledge.
2. *It is over-controlled, resulting in tunnel vision:* every communication addressed to the community has to be approved by a leader or moderator first.
3. *It is buried inside an unwieldy tool:* the lively discussion that is the essence of a Community of Practice is squelched by a tool that was designed for information storage and retrieval, not for discourse and collaboration.

---

<sup>1</sup> See Thomas Stewart & Victoria Brown (1996).

<sup>2</sup> if only due to the sheer number of contributors.

<sup>3</sup> See Jakob Nielsen (1999) for a good overview.



This paper attempts to give a general overview of some key issues faced by practitioners addressing the quality of information in virtual communities and showcases some solutions tried over the years on the Internet to solve information quality problems. Finally, some experience gained in projects charged with planning and setting up virtual communities over the course of the past decade is distilled into a few simple recommendations for raising the information quality in virtual Communities of Practice.

## 2. Networks vs. Hierarchies: The appeal of using Communities of Practice

Traditionally, we have used organizational systems based on hierarchy and authority to identify high-quality information. But with the world around us accelerating in its need for both time-critical and high-quality information, these traditional means are often no longer sufficient. For whereas traditional, hierarchical organization structures are very good at getting tasks done in a "divide and conquer" paradigm, the success of companies today increasingly depends not on *dividing the work*, but rather on *sharing the knowledge*. Yet for facilitating the free exchange of knowledge, networks are inherently better equipped than hierarchies. From an information sharing standpoint, a hierarchical, tree-like organization is a worst-case scenario because it is a collection of bottlenecks: There is only one "official" path between any two nodes in the graph and the likelihood of people sharing information can drop as a function of their distance in the corporate org-chart.

Faced with this dilemma, advanced companies have started overlaying their primary "command-and-control" structure not only with a subject-driven layer, forming a matrix, but also with "Communities of Practice": semi-formal networks of practitioners who exchange information on a common subject or problem of interest. They are alternately called "Competencies", "Communities", "Knowledge Networks", "Special Interest Groups", etc. and run all over the organization's chart (and sometimes even beyond a single organization) to facilitate the exchange of information and lessons learned among those who are dealing with a common set of problems or issues. That a *network* is extremely efficient for passing information is a phenomenon which has been studied and confirmed numerous times under the names "small world effect" or "six degrees of separation"<sup>4</sup>. But with a network where every member is also an instant publisher, a new challenge emerges.

## 3. Problem Outline: The vanishing cost of communicating

Before the proliferation of the Internet technology, there was a simple filtering system that kept the quality and relevance of transmitted information mostly above a certain threshold. It was called *cost*. Because every copy of information and every movement of that copy was tied to physical matter, it always incurred real cost that someone had to shoulder.

Therefore, only such information which was deemed by *someone* to be worth spending money on copying and distributing, had a fighting chance of ever being seen by more than a trivial number of people. Yes, there have always been tabloids whose information quality, when measured against objective criteria, did not do terribly well, but in terms of a *fitness for use* definition, this content still had to have a certain "quality", because people actually paid to read

---

<sup>4</sup> See Milgram (1967).

it. And while there have always been "nut cases" creating their own newspaper from their flat on a copying machine, your chances of ever coming across that content fell dramatically (probably to the tune of  $N^2$ ) with your physical or social distance from the source.

So traditionally, a piece of information had to pass through several layers of weeding, selecting and editing before it would get a fighting chance to come to the attention of a wider audience. Publishers and advertisers were effectively making a bet, with their very real money, on the quality (again in terms of *fitness for use*) of the content being suitable for its intended recipients, because every copy of information incurred real costs. Therefore, it *used* to be good business to limit the proliferation of your information to coincide with your intended audience, i.e. the target segment or market. And the better you were at addressing only your target segment with the information (and in turn designing the information to be relevant to the recipients) the less money you wasted. This incentive brought about editors, proofreaders, market analysts, etc.

The Internet, however, dramatically lowered the cost of copying and distribution – to practically zero. The last time such a major drop in the cost of information proliferation occurred was probably Gutenberg's invention of movable type<sup>5</sup>. As with Gutenberg's invention, this more recent drop in the economic cost of distributing information has created major shockwaves around the world. The success of the Internet as the undisputed global communications medium of the future and the rise (and often fall) of countless businesses built around it are an example of this kind of shockwave.

#### **4. Spam: A new word for irrelevance**

The vanishing cost of communication, coupled with commercial interest, has given rise to a bane of modern society, called "Spam"<sup>6</sup>. Whereas today "Spam" is mostly understood to mean "unsolicited commercial e-mail" or "unsolicited bulk e-mail" (the *electronic* junk mail that clogs your mailbox), it originally came up in the early nineties as a term for "massive crossposting": posting the same article to several thematically unrelated newsgroups on Usenet. When companies first stumbled upon the mostly academic global discussion network that drove the growth of the Internet in the pre-Web years, some were lured into abusing the infrastructure for posting commercial messages into thousands of newsgroups by the same economics that sustain e-mail spamming today: The cost incurred *by the sender* for sending information to be viewed by one recipient or 1,000,000 recipients, is practically identical, and in both instances negligible. This is because the bulk of that cost is carried by the providers of the infrastructure and the recipient, *not by the sender*. As Vint Cerf puts it:

*"Spamming is the scourge of electronic-mail and newsgroups on the Internet. It can seriously interfere with the operation of public services, to say nothing of the effect it may have on any individual's e-mail system. ... Spammers are, in effect, taking resources away from users and service suppliers without compensation and without authorization."*

---

<sup>5</sup> Not only did Gutenberg make it possible for Luther to copy his translation of the Bible, something that the copyshops of the day in central Europe - catholic monasteries - would not have approved of, it also created a whole new economy around typeset books and later newspapers.

<sup>6</sup> Named after the Monty Python sketch featuring the recurring canned meat product. Crossposting in the newsgroups amounted to reading the same message over and over and over again. See Eric Raymond (1993) for more information.

Certainly the most famous first instance of low-quality information massively posted to Usenet was the "Green Card Lottery" Spam perpetrated by two lawyers who decided to advertise their services to thousands of newsgroups simultaneously, creating a huge backlash by people who found their time and their resources abused. At the time, of course, network bandwidth and storage was much more expensive than today and a much higher proportion of people were on the Internet through dialup connections which had to be paid by the minute. But even today, the most precious resource is still very much threatened by low-quality information: *human attention*.

With communication increasing (and a lot of economic interests around) the "signal-to-noise ratio" dropped rapidly on Usenet newsgroups and other forms of virtual communities such as web-based discussion boards, chat fora such as IRC or instant messaging services such as ICQ or AIM. For example, the number of Spam postings to Usenet each month skyrocketed from below 100,000 to 1.8 million in just two years between 1995 and 1997.<sup>7</sup>

This deterioration of information quality, particularly on Usenet over the years, has caused many people to abandon public newsgroups and seek refuge in other, less open virtual communities, because finding relevant and high-quality information had become a hunt for the proverbial needle in the haystack. This information overload is a symptom of what can no-longer seriously be called the *information society*. Rather, it could be more appropriately called an *attention deficit society*.

What information consumes is rather obvious:  
It consumes the attention of its recipient.  
Hence a wealth of information creates  
a poverty of attention.“

- Herbert Simon (Nobel Laureate Economics)

It is this scarce resource, the *human attention*, which we must learn to better manage and direct toward the high-quality, relevant information in the exponentially growing haystack of low-quality information in collaborative environments.

## 5. Brooks' Law: Quality as the result of a single mind's integrity

These information quality problems, found in many virtual communities today, appear to confirm the old saying "*Too many cooks spoil the broth*". A modern and perhaps more "scientific" version of this notion has become known as "Brooks' Law", named after Frederick P. Brooks, author of the classic book "The Mythical Man-Month"<sup>8</sup>.

Brooks talks about the inherent complexities of coordination and states that as the number of involved programmers  $N$  rises, the work performed also scales as  $N$ , but the *complexity* and *vulnerability to mistakes* rises as  $N$ -squared, in accordance with the number of communication paths required to coordinate the contributors. To achieve quality, Brooks therefore recommends a *minimum* of contributors: "*Conceptual integrity in turn dictates that the design must proceed from one mind, or a very small number of agreeing resonant minds.*"<sup>9</sup> Conversely, Brooks' Law

---

<sup>7</sup> See the Cancelmoose Page at <http://www.cm.org/> for details.

<sup>8</sup> Brooks (1975).

<sup>9</sup> Brooks (1975), chapter 4.

predicts that "*a project with thousands of contributors ought to be a flaky, unstable mess*", as Eric Raymond put it.<sup>10</sup>

## 6. The new paradigm: Collaboration

Looking at some of the more prominent results of virtual collaboration, it becomes obvious that Brooks' Law cannot be the only force at work here. It used to be common knowledge that high-quality *software* could only be produced by a handful of highly skilled experts who are organized in the dedicated, hierarchic environment of big companies and headed by professional management.

However, during the 1990s, a radical new paradigm of collaboration, enabled by Internet technology, emerged, which seems to violate (or maybe naively ignore) Brooks' Law. This new paradigm has become known as the *Open Source development model*<sup>11</sup> and has brought about the creation of the successful free Linux operating system, which has been called "the impossible public good". Public Goods Theory predicts that a product which can be used by anyone, irrespective of whether they contributed to its creation or not, would never get created in the first place as everyone would attempt to free-ride. Instead, Public Goods Theory postulates that something like Linux could only be created with public money or by a government. But the dramatic drop in transaction costs suddenly allowed for the collaborative creation of such high-quality *software*.

Common knowledge still holds today that high-quality *information* can only be produced by a handful of highly skilled experts who are organized in the dedicated, hierarchic environment of universities or research centers and led by professional management.

But by using the same dynamics that made the "impossible" open source goods possible, people are already busily attacking this notion as well, suggesting that a loosely knit network of skilled amateurs can produce comparable or better quality information in a collaborative paradigm than traditional solitary authors, institutions or publishers are able to create.

As an illustration, consider the discussion around *PublicLibraryOfScience.org*, an initiative that is calling on journal publishers to hand control over published articles back to the scientific community after 6 months. Publishers have typically argued that information quality and integrity of publishes research can only be assured if they remain in exclusive control of the information. Practice, however, tells a different story: When some previously published articles were moved to public Internet repositories, several errors that had gone undetected during the original publishing were found and corrected, thereby *increasing* the quality of the information. As David Lipman, director of the National Center for Biotechnology Information, states: "*The more eyes to look at it and fingers trying to work with it, the more things you can find.*"<sup>12</sup>

## 7. Linus' Law: Quality as the result of massive collaboration

How was the new paradigm, involving massive collaboration, able to overcome the limits postulated both by Brooks' Law and Public Goods Theory? The explanation, again, lies in

---

<sup>10</sup> See *Revenge of the Hackers*, in Raymond (2001).

<sup>11</sup> See *Open Source Initiative* at <http://www.OpenSource.org>.

<sup>12</sup> See the Scientific American article by Karow (2001).

changing economics of information and has been called "Linus' Law", honoring Linus Torvalds, the former computer science student who spearheaded the Open Source development model. Linus' Law is usually stated in its informal version, which resembles Lipmans statement above: "*Given enough eyeballs, all bugs are shallow*".<sup>13</sup> The key to the success of the collaborative development model is based on the lowered transaction cost for information, allowing the separation of the *identification* and the *solution* components of quality problems and spreading both tasks over a much, much larger population than could sensibly be done in traditional hierarchic approaches. Says Linus: "*Somebody finds the problem and somebody else understands it. And I'll go on record as saying that finding it is the bigger challenge.*"

## 8. The Wiki Concept: Quality is what survives evolutionary pressure

The Wiki<sup>14</sup> concept is an example of taking aforementioned collaboration paradigm to its extreme by practically eliminating any transaction cost in changing or correcting information. On a Wiki web-site, *anyone* can view and edit *any* page, without any prior clearing process by an editor or moderator. There is nothing to stop a malevolent user from deleting passages, or even whole pages, of existing information, or just adding complete nonsense. At first encounter, especially in the context of information quality, this concept appears to be a recipe for dramatic failure, an information quality disaster just waiting to happen.

Interestingly, we are still waiting for the disaster to happen – and it is nowhere in sight. The key is that although *any* user can change any page, the changes are stored in a log and *any other* user can review that log and instantly undo any change that he or she does not approve of. Using this deceptively simple safety net, the Wiki concept can be a very powerful accelerator for collaboratively creating and improving information.

As one example, consider *Nupedia.com*, a project dedicated to creating a freely available<sup>15</sup> encyclopedia online. In accordance with their goal of high information quality, Nupedia.com adopted the traditional review process of publishers, where a (volunteer) author would first write an article and then submit it to Nupedia for review – a cumbersome process that resulted in only very few articles being contributed. When Nupedia.com still had only 20 articles to show for 18 months of operation, the founders realized that they had a problem and looked around for a solution. They found the Wiki concept and decided to start a complementary site, *Wikipedia.com*, as a hot-bed for collaboratively creating and improving articles. The best of these articles would then undergo the rigorous review process to become part of Nupedia.com. Something clicked and in merely 6 months from January 2001 until July 2001, Wikipedia.com has generated over 6000 articles, including many of very high quality, using this extreme interpretation of the collaborative paradigm. Instead of falling victim to vandalism, as might be expected, the site's wide-open concept quickly turned it into a thriving generator of information.<sup>16</sup>

---

<sup>13</sup> The official version is as follows: *Given a large enough beta-tester and co-developer base, almost every problem will be characterized quickly and the fix obvious to someone.* This emphasizes the separation between characterization and solution of quality problems as a means to achieve greater efficiency. See Raymond (2001).

<sup>14</sup> For more information, see <http://www.c2.com/cgi/wiki?WelcomeVisitors>

<sup>15</sup> Nupedia.com content is licensed under the GNU Free Documentation License.

<sup>16</sup> For more information, see <http://www.kuro5hin.org/story/2001/7/25/103136/121>

What drove this astonishing result? In my view, the key to the demonstrable success of the Wiki concept is based on two pillars. The first is the elimination of practically all transaction costs for collaboration. Instead of informing an editor of a change you'd like to see and talking him into accepting it (possibly taking many exchanges back and forth), a Wiki system lets you make the change yourself, on the spot, with minimum effort. The second pillar is the creation of an *artificial information economy* as a context for collaboration, which discourages low-quality or offending input, *because it is much "cheaper"<sup>17</sup> for person B to undo the low-quality change that person A caused, than it is for person A to cause it.* This process weeds out low-quality information in an evolutionary paradigm. As Richard Dawkins puts it: *"Life is the result of the nonrandom survival of randomly varying replicators"*. The evolutionary paradigm has been demonstrated to be so potent that it can create order out of apparent chaos even based on *random* mutations, given a *nonrandom* selection. The evolution in terms of the collaborative Wiki concept has the additional benefit of the changes being anything but random.

As an analogy, imagine a new method by which any passer-by could undo a night's work by a graffiti "artist" simply by snapping his fingers, if he thought that the house looked better the way it was before, without the new "decoration". What would the effect be?

1. There would be very little incentive for people to create low-quality graffiti "content", because they have to labor for hours, only to have their effort casually nullified by the next person to walk by.
2. The content that *survives* review by many people over a long period of time is likely to be of high quality, in the sense that there is widespread agreement that the wall looks better *with* the new graffiti than it did without it.

## 9. "It's the economy, stupid!"

A key problem with hierarchic approaches to information quality is that they don't scale well – and you have the issue of who chooses the editors, peers or raters. Instead, the solution may lie in creating an information economy that uses an evolutionary paradigm to *grow* and *evolve* high-quality information collaboratively, rather than to have a single author *construct* it. In this information economy, there should be a high incentive for contributing and maintaining high-quality information and a disincentive for contributing poor-quality information. The challenge then becomes one of creating such an information economy that produces high-quality information. Some important elements for this to work are the following five factors:

1. Accountability for contributions as a basis for reputation
2. A thematic focus and "culture" for high quality contributions
3. A sense of trust and identity through personal profile pages
4. A common memory or knowledge repository which is developed in collaboration
5. Membership criteria to keep the level of discourse high and on topic

---

<sup>17</sup> "Cheaper" economically, i.e. in terms of low cost in attention, time or reputation.

## 10. Conclusion

The Internet has provided us with the means to effectively collaborate across time and space with vanishing transaction costs. As the success of open source software such as Linux has proven, this new paradigm has the potential to break through well-established limits. Virtual Communities of Practice have the chance of using these same economics to redefine how high-quality information is created and shared in an organization. But one of the greatest obstacles to adopting this new paradigm are traditional notions in our own minds of how information quality is achieved, limiting our thinking about the non-linear potential of collaboration.

I believe that companies, which are today faced with increasingly well-informed customers, need to actively embrace and support virtual Communities of Practice as a way to bypass information bottlenecks, to speed up internal knowledge creation and sharing, and because they need to keep up with their increasingly well-informed customers. Or, as The Cluetrain Manifesto postulates: "*Because markets, unencumbered by corporate bureaucracy and the need to ask permission at every turn, are learning faster than organizations.*"<sup>18</sup>

## 11. Outlook: Networks of knowledge

The future of addressing information quality in virtual communities may well lie in supporting collaboration by mapping and analyzing the *underlying social networks*, revealing the now mostly invisible links between people and communities. To some extent, this is already being done, i.e. by the Google search engine<sup>19</sup>. Some advanced research on analyzing social networks to assess information quality is also carried out in the CLEVER Project at IBM's Almaden Labs<sup>20</sup>.

The essence of the mapping approaches is to analyze the micro-decisions made by people pointing to resources and to aggregate this information over a large number of people to derive information quality measures based on implicit human decisions. Besides opening up an exciting new way to tap human expertise for determining quality, these approaches also bring up a new class of challenges, especially in the area of privacy, that have only recently received attention.<sup>21</sup>

## Appendix: A brief community cookbook

Feedback loops are important for efficiently producing high quality information. Therefore, the system that allows for better feedback is the one with the potential to provide the better quality. The Internet technology provides us with such a system for much faster – and much broader – feedback loops than were previously possible. But the new technology is only the *enabler* for the virtual Communities of Practice, facilitating collaboration and feedback. The *driver* for the new paradigm is the organization, trust, commitment and interaction between the community members. This "soft" or *human* side often turns out to be the trickier part. Therefore, here are a few suggestions to help managing these softer issues.

---

<sup>18</sup> Levine et al. (2000).

<sup>19</sup> [www.google.com](http://www.google.com)

<sup>20</sup> See Kleinberg (1997) and Kumar et al. (1999) for more details.

<sup>21</sup> See Lada A. Adamic & Eytan Adar (2001).

## **Five simple steps**

Here are five simple steps you can take to support a strong collaborative culture and improve the quality of information in your virtual Communities of Practice.

### **1. Accountability: The prerequisite to reputation**

When someone makes a change to the knowledge, this change must be tracked so that there is accountability, i.e. the actor can be adequately credited with the cost or benefit of the change, and in order to allow selective reversing of changes. This way, other community members can make intelligent choices regarding how they spend their attention with respect to this user.

*Impact on information quality:*

As an example, I may decide whether or not to review a contribution or a change to the pool of knowledge depending on whether the actor has produced high-quality information in the past. While "blacklisting" is difficult in a possibly pseudonymous virtual environment, accountability is still very important for "whitelisting" – because a positive reputation is an asset that the owner has an incentive to protect.

*Recommendation:*

Addressing accountability can range from something as simple as making sure everyone has to log in with a username and password before contributing and keeping a log-file to highly complex rating and reputation systems.

### **2. Focus and culture: A community charter**

A charter including clear rules on what behavior is expected and what may be done with the content created. One key to achieving high quality is to realize that a good community of practice is a self-regulating entity that will improve the information quality by peer-pressure. Traditions and customs governing what kind of information is accepted will develop and these will be enforced by the members as part of the community's culture.

*Impact on information quality:*

A charter sets the tone for the discourse in the community of practice. It should be created jointly with the community members to ensure adoption. A strong culture on what kind of quality is expected from the information will go a long way to ensuring, via peer pressure, that the quality of contribution remains high. Explicit rules on re-use of posted information outside the community are necessary for a feeling of trust and comfort to develop, where people are willing to ask "stupid" questions or go out on a limb. I.e. "no external re-use without asking permission from the author first" could be such a rule. Because a community's membership changes, its activities ebb and flow, and its leaders change, a charter is a good way to provide a scaffolding that does not depend on individual members.

*Recommendation:*

Write at least a draft charter for the community that sets a standard for behavior, expected quality, and in which circumstances information created in the community may be used outside it. For inspiration, have a look at some examples of such manifested traditions, like



the famous "Netiquette" texts that are posted to *news.announce.newusers* group<sup>22</sup>. They explain customs and the reasons behind them to new members. Other examples of standards for quality-checking information are the humorous "Crackpot Index"<sup>23</sup> circulated in the *sci.physics* newsgroups and the "Gullibility Virus"<sup>24</sup> warning.

### **3. Trust and Identity: Personal Profile Pages.**

Trust is a problem in virtual environments. The trust that forms very easily in face-to-face meetings is much harder to achieve when all you know about the other members is their e-mail-address. A key step to creating the trust and sense of identity necessary for a thriving community of practice can be taken by providing a personal profile page for each member and encouraging its use. This page should include a picture, some self-description, and room for (links to) other resources relevant to the member's professional and maybe personal life, which they wish to share.

*Impact on information quality:*

These profiles facilitates the exchange of ideas and the creation of trust between members, as they allow people to get a better concept of the other person's expertise and interests. Resources linked to from a profile are directly available to the other members, without having to ask and wait for an answer. It ensures that members get a good idea what the skills in their community of practice are and who to turn to with which kind of question.

*Recommendation:*

Make sure your community tool supports such personal profile pages and encourage members to use them. When a few key members present themselves in this fashion, the others usually follow suit.

### **4. Collective Memory: FAQs as efficient knowledge repositories**

FAQs are a very powerful way to distill lessons-learned in virtual communities. They were originally invented out of sheer necessity: With the rising popularity of the Usenet in the late 80s came a problem: Newsgroups, which were home to communities discussing their chosen subjects, usually at a very high level of expertise, were faced with an influx of new users almost on a daily basis. This caused the ongoing discussion to be brought back down to basic, beginner's questions frequently, as a stream of new users, unaware of the discussion's history, asked the same questions over and over again. The regulars realized that they could only solve this by compiling the answers to those *frequently asked questions* in a file that could then be referenced in reply to those questions. Out of this necessity, accidentally, a very powerful didactic tool was born: In contrast to practically all other forms of documented knowledge, the FAQ is structured *not* from the perspective of the "knower", but is collaboratively created over time and structured from the perspective of the "knowledge seeker". It is therefore a much more efficient way of educating people and bringing them to a common level of understanding than was available before. In a way, it is accelerated education. An FAQ is usually maintained by one or more people who have an interest and

---

<sup>22</sup> See *news.announce.newusers* FAQs in the references below.

<sup>23</sup> See John Baez (1998) for details.

<sup>24</sup> This text by Robert Harris (2000) warns readers not to become multipliers for false information – in a very original way.

some expertise in the subject. Those people become a natural focal point for both questions and new answers regarding the subject at hand, starting a virtuous cycle.

*Impact on information quality:*

By providing a focal point for the community's knowledge on given subjects, people can stop reinventing the wheel and instead focus on creating the best wheel for everyone. FAQs of active communities are typically of a much higher quality than i.e. Textbooks, simply because there are so many more eyes for scrutiny and the combined know-how of the community helps to polish the text over time, instead of having just a few authors write a text that only gets revised every other year at the most.

*Recommendation:*

Create an infrastructure for maintaining FAQs and encourage members to start FAQs on their pet subjects. This way you will quickly seed the creation of efficient knowledge repositories that can quickly grow from half a page to several dozen pages in size and allow you to easily capture lessons-learned in the community of practice.<sup>25</sup>

## 5. Membership Criteria

A community lives off its peers. If you get the right people together, you start a virtuous circle that draws in more of the right people simply by word-of-mouth. If you let everyone in indiscriminately, you soon have an unfocused group of members, dropping information quality (in the fitness for use sense) and the experts, whose discussion you wanted to tap into, will drop out of sight again. This is why good clubs have bouncers who perform a very important *function* of quality control (which is not to say they always do a good *job*).

*Impact on information quality:*

The impact of controlling membership is very straightforward: Having the right members in the community goes a long way to ensuring a good signal/noise ratio. This is doubly important as few experts wish to waste their time in a community where they feel they do not have a lot to learn themselves, but rather always serve as unpaid teachers to the rest.

*Recommendation*

Think about some kind of barrier to entry. It does not need to be high, just something that keeps people with only a passing interest out. This can be as simple as asking people to send a CV or just give a few statements about why they believe they would make a valuable contribution to the Community of Practice. This encourages a self-selection that will increase the quality and focus of your community's members. For an example, observe the membership application in Howard Rheingold's successful "BrainStorms" community<sup>26</sup>. Managing the quality of members is possibly the most important single aspect and deserves thought. While doing this, consider the whole lifecycle: Individual members will become more and less active over time. Have a policy for weeding out people who have abandoned the community to avoid the sense of anonymity that comes with a community having too many members with whom no-one has communicated in a long time.

---

<sup>25</sup> To get a feeling for the vast amount of high-quality information thus captured on Usenet over the years, visit [www.faqs.org](http://www.faqs.org)

<sup>26</sup> See Howard Rheingold at [www.rheingold.com](http://www.rheingold.com)

These steps should bring your Community of Practice closer to becoming a thriving, collaborative source of high-quality information – and they are largely independent of the underlying technical infrastructure employed.

Keep in mind that a community is a social creature that cannot be "created" in a traditional sense. Rather, you need to provide the right context for a community to prosper. Therefore, you need to be very careful how much control you seek to exert. Try to let the community organize itself to the greatest degree possible, rather than trying to micro-manage it. Community leaders will emerge naturally: Those members who are perceived as the right mixture of being very knowledgeable, accessible and *active* in the continuing dialogue. Also, even in work-related communities of practice, "off topic" discussion should not be squelched, but accepted as a the necessary "social lubricant" that any efficient knowledge network needs. After all, *knowledge is human*.

## References

- Adamic, L.A. & Adar, E. (2001)** *Friends and Neighbors on the Web*. Web-site: <http://www.parc.xerox.com/istl/groups/iea/papers/web10/>
- Baez, J. (1998)** *The Crackpot Index*. Web-site: <http://math.ucr.edu/home/baez/crackpot.html>
- Brooks, F. (1975)** *The Mythical Man-Month*. Reading, MA: Addison-Wesley.
- Cancelmoose (1997)** The Cancelmoose[tm] Homepage. Web-site: <http://www.cm.org/>
- Euro Cauce (2001)** Euro Cauce. Web-site: <http://www.euro.cauce.org/en/index.html>
- Harris, R. (2000)**. *The Gullibility Virus*. Web-site: <http://www.virtualsalt.com/warning.htm>
- Karow, J. (2001)**. *Publish Free or Perish*. In: Scientific American, April 2001. Web-Site: <http://www.scientificamerican.com/explorations/2001/042301publish/>
- Kleinberg, J.M. (1997)**. *Authoritative Sources in a Hyperlinked Environment*. IBM Research Report RJ 10076
- Kumar, S.R., Raghavan, P., Rajagopalan, S., and Tomkins, A. (1999)**. Trawling the Web for emerging cyber-communities.
- Levine, R., Locke, C., Searls, D. & Weinberger, D. (2000)**. *The Cluetrain Manifesto: The end of business as usual*. Cambridge, MA: Pegasus Books
- Milgram, S. (1967)** The Small World Problem. *Psychology Today*, May 1967.
- news.announce.newusers FAQs**. Web-Site: <http://www.ibiblio.org/usenet-i/groups-html/news.announce.newusers.html>
- Nielsen, J. (1999)** *Reputation Managers are Happening*. Web-site: <http://www.useit.com/alertbox/990905.html>
- Open Source Initiative (2001)**. *The Open Source Initiative (OSI)*. Web-site: <http://www.opensource.org>
- Raymond, Eric S. (1993)** *The New Hacker's Dictionary*, 2<sup>nd</sup> ed. Cambridge, MA: MIT Press

**Raymond, Eric S. (2001)** *The Cathedral and the Bazaar*. Cambridge, MA: O'Reilly & Associates (Web-Site: <http://www.tuxedo.org/~esr/writings/cathedral-paper.html>)

**Raymond, Eric S. (2001)** *The Jargon File*. Web-site: <http://www.jargon.org> (The online-Version of the New Hacker's Dictionary).

**Resnick, P. et al. (1994)**. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work, Chapel Hill, NC: Pages 175-186

**Rheingold, H. (2001)**. The Virtual Community. Web-site: <http://www.rheingold.com/vc/book/>

**Rheingold, H. (2000)**. Join the BrainStorms Community? Web-site: <http://www.rheingold.com/community.html>

**Sanger, L. (2001)**. Britannica or Nupedia? The Future of Free Encyclopedias. Web-Site: <http://www.kuro5hin.org/story/2001/7/25/103136/121>

**Stewart T. & Brown., V. (1996)**. The Invisible Keys To Success – Communities of practice are where learning and growth happen. Fortune, August 5, 1996.

**WikiWiki (2001)**. Web-Site: <http://www.c2.com/cgi/wiki?WelcomeVisitors>

### **Suggested further reading**

**Brown, J.S. & Duguid, P. (2000)**. *The Social Life of Information*. Cambridge: Harvard Business School Press.

**Kim, Amy Jo (2000)**. *Community Building on the Web: Secret Strategies for Successful Online Communities*. Berkeley, CA: Peachpit Press

**McKeon, D. (1997)**. Moderated Newsgroups FAQ. Web-site: <http://www.faqs.org/faqs/usenet/moderated-ng-faq/>

**Oram, A. [Ed.] (2001)**. *Peer-to-Peer: Harnessing the Power of Disruptive Technologies*. Sebastopol, CA: O'Reilly & Associates.

**www.faqs.org**: A vast resource of Usenet-community generated content, including Netiquette, Emily Postnews, Newusers Introductions, FAQs, RFCs, etc.

**www.isoc.org**: The Internet Society. An authoritative source for information on "The Internet", significantly higher quality than most other sources.

## **The issue of IQ in internet-based early-warning systems for trend management**

Daniel Diemers  
University of St. Gallen HSG, Switzerland  
(Research Assistant)  
Wedoso gmbh  
(Managing Partner)

SfS-HSG, Ackerstrasse 44  
8005 Zürich, Switzerland  
e-mail: daniel@diemers.net

Practice-oriented Paper

### **Executive Summary**

This contribution presents our latest experiences with internet-based early-warning systems that are used in firms in different areas with different functionalities to facilitate trend management and risk management. In these systems, intelligent software agents are deployed to gather and analyze information and monitor sites on the internet.

A central problem in this context is to calibrate these systems in a way that data mining and analysis processes deal with all mission-critical data available, but are still able to operate with a manageable amount of data. Thus, IQ plays a central role in measuring the performance of such a system and in achieving a sound balance between quality information and information overload.

Our methodology includes a distinct framework of IQ and respective procedures and algorithms that are based on a socially biased, community-oriented perspective of IQ that has been presented in past contributions to this conference.

# “The issue of IQ in internet-based early-warning systems for trend management”

Dr. Daniel Diemers

University of St. Gallen (HSG), Switzerland  
Institute of Sociology SFS

## Structure

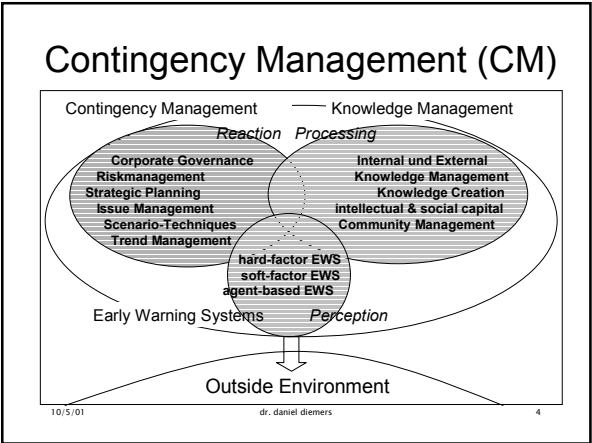
1. Research Context: Trend Management
2. Internet-based Early Warning Systems
3. Our Framework for IQ
4. Practical Experiences with IQ issues
5. Conclusions and open questions

10/5/01 dr. daniel diemers 2

## 1. Research Context

- Contingency Management, Knowledge Management, Early-Warning Systems
- Research Question:
  - „How can we measure and quantify the „quality“ of an identified Web source within an early-warning system?“

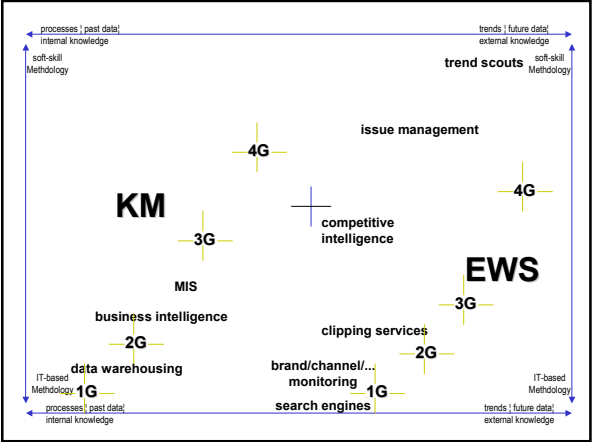
10/5/01 dr. daniel diemers 3



## Locating EWS in a CM Framework

- Internal view:
  - Knowledge Management, Data Mining, MIS, Business Intelligence
- External View:
  - Trend Scouts, Issue Management, Clipping Services, Monitoring Service, Search Engines
- soft-skill vs. IT-based Methodology
- future/external vs. past/internal data

10/5/01 dr. daniel diemers 5



## 2. Early Warning Systems

- Early Warning Systems shall help firms in their perception of the contingent corporate environment
- Early Warning Systems will play an increasingly important role in coping with the volatility and dynamics of markets

10/5/01

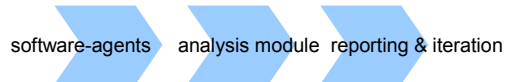
dr. daniel diemers

7

## a feasible EWS Methodology

- 3 Phases:
  - Focused Crawling: Community Topography
  - Scanning/Monitoring/Analyzing
  - Border Control and Topography Updating

- 3 Modules:

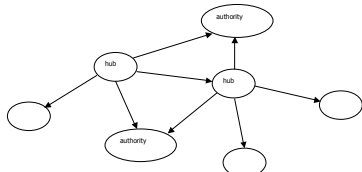


10/5/01

dr. daniel diemers

8

## The Community Perspective



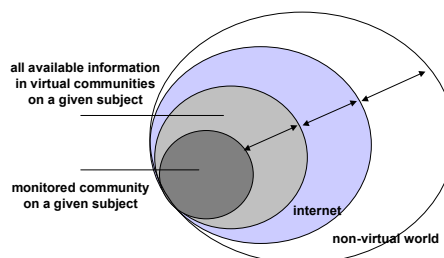
- Diemers Theorem: "only information that is referenced to or embedded in the respective community can develop any potential relevance".
- Identify Hubs & Authorities within topical communities

10/5/01

dr. daniel diemers

9

## Why Community Knowledge?

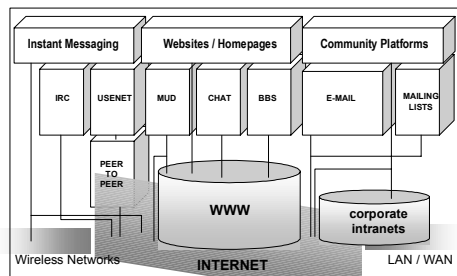


10/5/01

dr. daniel diemers

10

## Modelling Virtual Spaces

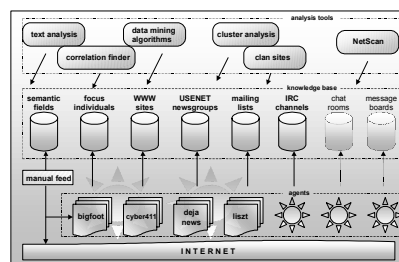


10/5/01

dr. daniel diemers

11

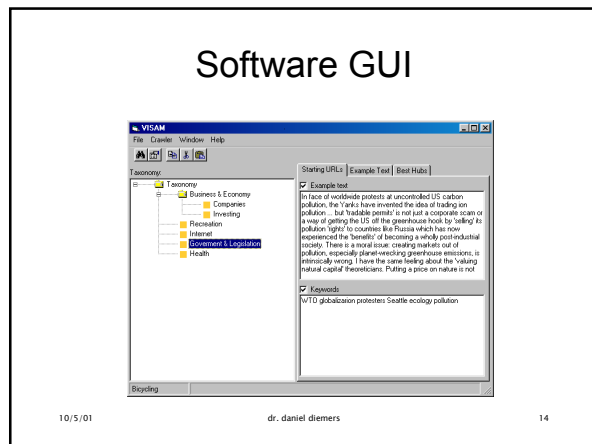
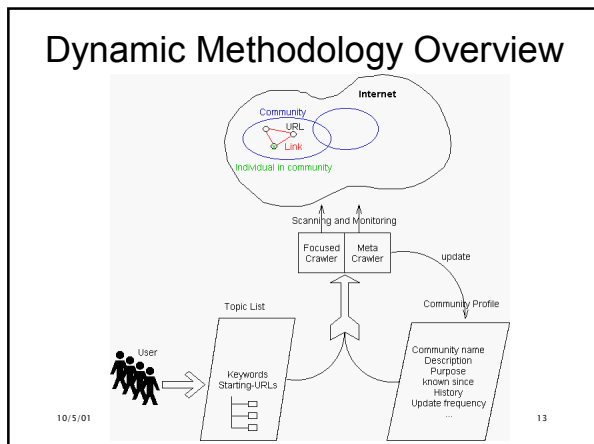
## Static Methodology Overview



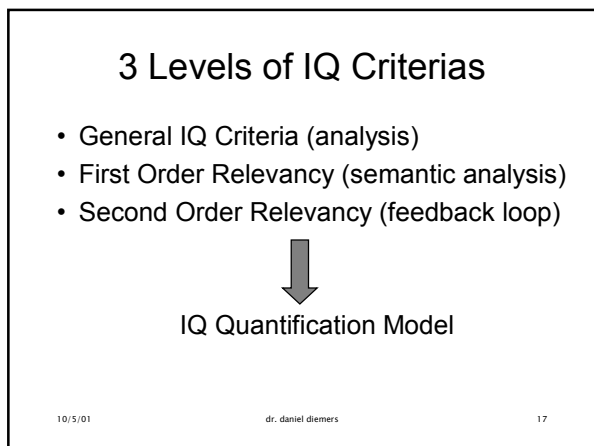
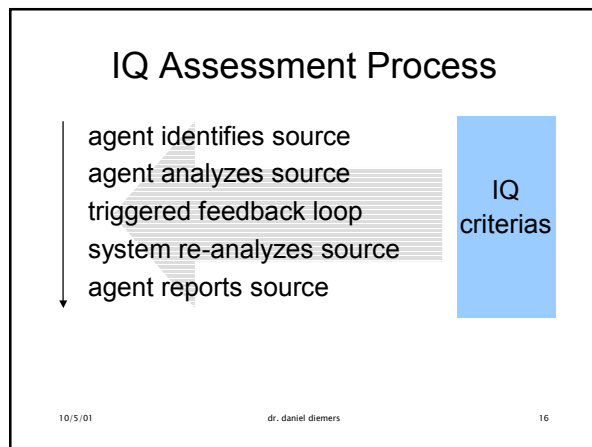
10/5/01

dr. daniel diemers

12



- ### 3. Our Framework for IQ
- How can we measure and quantify the „quality“ of an identified Web source within an early-warning system?
  - Agent-based IQ assessment process
  - 3 levels of IQ criterias applied
- 10/5/01 15



- ### IQ Cat. I: General IQ Criteria
- |                                 |                 |
|---------------------------------|-----------------|
| 1 Latency (Agents sleep/frozen) | #seconds        |
| 2 Server Performance            | #Kbits/s        |
| 3 Relative Size of Source       | #Kbyte/Average  |
| 4 Age of Site                   | #days           |
| 5 Age of Last Update            | #days           |
| 6 IP Address / Domain Name      | semantic rating |
- 10/5/01 18





## Practical Experiences (cont.)

- Establish a learning-system, especially in respect to IQ
- In Scanning Mode: achieve „minimum tolerance“, 90% „no blind-spots“ reliability
- In Monitoring Mode: achieve „zero tolerance“, 100% speed and reliability

10/5/01

dr. daniel diemers

25

## 5. Conclusions

- Internet-based early-warning systems are increasingly becoming accepted and useful tools for management
- monitoring and scanning virtual spaces requires highly sophisticated methodology and agent/software technology
- reliable results depend mainly on a sound and applicable framework for IQ

10/5/01

dr. daniel diemers

26

IQ'2001 community!  
Thank You For Your Attention!

contact information:

dr. daniel diemers  
*institute of sociology sfs, university of st. gallen (hsg), switzerland*  
*partner, wedoso gmbh, zurich, switzerland*

ackerstrasse 44, 8005 zurich, switzerland  
+41 1 271 18 35, e-mail: daniel@diemers.net

10/5/01

dr. daniel diemers

27

# **AN EXPLORATORY INVESTIGATION INTO THE IMPACT OF INFORMATION QUALITY UPON THE PERCEIVED VALUE OF INFORMATION**

**Graham DOIG,**

The Business School, Loughborough University, UK & Microsoft Ltd.  
Tel: + 44 (0) 118 909 3053  
doigg2000@hotmail.com

**Neil F. DOHERTY,**

The Business School, Loughborough University, UK;  
Tel: +44 (0) 1509 223128  
n.f.doherty@lboro.ac.uk

**Chris G. MARPLES**

The Business School, Loughborough University, UK;  
Tel: +44 (0) 1509 223131  
c.g.marples@lboro.ac.uk

**Abstract:** A key challenge for every organisation today is to recognise and unlock the value of its substantial collection of information resources and to utilise them for the maximum benefit of the organisation and its stakeholders. Improving the quality of information is potentially one of the most critical factors that can positively impact upon the perceived value of information within an organisation. The research presented in this paper seeks to provide a deeper understanding of the nature of the relationship between information quality and value. Based upon eight in-depth case studies, conducted at large UK-based, multinational companies, this report concludes that the accuracy, timeliness and consistency of information all have a direct impact upon its perceived value. The paper concludes by articulating some of the most important lessons that have been learned, as a result of this research project, about improving the quality of information.

## **1. INTRODUCTION**

For information to be recognised as a valuable asset that makes a significant contribution to the success of the organisation it is important that the information provided within the organisation is of good quality. Most organisations are reliant upon information for running the business and for making important strategic decisions and for monitoring corporate performance. (Eccles, 1991; Goodman, 1993; Kaplan and Norton, 1996; Finlay 2000). With such an important role to play, the quality of the information being used is a critical factor in the long-term success of the organisation. If the information provided is inaccurate, inconsistent, or not provided in a timely manner then there is every likelihood that the managers of the organisation will make poor decisions or steer the firm in the wrong direction. (Crockett, 1992; Goodman, 1993). For these reasons it is essential that organisations have good quality information available for running the business, supporting the decision making processes, and for monitoring corporate performance. As O'Brien (2001) observed '*information that is outdated, inaccurate, or hard to understand*

*would not be very meaningful, useful, or valuable ... people want information of high quality, that is, information products whose characteristics, attributes, or qualities help to make the information more valuable to them*'. If poor quality information is provided which results in poor decision making, it is very likely that there will be a negative opinion within the organisation on the perceived value of information.

Whilst a tremendous volume of literature has been produced regarding the quality of information, little work has been published which seeks to directly investigate the relationship between information quality and the perceived value of information. The following section reviews the relevant literature, before the research method is discussed in section three. The research results are presented in the fourth section and their importance is assessed in the final sections.

## **2 LITERATURE REVIEW AND RESEARCH OBJECTIVES**

The aim of this section is to present a discussion of the literature with regard to the value of information and the way in which this is dependent upon its quality. In so doing, the motivations and academic justification for this research are established.

### **2.1 The Value of Information**

*'Information is the lifeblood of the organisation'* (CBI, 1992). This powerful statement establishes a foundation for the argument, repeated over many years, that information provides value to the organisation (Porter and Millar, 1985; Glazer, 1991; Glazer, 1993; McPherson, 1994; Rayport & Sviokla, 1995; Hawley, 1995). A number of these opinions have gone as far as suggesting that information should be considered to be an asset and included on the corporate balance sheet. (Willard, 1993; McPherson, 1994). However, making bold statements of this nature is all well and good, but what basis do these statements have? What evidence exists to suggest that information does have value and why is information value important?

Over the years there have been clear indications that information does add value to the organisation. This has led to a number of researchers investigating the use of information within organisations and the value that the information adds. From this research it has become apparent that information can be perceived as valuable for the following reasons:

- information is a product in its own right (Porter and Millar, 1985; Davenport and Cronin, 1988; Mowshowitz, 1992; Rayport and Sviokla, 1995);
- information is incorporated into products to enhance their usability and value (Davenport and Cronin, 1988; Hopper, 1990; Goodman, 1993; Davies and Botkin, 1994);
- information is used for strategic purposes within the organisation (Porter and Millar, 1985; Hopper, 1990; Bowonder and Miyake, 1992; Kaplan and Norton, 1996);
- information supports the day to day operations of the business (Porter and Millar, 1985; Davenport and Cronin, 1988; Glazer, 1991; Bowonder and Miyake, 1992; Souchon and Diamantopoulos, 1996)

When considering the importance of information to all these types of activities Davenport (1993) argued that 'information is both an essential element without which they could not take place, and has the potential to confer value'.

However, it has also been identified that information has traditionally been treated as an overhead of the organisation that has to be managed as a cost rather than as an asset. (Strassmann, 1985; Willard, 1993, Orna, 1999). This perspective appears to contradict the view that information has value and can have a positive impact on corporate performance. But why should this be important, why should recognising information value be important? Most organisations invest in the assets and resources that they believe will assist them in being successful and gaining competitive advantage. (Grant, 1991; Collis and Montgomery, 1995;). The traditional view of information being an overhead has created a situation where many organisations are reluctant to invest in the resources that are required to develop information as an asset. On many occasions the activities of an organisation that are cutback when business is in decline are the information exploitation activities. If information really does provide substantial value to organisations then these cutbacks are a false economy. In identifying this situation Orna (1999) believed that ‘if we can’t find a way of putting information in the same framework as physical resources, they will not get serious sustained attention from decision makers and controllers of resources and they will not receive the investment they need for putting them to productive use’.

Recognising that information provides significant value to the organisation appears to be the major problem even although conventional wisdom clearly identifies that it does. The problem appears to be that this message does not register with senior management and decision makers. This problem is compounded further by the fact that in most organisations information resources appear to be under utilised. There is much more value that could be obtained if the appropriate actions were taken. As Rostick (1994) identified, ‘today’s invaluable corporate information asset lies like sediment at the bottom of an ocean’. Information value is important because recognition of this value is the catalyst for investments in information resources that lead to the utilisation of information as a strategic resource and the generation of substantial benefits. In most organisations today the value of information resources are not fully recognised by management. Many organisations are unaware of the full extent of their information resources, the information resources are under-utilised, and the additional value that could be obtained by utilising this information is being lost.

One of the most likely reasons for information not being perceived as valuable, as is discussed in the following section, is that far too often its quality is poor. Consequently, if the quality of information can be improved, then there is the potential for substantially enhancing the value of information being generated and utilised within the organisation.

## **2.2 The Information Quality Literature**

Information quality can be measured by a wide variety of distinct dimensions. For example, Davenport et al (1992) identified integrity, accuracy, currency, interpretability and overall value as being key dimensions. Based upon the opinions of the users of data, Wang and Strong (1996) formulated a fifteen dimension framework of data quality that included accuracy, timeliness, completeness and representational consistency. As it is beyond the scope of this paper to review all the dimensions of information quality touched upon in the literature, the remainder of this section will focus upon the three dimensions ultimately utilised in this study. The justification for the selection of these three is discussed in the research method section.

- 1. The accuracy of information:** Accuracy is generally acknowledged as one on the most critical dimensions of information quality (Klobas, 1995; Strong et al, 1997; Raghunathan,

1999; Alter, 1999. As defined by Alter (1999) it relates to 'the extent to which information represents what it is supposed to represent'. Having consistent information available at the right time is very important, however, if the information provided is not accurate then the value of the information will be questioned. For example, Goodman (1993) found that '*managers frequently plan, solve problems and make decisions based upon incomplete and sometimes inaccurate information*'. Similar problems were identified by McKinnon and Bruns (1992). They provided examples of managers ignoring information being provided because of the known inaccuracies. In one example 'entire columns of data were being dismissed because they had not been updated for six months'. A key focus area of the Hawley report (1995) was the accuracy of information. This report identified that there were widespread concerns about the quality of information in organisations. Of especial concern was its accuracy and the extent to which it is misinterpreted or misunderstood because the source data is flawed or the information partial. A key recommendation produced by Hawley was, '*the organisation should review the information required at each stage of each process in its business to ensure that necessary and sufficient information is available as required for effective operation, and no more*'.

- 2. The timeliness of information:** Another of the information quality dimensions that is regularly identified in the literature is timeliness (Strong et al, 1997; Alter, 1999; Maltz, 2000), which as Maltz (2000) defines is the dimension that 'refers to whether information is transmitted quickly enough to be utilized'. The importance of timely information being provided for measuring the performance of the organisation was discussed by both Eccles (1991), and Kaplan and Norton (1996). It was Kaplan and Norton who identified lack of timeliness in providing relevant information as being a constraint to effective performance measurement. Similar problems related to executive decision making were identified by Crockett (1992), and Souchon and Diamantopoulos (1996).
- 3. The consistency of information:** Inconsistency of the information that is available is one of the most significant problems that business managers are confronted with. As a result of the haphazard ways that computer systems have been developed in most organisations over the last thirty years the source data that is available is often very inconsistent. As Goodman (1993) recognised when considering the plight of general managers 'their most fundamental challenges are sorting out the uncertain, diverse, and enormous amount of potentially relevant information'. These inconsistencies consist of multiple occurrences of the same data items and different values being maintained for these same data items, such as mis-spelt names, multiple codes being assigned to one entity, and inconsistent date values and formats. Multiple meanings for the same items of information is another common problem. As Wang and Strong (1996) identified 'a major manufacturing company found that it could not access all sales data for a single customer because many different customer numbers were assigned to represent the same customer'. Similar situations were identified by Lingle and Schiemann (1994), and Davenport (1994).

### **2.3 Summary of Literature and Research Objectives**

There are potentially many factors that can influence the perceived value of information. It can be inferred from the review of the literature that information quality is one of the most critical of these. If information quality is generally regarded as being poor this will undoubtedly have a negative impact on the perceived value of information. This is because information quality

dimensions can have an immediate and significant impact on the perceptions that users of information develop, regarding its value. For example, if the information being provided fails to meet the expectations of the user in terms of its accuracy, timeliness and consistency, it is likely that the quality of the information will be regarded as being poor and the perceived value of the information will diminish. If this perception persists over a significant period of time then it is unlikely that information will be regarded as being a valuable asset, and the perception of information being a costly overhead providing little value will prevail. Whilst a strong causal logic can be derived between the quality of information and its perceived value, there is little previously published research, which explicitly tests this logic, has been identified. The aim therefore of this research is to *'investigate whether the perceived value of information is influenced by the quality of information provided to those who use it within an organisation'*.

### **3 RESEARCH DESIGN AND METHODS**

The aim of this section is to review the overall research design, describe the targeting and execution of the case studies and then review data analysis strategy.

#### **3.1 Research Design**

Having established a clear research objective, it was necessary to choose an overall research approach that would be best suited to its exploration. Due to the lack of prior empirical research in this area, an exploratory research design was chosen, as *'it is appropriate to any problem about which little is known'* (Churchill, 1991: p 149). More specifically, it was envisaged that the research objective could best be explored by adopting a multiple case study approach, which has been defined as *'an empirical enquiry that investigates a contemporary phenomenon within its real life context'*, which *'relies on multiple sources of evidence'* (Yin, 1994; p 13). Such an approach was considered ideal for studying the impact of the quality of information on its perceived value, *in situ*, within a variety of large and highly sophisticated commercial organisations. When addressing the question of information value, an obvious question is why choose a qualitative approach when a quantitative analysis might provide more immediately meaningful results. However, as Orna (1999), amongst others, has noted attributing value to information is *'a notoriously difficult subject, which economists have grappled with rather unsuccessfully for many years'*. Consequently, the adoption of a qualitative research design was considered to be more likely to provide useful results.

The detailed design of the research strategy was very strongly influence by the fact that one member of the research team was employed in the role of *'Principal Business Analyst'*, for a large software firm. More specifically, he was responsible for the development and implementation of data warehouse systems, for a wide variety of commercial clients. In this position he had unrestricted access to a wide variety of relevant information and key personnel, within a variety of systems development projects, each of which would make a highly appropriate case study. Consequently, he was able to gain unique insights into this increasingly important phenomenon. The research approach adopted was, however, more akin to *'participant observation'*, than *'action research'*, as the aim of the study was to assess the whole of the organisational change process, involving numerous individuals, rather than to focus primarily on the role of the *'Principal Business Analyst'*.

To ensure that the data collection process was focused and structured, it was necessary at the project's outset to identify the dimensions of information quality that were to be investigated, in each case study. As noted in the review of the literature, there is a wide range of dimensions that

contribute towards information quality. This study focussed upon just three dimensions, namely accuracy, timeliness and consistency. These three dimensions were ultimately chosen as they were the ones most commonly being mentioned in the early, exploratory phases of this research project. Moreover, based upon the numbers of citations, it can be suggested that these are also considered to be the most important in the literature (e.g. Strong et al, 1997; Wang & Strong, 1996).

This research study set out to investigate the research question, presented in section 2.3, within the case study organisations, by considering whether the information provided was of good quality and whether this had a consequential impact upon the perception of information value.

### **3.2 Case Study Targeting and Execution**

Over a five-year period the '*principal business analyst*' was employed on eight major systems development projects. In all cases, the aim the project was to develop and implement a large-scale data warehouse that would dramatically improve the quality and availability of information within the host organisation. Each project necessitated him being seconded to the project for a minimum of three months and spending much of his time working, *in situ*, at the client's site. Consequently, each of the case studies was chosen on the basis of convenience, rather than more objective criteria. However, as each of the case study organisations was a large, highly sophisticated, UK-based public limited company, they constituted a sufficiently homogeneous group to allow meaningful comparisons and contrasts to be made. More specifically the case study organisations were a clearing bank, a commercial bank, a retailer, a car manufacturer and four insurance companies that are labelled 'A' - 'D' in the remainder of the paper.

When conducting a case study, Darke et al (1998) suggests that data should be collected in a variety of ways, including '*formal interviews, questionnaires, observation, and document analysis*', so that the findings can be triangulated. In the context of this study, it was not possible to undertake formal surveys. However, it was possible to employ both formal and informal interviews, observation and document analysis techniques. More specifically, when working on each case study project, the following data collection techniques were employed:

- **Document reviews:** The principal researcher had access to a wide variety of documents, including IT, marketing and corporate strategy reports, staff communication documents and detailed design documents.
- **Interviews:** Formal interviews or informal discussions were conducted with a wide variety of stakeholders in each project, ranging from users through to very senior managers.
- **Observation:** Being an active participant in each project, the principal researcher was able to observe their day to day execution at very close quarters, including participation in the vast majority of important project meetings.

A series of note-books were compiled to ensure that a complete, coherent and contemporaneous set of evidence was captured. Furthermore, the advice of Nandhakumar & Jones (1997) was followed and time was set aside to periodically '*step back from the research context*', to write-up key findings and objectively review them with the other researchers.



### 3.3 Data Analysis Strategy

The source data to be analysed, for each case study, was comprised of a selection of notebooks, formal business documents and the verbatim transcripts from interviews. The first stage of the data analysis exercise was to use the '*QSR NUD\*IST Vivo*' (Nvivo) software to facilitate the coding of all the source documents and the retrieval of data from them. Nvivo was chosen as it provides a range of tools for handling rich data records and information about them, for browsing and enriching text, coding it visually, and for grouping the data records by many categories. A tool of this nature was required to gather, manage, and facilitate the analysis of the wide range of data that was collected during the individual case studies and for assembling the data for cross-case analysis. The next stage of the qualitative data analysis was to create '*within-case*' displays, using the '*ladder of analytical abstraction*' approach (Miles and Huberman, 1994; p 92). More specifically, '*check-list matrices*' and a '*thematic conceptual matrix*' were created for each case study organisation. Having organised and summarised the data, at the case level, it was then possible to embark upon the '*cross case*' analysis, the key component of which was the creation of '*thematic conceptual matrices*' (Miles and Huberman, 1994; p 131). This latter analysis focused upon evaluating the levels of consensus or variability that existed, with respect to the impact of the three different dimensions of information quality upon its value, across the eight case study sites; a '*variable-oriented approach*' to cross case analysis (Runkel, 1990).

## 4 RESEARCH FINDINGS

The aim of this section is to provide an overview of the findings, before presenting a more detailed analysis of the impact of each of the three dimensions of information quality on the perceived value of information.

### 4.1 Overview of findings

The data and evidence was carefully reviewed, to evaluate how significant each quality dimension was to the participating organisations, using a four point scale: highly significant, significant, moderate and none. More specifically, the assessment of the level of impact was based upon:

1. **Current experiences:** the degree to which the case study organisation has been experiencing quality problems, with respect to each dimension;
2. **Desired outcomes:** The extent to which the data warehousing projects have explicitly targeted information quality, and would, once operational, deliver improvements to the accuracy, consistency and timeliness of information, within each case study organisation.

The analytical exercise considered current experiences and desired outcomes in tandem, as organisations who were experiencing very significant quality problems, for example in the area of accuracy, generally had high expectations that the introduction of a data warehouse would deliver significant improvements to the accuracy of their information. The results of this exercise, see table 1, therefore present a unified view of the level of impact, based upon both an organisation's current experiences and desired outcomes.

It is interesting to note that whilst information quality issues were high on the agenda of most case study companies, in two of the organisation, namely **insurance company 'A'** and the **car manufacturer**, information quality was not a significant consideration. It was not so much that these companies were uninterested in quality, but more that as their focus was very firmly elsewhere, quality was generally down-played. For example, **insurance company 'A'** had major problems simply gaining access to any data; information availability was their main concern and

quality would have been the 'icing on the cake'. Similarly, the **car manufacturer** was primarily focussed on improving the performance of its sale to delivery performance by integrating more tightly its sales and manufacturing processes. The availability of relevant information was critical to this initiative but wider quality issues were not of primary importance.

**Table 1:** The Significance of Information Quality within Case Study Organisations

Case	Consistency of information	Timeliness of information	Accuracy of information
Clearing Bank	***	***	***
Retailer	***	***	*
Insurance Co. 'A'			
Insurance Co. 'B'	***	**	***
Commercial Bank	**	**	
Insurance Co. 'C'	**	**	*
Car Manufacturer		*	
Insurance Co. 'D'	*	***	***

**Key :** \*\*\* highly significant impact; \*\* significant impact; \* moderate impact

As there is evidence that all three of the dimensions of information quality have an impact on its perceived value, each of these areas is more fully discussed in the remainder of this section. More specifically, the sections 4.2 - 4.4 seek to provide evidence that specific organisations have focussed strongly on a particular dimensions of quality, in the instances where the highest levels of significance have been detected. This analysis is followed, in section 4.5, by a presentation of evidence that supports the hypothesis that enhanced information quality will result in a concomitant rise in its perceived value.

#### 4.2 Consistency of Information

Providing the business with consistent company-wide information to support decision making and corporate performance measurement was an explicit objective, within the **clearing bank**. As one manager noted: *'the bank has a growing requirement for timely, flexible and consistent information to enable strategic decisions to support the business'*. The major problem that confronted the bank was that although there was a clear understanding of what information was required, very little of this information was being provided at an appropriate quality. One manager's view of the problem was that *"there are issues of data consistency, both across the varied operational systems, and also over time ... these problems of lack of accuracy and consistency, in the underlying data, limit the value of any analysis or report based on the data"*.

The **retailer** had similar requirements to the clearing bank for high quality information to support decision-making and for monitoring of business performance. The problem was summarised by one manager who believed that *'the information that has been provided, had been collected in a piecemeal manner and distributed by many different methods ... this resulted in a lack of consistency and quality'*. The inconsistent nature of the information available made it very difficult for the retailer to have a clear understanding of the information that was available or the ability to provide the business with a consistent view of corporate performance. A management

report on the situation concluded that *'management reports suffer from data inconsistency ... it is difficult to help users with queries regarding the data ... it is also not possible to advise the business on the information currently available'*.

**Insurance company 'B'** also had the challenge of providing information of an improved quality to support decision-making and the measurement of corporate performance. As the company was experiencing a period of dramatic change and internal reorganisation, it was also recognised that good quality information had to be made available to all areas of the business to ensure that the changes were being implemented consistently across the organisation. This philosophy of ensuring that information was provided to all areas of the business that required it was fundamental to the changes taking place within the organisation. This was summarised in the view that *'a comprehensive source of consistent data would be available to all users without internal constraints as to how the data elements can be accessed or viewed together'*.

#### 4.3 Timeliness of Information

In its attempts to support decision making and improve corporate performance measurement, it had also been recognised, within the **clearing bank**, that the timeliness of information provision was a critical factor. However, the reality of the situation was that the provision of information in a timely manner was a significant problem. These problems were identified in a number of areas of the bank. The marketing function was one of the hardest hit areas where *"the ability of product managers to access information about customers and products is severely limited at present ... such information as is available is difficult to get at or provides only a snap-shot of the data which is often too old to be of practical use"*. These problems were reiterated by a number of managers in the bank who had observations such as *"it is often difficult to extract information in a timely fashion in a usable form"*, or *"frequently, by the time the report is available the information it contains is out of data and therefore of no benefit to the user"*, and *"the long lead times to gather information also prevents iterative exploration of information"*.

Timeliness of information at the **retailer** was an issue, not because of a lack of timeliness but rather because of the effort and cost that was involved in delivering the information in a timely fashion. Most information was delivered to management within reasonable time-scales and on the surface the timeliness of information provision was satisfactory. For most senior managers information was being provided when they required it. However, this exterior view of the provision of adequate information in a timely manner was the tip of a very large iceberg that was hidden from the view of senior management. This iceberg was discussed in a consultancy report that stated, *"this iceberg is a substantial and complex system which requires a considerable amount of management to ensure that the necessary data is being provided to the user departments in a consistent and timely fashion"*. This 'system' consisted mainly of a substantial manual effort of extracting data from many different disparate sources, transforming it into the required formats, then re-keying it into other set of computer systems.

The provision of timely information had been identified as critical for the two main strategies being adopted by **insurance company 'D'**. As the new strategies had been developed it had been recognised that the information the company had available was inadequate for what was required. As studies within the insurer had identified *'much of this information is inaccurate, untimely, and in many instances, irrelevant to the actual business needs. Moreover, 'with the development of new business strategies for the IFA and direct divisions it had become apparent that the provision of good quality, timely and relevant information is essential if the strategies*

are to succeed'. Among the critical problems that needed resolution was the ability of the insurer to provide senior managers with up-to-date information for managing the operation. The failure to provide this information was forcing managers to manage 'in the dark'. As an internal study identified '*some of the data is only transferred on a monthly basis, leaving management unaware of business levels during the month*'.

#### 4.4 Accuracy of Information

With the objective of providing the business with the data it required for improved decision-making and monitoring corporate performance the accuracy of the information was an important dimension for the **clearing bank**. Individual managers had identified the importance of being provided with accurate information. As one observed "*the quality of data must be appropriate in the first instance ... information which is both relevant, and accurate enough, for the purposes of the business*". As has already been identified the bank had identified what was required in terms of information quality dimensions but was having difficulty in delivering the requirements. As one manager identified "*there are known problems with the accuracy of data in certain fields such as date of birth where though data is present and in the correct format, the validity of the data is questionable*". As a consequence, it was recognised that "*as well as an increased demand for information, there is likely to be increased demands for more detail, accuracy and consistency of information*".

The improvement of management decision-making was also an objective of **insurance company 'B'**. The provision of accurate information was seen as a key prerequisite to achieving this objective. As internal analysis of the situation identified '*there is also an expectation that by providing consistent, timely, and more accurate information there will be noticeable improvements in the decision making process*'. However, as with most of the other case study organisations the general opinion was that the information that was currently available was not of the required quality. In this case there was an opinion that '*a number of data quality issues were identified which still needed to be tackled ... a key problem was a perceived lack of accuracy*'. There was therefore a pressing requirement to provide "*executive access ..... to a single agreed and validated set of data ... access should be to controlled data which has a known degree of accuracy*".

**Insurance company 'D'** was experiencing similar problems. As discussed earlier the development of its two main strategies was dependent of the provision of good quality and accurate information. However, the information being provided was not of the required standard and the company had started to recognise '*much of this information was inaccurate, untimely, and in many instances, irrelevant to the actual business needs*. As one executive noted, "*manual effort is high; in an attempt to make the data as accurate as possible, great efforts have to be made on a regular basis in order to provide the best possible picture*". For many managers this was a fundamental problem that had to be resolved. There was a view within the company that '*the provision of accurate new data should be seen as a minimum basic requirement*'. Without this there was a feeling that the strategies being adopted could result in failure.

#### 4.5 The relationship between information quality and its perceived value

Having demonstrated that accuracy, timeliness and consistency are considered to be key dimensions of information quality, it is important to present the evidence that supports the hypothesis that there is a causal relationship between information quality and its perceived value. For example, at the **commercial bank** it was recognised that "*the importance of information had*

started to be recognised and senior management has realised that good quality information was of great importance and that without it they would have difficulty competing and maintaining their leading position in the UK market". A similar opinion of the value of high quality information was formed at **insurance company 'C'** where *'the importance of information is being recognised by the company as being imperative to future success and serious steps were now being taken to ensure that information is regarded as being a valuable asset'*. More specifically, the evidence from the case studies suggest that the enhancement of information quality can deliver value in four key areas: decision-making, strategy formulation and monitoring, flexibility and integration.

**Enhanced decision-making:** At **insurance company 'B'** there was an explicit objective to *'deliver in a flexible and convenient fashion high quality business information to all staff, agents, and managers who need it, regardless of their level of computer literacy'*. There was a view that by *'increasing professional potential by improving the quality of management information would lead to improved decisions, and thus provide the basis for improved job performance'*. It was also recognised that high quality information was essential to the enhancement of decision-making at the **retailer**. As one manager identified *"the benefit should primarily be better decision making and better buying decisions if the data is used effectively ... good quality information should also lead to better selling decisions in terms of reductions and locations"*. This type of usage was discussed in more detail with ideas being proposed such as *'if better quality information could be supplied improved buying and selling decisions could be made. The measurement of stock utilisation could improve which would lead to better selling decisions. Better quality information would also enable the business to identify unprofitable business and to improve range and space optimisation and stock space utilisation'*.

**Strategy formulation and monitoring:** There was a great deal of evidence to suggest that the delivery of high quality information is essential for monitoring strategic performance and formulating effective corporate strategies. For example, there was recognition, within the **clearing bank**, that information was required for monitoring the success of the corporate strategies. As one study conducted in the bank identified *'the provision of accurate and consistent information would be required to enable measurement of performance against the bank's critical success factors'*. This view was echoed by a manager from the clearing bank who noted that *'there is a business need for good quality, consistent and timely information to support and monitor progress towards the achievement of the corporate objectives'*. The role of high quality information in strategic performance monitoring was also recognised at the **retailer**, where one manager suggested that *'by gathering good quality data senior management could be provided with Group/Corporate views of company performance'* with the result that *'improvements could be made in strategic decision making'*.

**Flexibility:** As an internal study at **insurance company 'B'** identified *'making consistent information widely available to support quality business decision making is key to achieving flexibility and maintaining strategic alignment'*. The **car manufacturer** also believed that improved flexibility could be achieved by the provision of better information. There was a view that *'good quality information was required to improve the performance of key processes and to achieve nimbleness'*.

**Process Integration:** Another area in which there was a belief that improved information could make a substantial impact was in the improved integration of key processes. Managers at the **car**

**manufacturer** believed that by providing the right quality information they could achieve *'integration with key functions and initiatives, such as a single common global interface with material planning and logistics'*. They were convinced that *'these interfaces could be developed and improved by the use of better quality information'*.

## DISCUSSION: LESSONS LEARNED

The research presented in this paper has reinforced the message that accuracy, timeliness and consistency are all important dimensions of information quality, and that as such they have an important impact on users' perceptions of the value of information. The aim of this section is to highlight some of the strategies to improve information quality that have been identified by the case study companies. It should be noted that each strategy is very specific to the cited case study company, rather than being representative of all the companies:

- **User education and data integrity procedures are essential:** As the **commercial bank** had identified in internal studies *'the quality of the data being gathered was a major concern and a decision was made that steps would be taken to ensure that the data being gathered was of good quality'*. This was facilitated by providing focussed education and training, and implementing data integrity procedures, in all the bank's branches. As the bank identified *'the steps were all aimed at stressing the importance of data quality at the branch and the need to ensure that high quality standards were maintained'*.
- **Need to clean-up data sources:** The introduction of new information sources, such as data warehouses, will only improve the overall quality of information if the quality of all up-stream data sources are also tackled. As an internal report at **insurance company 'B'** identified *'the business need for good quality information is being continually hampered by these diverse (source) systems with inconsistent and inaccessible data'*.
- **Design for quality:** Data quality issues must be prioritised during the design and development of information systems. As the **retailer** noted *'when a new requirement for information is identified the solution is often viewed as 'tactical' and hence speed of development and reducing the cost of the build are usually viewed to be higher priority than resilience, cost of maintenance, future flexibility and data integrity'*.
- **Don't tamper with the source data:** Another major factor, which contributed to the information quality problems that faced the **retailer**, was the practice of manipulating data before information was passed to management. As one manager acknowledged *"in the past much of the data used for strategic decision making has been manipulated which has created a false inference ... by the time it has been through the mill it has changed"*. This resulted in a spider's web of mismatching and incomplete information with many answers being unavailable and many answers being available for the same question, and with no one really knowing what the correct answer was.
- **Provide managers with appropriate tools:** In many cases problems were highlighted with regard to the compilation of data into meaningful information sets. For example, at **insurance company 'D'** much of the information that was required was not available in a complete form and a significant effort was required to assemble what the business was

requesting. As a study of the problems highlighted '*additional problems included data content being poor, data being fragmented forcing the business to use multiple sources, and a high level of manual effort being required to pull together the necessary information*'. Consequently, it is necessary to provide managers with appropriate tools and training, so that they can construct the reports they need.

In addition to highlighting some practical steps that organisations can adopt to improve their information quality, and in so doing their information value, this research also highlights a fundamental lesson about the nature of information quality problems. The evidence from the case study organisations suggests that information quality problems typically manifest themselves across a range of dimensions rather than a single dimension. This view was well summarised by one manager who noted '*while the Bank may have immense volumes of operational data, the current processes for extracting business information from this data are inadequate ... these processes act as a bottleneck, delivering inaccurate / inconsistent data too late*'. This suggests that data / information quality problems might best be tackled holistically rather than at the level of the individual dimensions.

## **6 CONCLUDING REMARKS**

Whilst the literatures with regard to the quality of information and its perceived value are both growing, little previous research has sought to explicitly link these two themes together. The research, reported in this paper, therefore makes an important contribution in that it shows how the quality of information is being explicitly targeted to improve its value, in eight large, multi-national organisations. Moreover, it provides some important insights into how this process can best be achieved.

Research into the role of information, within the organisational context, is an ambitious undertaking, and therefore contains a number of inherent limitations. In particular, the adoption of the case study format reduced the number of organisations that could realistically participate and there is also potential bias with respect to the way in which the principal researcher interpreted the situations to which he was exposed. Moreover, the research only addressed three of the many dimensions of information quality. Consequently, whilst the study provides many interesting and novel insights, these limitations do highlight the need for follow-up studies to be conducted that adopt different methods, and target different populations and respondents, to investigate the generalisability of the results.

## **REFERENCES**

- Alter, S., (1999), *Information Systems – A Management Perspective*, Reading MA, Addison-Wesley.
- Bowonder, B. and Miyake, T. (1992), 'Creating and Sustaining Competitiveness: Information Management Strategies of Nippon Steel Corporation', *International Journal of Information Management*, Vol. 12.
- Churchill, G. A., (1991), *Marketing Research Methodological Foundations*, 5<sup>th</sup> edition, Orlando: Dryden Press.
- Collis, D. J. and Montgomery, C. A. (1995), 'Competing on Resources in the 1990's', *Harvard Business Review*, Jul-Aug.
- Confederation of British Industry (CBI), (1992), *IT The Catalyst for Change*, London, CBI.

- Crockett, F. (1992), 'Revitalising Executive Information Systems', *Sloan Management Review*, Summer.
- Darke, P., Shanks, G. & Broadbent, M. (1998)'Successfully completing case study research: combing rigour relevance and pragmatism', *Information Systems Journal*, Vol 8, No 1, pp273 - 290.
- Davenport, T.H. (1993), *Process Innovation. Reengineering work through information technology*, Boston MA, Harvard Business School Press
- Davenport, T.H. (1994), 'Saving IT's Soul: Human Centered Information Management', *Harvard Business Review*, March-April.
- Davenport, T.H., Eccles, R.G., Prusak, L. (1992), 'Information Politics', *Sloan Management Review*, Fall
- Davenport, L. & Cronin B. (1988), 'Strategic Information Management - Forging the Value Chain', *International Journal of Information Management*, Vol 8 No 1
- Davis, S. and Botkin, J. (1994), 'The Coming of Knowledge Based Business', *Harvard Business Review*, Sept-Oct.
- Eccles, R. G. (1991), 'The Performance Measurement Manifesto', *Harvard Business Review*, Jan-Feb
- Finlay, P. (2000), *Strategic Management*, Harlow, Pearson Education.
- Glazer, R. (1991), 'Marketing in an Information Intensive Environment : Strategic Implications of Knowledge as an Asset', *Journal of Marketing*, Oct.
- Glazer, R. (1993), 'Measuring the Value of Information: The Information Intensive Organisation', *IBM Systems Journal*, Vol 32 No 1.
- Goodman, S.K. (1993), 'Information Needs for Management Decision Making', *Records Management Quarterly*, October.
- Grant, R. M. (1991), 'The Resource Based Theory of Competitive Advantage: Implications for Strategy Formulation', *California Management Review*, Spring.
- Hawley Report (1995), *Information as an Asset: The Board Agenda*, London, KPMG Impact Group
- Hopper, M. D. (1990), 'Rattling SABRE - New Ways to Compete on Information', *Harvard Business Review*, May-Jun.
- Kaplan, R. S. and Norton, D. P. (1996), 'Using the Balanced Scorecard as a Strategic Management System', *Harvard Business Review*, Jan-Feb.
- Klobas, J. E., (1995), 'Beyond information quality: Fitness for purpose and electronic information resource use', *Journal of Information Science*, Vol. 21, No. 2.
- Lingle, J.H. and Scheimann, W.A. (1994), 'Is Data Scatter Subverting Your Strategy', *Management Review*, May.
- McKinnon, S. M.and Bruns, J. (1992), *The Information Mosaic*, Boston MA, Harvard Business School Press
- McPherson, P. K. (1994), 'Accounting for the value of information', *Aslib Proceedings*, Vol 46 No 9.
- Maltz, E., (2000), 'Is all communication created equal?: An investigation into the effects of communication mode on perceived information quality', *Journal of Product Innovation Management*, Vol. 17, No. 2.
- Miles, M. B. & Huberman, A. M. (1994), *Qualitative Data Analysis*, Beverly Hills CA, Sage.
- Nandhakumar, J. & Jones, M. (1997) "Too close for comfort? Distance and engagement in interpretive information systems research", *Information Systems Journal*, Vol. 7, No. 9, pp 109-132.



- Mowshowitz, A. (1992), 'On the market Value of Information Commodities, Parts I - III, Journal of the American Society for Information Science, Vol. 43, No. 3.
- O'Brien, J. A., (2001), *Introduction to Information Systems – Essentials for the Internet worked E-Business Enterprise*, Singapore, McGraw-Hill.
- Orna, E., (1999), *Practical Information Policies*, Aldershot, Gower.
- Porter, M.E. and Millar, V.E. (1985), 'How Information Gives You Competitive Advantage', *Harvard Business Review*, Jul-Aug.
- Raghunathan, S., (1999), 'Impact of information quality and decision-maker quality on decision quality: a theoretical model and simulation analysis', *Decision Support Systems*, Vol. 26.
- Rayport, J. F. and Sviokla, J.J. (1995), 'Exploiting the virtual value chain', *Harvard Business Review*, Nov-Dec.
- Rostick, P. (1994), 'An Information Manifesto', *CIO*, Sept.
- Runkel, P. J. (1990), *Casting nets and testing specimens: Two grand methods of psychology*, New York NY, Praeger.
- Souchon, A. L. and Diamantopoulos, A. (1996), 'A Conceptual Framework of Export marketing Information Use : Key Issues and Research Propositions', *Journal of International Marketing*, Vol. 4, No. 3.
- Strassmann, P. A. (1985), *Information Payoff*, New York NY, The Free Press.
- Strong, D. M., Lee, Y. W., Wang, R. Y. (1997), 'Data Quality in Context', *Communications of the ACM*, May, Vol. 40, No. 5.
- Wang, R. E. and Strong, D. M., (1996), 'Beyond Accuracy: What Data Quality Means to Data Consumers', *Journal of Management Information Systems*, Spring, Vol. 12, No. 4.
- Willard, N. (1993), 'Information Resource Management', *Aslib Information*, May.
- Yin, R. K., (1994), *Case Study Research*, Thousand Oaks, Sage Publications.

## **Data Quality in the Small: Providing Consumer Information**

Arnon S. Rosenthal, Donna M. Wood, Eric R. Hughes, Mary C. Prochnow

The MITRE Corporation

202 Burlington Road

Bedford, MA 01730, USA

(813) 831-5535 (813) 835-4661 (fax)

{arnie, dwood, hughes, maryproc}@mitre.org

**Abstract:** Data quality, defined here as fitness for use, is increasingly seen as a serious problem in government and private sector databases. In this paper, we survey available techniques, and then describe our own work.

We are in the process of adapting general data quality techniques suited to large government relational databases, focusing on an aspect rarely seen in the literature, i.e., helping a user assess the quality of individual data records. Our emphasis is on developing solutions to the problem of providing better consumer information on each value used. We provide such information, so the consumer can determine whether the data are good enough for the intended purpose. The primary concern is with individual data items that drive major decisions, where erroneous data have high cost (e.g., human lives). The broad aim is to enable better decisions. A narrower aim is for consumers to trust data when appropriate, thereby reducing the incentives to ignore the data or expend effort on workarounds for data of unknown quality. This paper explains where our approach fits in the spectrum of data quality approaches, and describes a methodology for providing consumers with information needed to guide how they use each data value in making decisions. The methodology encompasses the following aspects:

- Providing an infrastructure to define, store, and make available quality attributes on various data records
- Obtaining values for quality attributes on important data granules
- Making the quality attribute values available to systems and people that use each data granule
- Tracking the impact of providing the quality values on decision-makers and decisions

**Key words** - data quality annotations, quality annotation methodology

## 1. Introduction

*Data quality*, defined as *fitness for use*, is increasingly seen as a serious problem in government and private sector databases. We are currently involved in a research project that has produced a methodology for providing consumers with information needed to guide how they use each data value in making decisions. This paper explains where our research fits in the spectrum of data quality approaches, and discusses our results and future plans.

## 2. Overview of Data Quality Approaches

There are two basic approaches to improving a system's data quality: *defect reduction* and *consumer information*.

*Defect reduction* efforts receive more attention in the literature. The mainstream of data quality research and products seems driven by data warehousing, enterprise resource planning systems, customer relations, and direct mail. For such efforts, one typically gathers impressions or statistics about the quality of large sets of data (e.g., all customer deliver-to addresses), the benefits of improved quality for each category, and the likely costs of improvement. One then alters the data acquisition and cleaning processes to improve the data values stored within the database [Red97]. Many government applications use non-rigorous, informal methodologies for defect reduction.

*Consumer information* efforts aim to make the existing data more usable, by adding information. One aspect is to better document how one interprets the *meaning* of the data (for example, just how 'Threat' is defined in Army applications or whether a French unit reports distance in meters, feet, or kilometers). Understanding the meaning is particularly important when connecting an automated application, which may not realize that 5 feet is a ridiculous distance for a tank sighting report. Because meaning is covered in the extensive data integration literature [Bln86, Rah01, Mil01], we will not consider it further here.

We focus instead on an aspect rarely seen in the literature, i.e., helping a data consumer assess individual data values. We are concerned with individual data items that drive major decisions, where erroneous data have high cost (e.g., loss of life). However, we find that the same quality measures can be used for both defect reduction and consumer information.

Our task therefore goes beyond the data quality marketplace. Traditionally, data are byproducts of providing goods or services; for our customers, information may be the primary product. Traditional efforts often use data for routine automated transactions; there is little human involvement with each data instance. Errors there are costly in the aggregate (e.g., wasting 10% of a direct mail campaign), but a single wrong data value rarely causes loss of life (or the equivalent in corporate motivation and survival, catastrophic financial loss). In government settings, a human typically inspects the data before a decision is made.

*We aim to provide the consumer with a better picture of an individual item's quality so he can determine whether the data are good enough for the intended purpose. The broad aim is to enable better decisions. A narrower aim is for consumers to perceive the databases' contents as trustworthy, thereby reducing the incentives to ignore the data or expend effort on workarounds.*

*In other words, our methodology addresses both the quality of aggregate data sources as well as that of individual data items.*

## **2.1 Project Setting**

Our research focuses on two operational databases, their interactions with each other, and a subset of their data producers and consumers. In our domain, data necessarily give an imperfect picture of the external world. While we are performing informal studies for defect reduction, our efforts have focused on providing information so data consumers can more appropriately and confidently employ the data that are available.

For defect reduction, we conferred with data providers, system managers, and some users of these databases to identify individual data attributes whose quality was perceived as problematic. Several common issues emerged:

- The effect of semantics (data item meaning) on data quality
- The effect of business rules on data quality
- The cause and effect of inconsistencies between databases
  
- The effect of [poor] data quality on the enterprise
- The effect of database structure on data quality

We selected the quality measures (derived from the literature [Str94]) that would describe the problem to guide future efforts. The central idea is to allow consumers to see quality values for the data they retrieve. We define each step so that the process can be repeated. It is interesting that when compared with prior data quality methods (e.g., [Wang93]), the steps line up fairly exactly, but many of them took a radically different form.

Figure 1 illustrates two variations of the specific case that we investigate. In both instances, no quality annotations are provided in the database. When a consumer accesses data directly from the source, as depicted in the top flow of Figure 1, it is obvious that the consumer cannot ascertain quality. The second case, depicted in the lower flow of Figure 1, presents the problem of consumers accessing data that have been derived from a source database. In this case, even if the producer database contains quality indicators and the consumer is aware that the quality indicators exist, there is no mechanism to derive quality values as information flows from one system to the next.

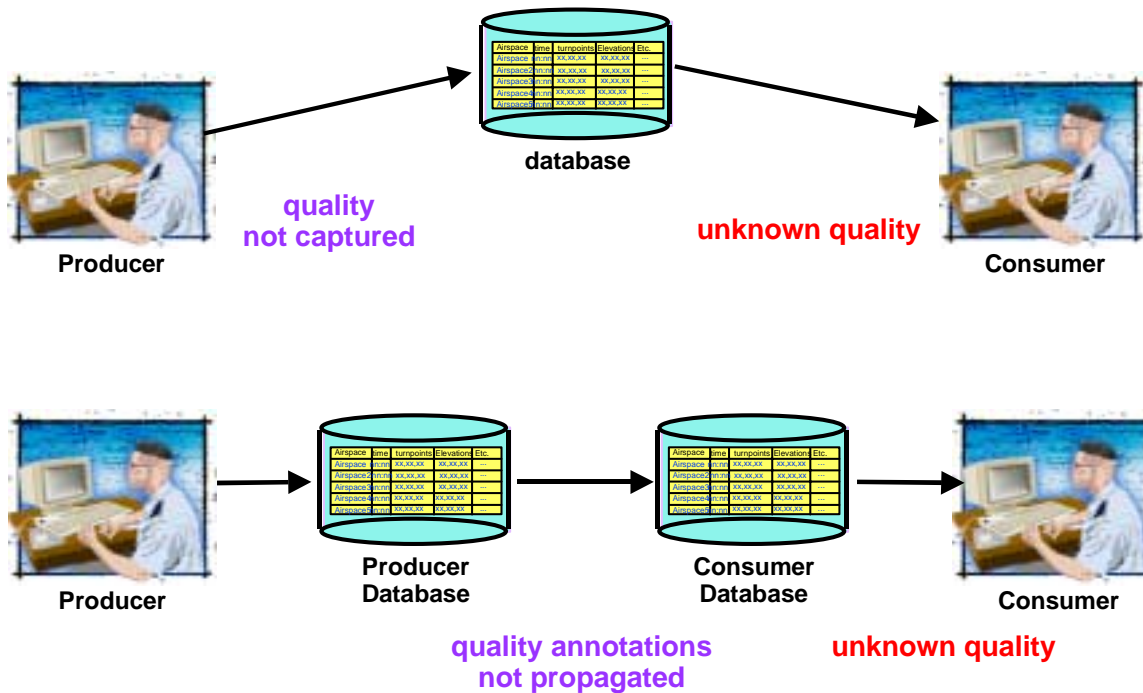


Figure 1. Problems with Understanding Data Quality

## 2.2 Overview of Methodology for Providing Consumer Information

Our research has revealed that in order to provide the data consumer with the information needed for critical decisions, these tasks must be accomplished:

- Provide an infrastructure to define, store, and make available quality attributes on various data items
- Obtain values for quality attributes on important data granules
- Make the quality attribute values available to users and systems
- Track the impact of providing the quality values on decision-makers and decisions

We assume that these technical tasks will be performed in the context of a business process reengineering effort designed to fix the problems with the processes that are used to provide, store, and access shared information. The next section provides more detail.

## 3. Methodology for Providing Consumer Quality Information

We begin with a data quality annotation infrastructure. The infrastructure's basic task is to allow administrators and other authorized users to attach values of data quality metrics, as annotations, to various chunks (*granules*) of the database, and to make this information available to other users of those granules. These concepts are defined below.

A *data quality (DQ) metric* describes the usefulness of some data. Popular examples include measures of accuracy, precision, source, completeness (for sets), and time of observation

[Fox95]; others may be derived from these; still others may be collected. Our infrastructure can contain quality metadata on quality values – after all, quality values are data. However, our investigations have not pursued this second order effect. In the future, we may track information *utility* (i.e., benefit of having it), both for ordinary data and for metrics; utility can be a function of quality.

An *annotation* is a triple, (annotated-object, annotation type, value) that is logically attached to some granule of a database. An alternative logical view might include some of the annotations as part of the regular database structure. The ordinary interface does not show whether an annotation is physically co-located with the annotated object (e.g., for image metadata) or stored separately. We use typed annotations so that many systems can use the same definitions of quality measures, which might be standardized for particular domains. The next section describes this concept in more detail.

### **3.1 Provide an Infrastructure to Define, Store, and Make Available Quality Attributes on Various Granules of Data**

While not strictly part of the methodology, it is interesting to understand the infrastructure provided to support the data quality work. The primary requirement is to be non-intrusive. The infrastructure is able to employ existing data that provide quality information (e.g., dates of capture, error bounds), as well as store separately-provided knowledge, at cell, column, row, and table granularities. A more sophisticated infrastructure could capture knowledge as rules (e.g., If Year=1996 and company=MITRE then EarningsAccuracy < 0.95). Added metrics are stored separately from the application tables. The infrastructure provides operations to administer, update, and retrieve data quality metrics.

*Administration* comprises definition, annotation administration, and physical administration. One can define types of annotations (e.g., *accuracy*, *time-captured*, *source*) as ordinary data attributes. For each, one supplies a data type, value constraints, and prose describing its meaning. For each attribute that receives that type of annotation, an administrator specifies 1) rules that derive values from contents already in the database, or 2) that explicit storage be allocated. For example, one of our subject databases contains many attributes that describe how a datum was obtained, and an estimate of its currency. These are logically derived into annotations, but need not be physically replicated. Annotation administration controls are shown, by default, as part of the annotation user interface.

The infrastructure maintains the relationship between an annotation value and a granule in the database, e.g., a table, row, column, or cell. Annotations are updated as ordinary database data. Access permissions can either be derived from those for the annotated data, or managed as for any other data. For read, one can get annotations exactly on a granule, or include super- and/or sub-granules. The infrastructure provides a generic query interface that presents annotations as additional columns of the annotated table. The semantics are those of an ordinary database view.

Finally, we note that the infrastructure works for any kind of annotations one wishes to attach to data values – it is not specific to quality information other than the types of annotations we

have defined. Database researchers are making interesting progress on all sorts of annotations [Delc01, Bird00].

### **3.2 Obtain Values for Quality Attributes on Important Data Granules**

Again, intrusion and extra work are minimized.

The first step is to determine what quality metadata is already provided in the database schema. If possible, we will get providers' agreements to continue supporting these attributes, and to provide fill for them. (One of our subject databases contains many attributes describing data acquisition and processing, and these provide much of the necessary information.)

Beyond this, we intend to capture wholesale rules that describe all the instances provided by a data feed. This is much cheaper than manually creating each instance. In one of our subject databases, this approach can be used to derive completeness and consistency measures, and estimate precision and accuracy for geospatial coordinates.

To plan gathering of further quality information, one works from two sides – need and ease of capture. For need pull, we determine what quality data would make a difference, and be desirable to obtain. As a form of push, we capture quality annotations that are cheap to get (e.g., time of entry, source).

Builders of data capture software have the option of enforcing data constraints, which sometimes improve data quality. These include value constraints and referential constraints (i.e., that subsidiary data must refer to entries already in the table). Ease of use must be considered. In Desert Storm, data providers disliked the constraints, and moved much of their content to free-text fields. An alternative, supported by our approach, is to record constraint violations as annotations for later attention, e.g., to check for alternate spellings.

But even the best data capture software cannot provide fill where none is available, nor recheck to determine if an office has moved or a company has a new vision. Some observations are inherently unreliable (e.g., number of people in an organization). For these cases, quality metrics should be provided.

Where part of a record fails the quality checks, we want a means of capturing the good part. (In the past, one government system lost considerable data that its data-passing interfaces found somehow faulty; the overall effect of these interfaces on data quality was detrimental.) One approach is to set “bad” values to null, with an annotation holding the suggested value so it is not lost. Automated applications will need to be null-aware, i.e., to behave correctly with null data.

### **3.3 Make the Quality Attributes Values Available to Users of Each Data Granule (Including Both Humans and Queries)**

We explore two means of providing quality values to users -- non-intrusively. Figure 2 illustrates our concept. The producer provides quality annotations, which are captured by our tool in a separate but related database. These annotations are then propagated either directly to the

consumer, or to the consumer database. We note that the consumer needs control over whether screen space is devoted to these extra columns when displaying the results of database queries. In both cases, we have provided a very basic implementation. We also modified two existing user interfaces: one for querying, and the other for map display. In both cases, once the implementers understood the user interface's code, a few hundred new lines sufficed. We are convinced that a fuller implementation need not be very difficult. Generic interfaces for annotations should be provided for the most common forms, i.e., relational (which we have completed) and Extensible Markup Language (XML).

### 3.4 Track the Impact of Providing the Quality Values on Decision Makers and Decisions

We anticipate that it will be very hard to track the impact of quality metadata on user decisions. Several techniques seem natural. For now, we lean toward using only the first, which is least intrusive:

- Use interfaces that make display of quality values optional, generating different queries based on what quality values the user wants retrieved. We can track whether users include quality annotations in their displays (though not its influence on their decisions).
- Survey users about what they use and how valuable it is.
- Provide a box for rating the utility of metadata, as part of the user interface. (For example, Amazon.com lets users rate the utility of feedback from a reviewer.)

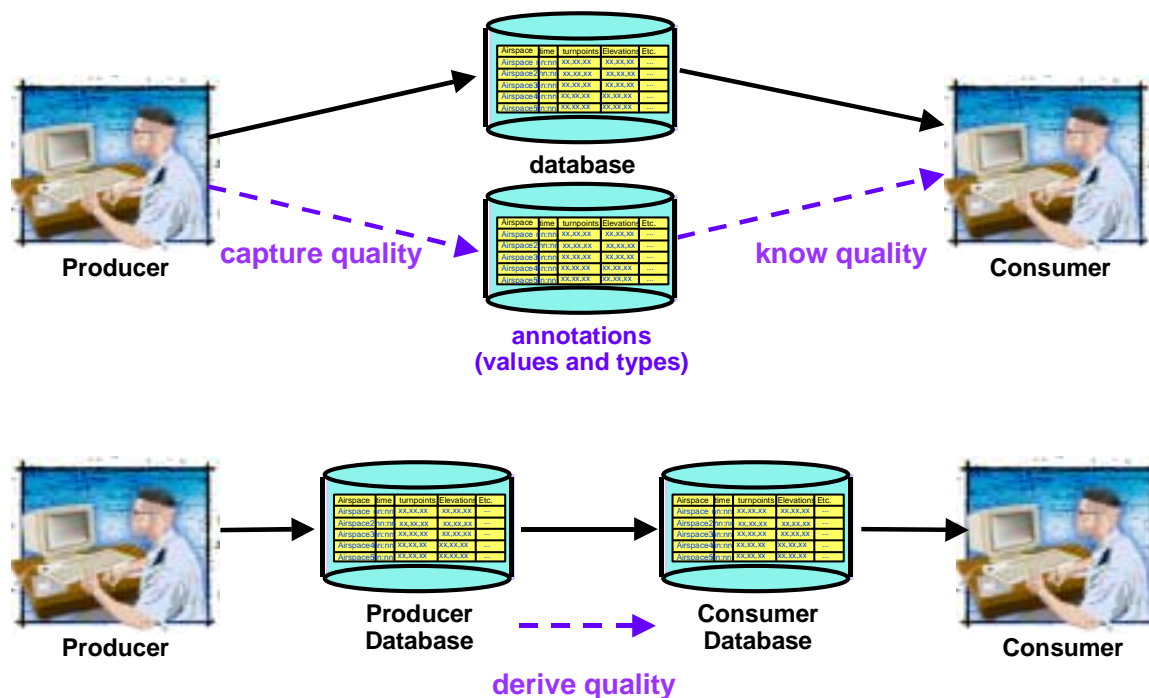


Figure 2. Benefit of Consumer Quality Information



#### 4. Conclusions and Future Work

We have shown how database systems can be extended to manage data quality annotations on base data. Our critical next step is to engage real users, and improve the approach based on their feedback.

Figure 3 illustrates some additional future directions. We hope to provide the means for creating and managing tailored views (comprising both query and display) for communities of interest (COIs), and for adding data quality capabilities to these views. We aim to reduce the cost and delays in producing and maintaining tailored interfaces, thereby enabling better ones to be provided. To do so, we plan to build a componentized view capability that can address both query and display aspects of an interface. We will show how this view capability will allow COIs to form dynamically, collaborate through a view, and update data via the view. We also intend to investigate mechanisms that allow users to provide feedback on quality annotations. In addition, we will consider techniques to dynamically choose source information with quality annotations considered.

Portions of this paper have been derived from A. Rosenthal et al., "Methodology for Intelligence Database Data Quality", AFCEA Database Colloquium, August 2001.

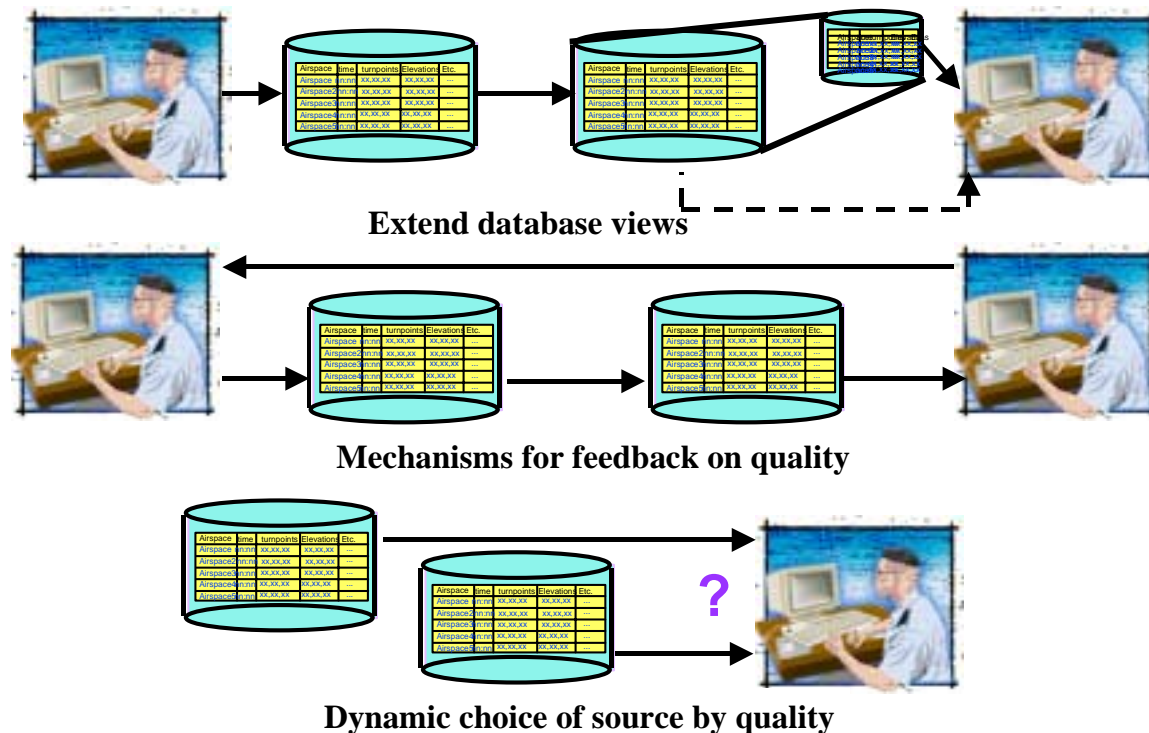


Figure 3. Future Directions

## References

- [Bird00] S. Bird, P. Buneman, W-C Tan, "Towards a Query Language for Annotation Graphs," *Conference on Language Resources and Evaluation 2000*.
- [Bln86] C. Batini, M. Lenzerini, and S. Navathe, "A Comparative Analysis of Methodologies for Database Schema Integration," *ACM Computing Surveys*, 18(4):323-364, 1986.
- [Delc01] L. Delcambre, D. Maier, et. al., "Bundles in Captivity: An Application of Superimposed Information," *IEEE International Conference on Data Engineering 2001*.
- [Fox95] C. Fox, A. Levitin, and T. Redman, "The Notion of Data and Its Quality Dimensions," Massachusetts Institute of Technology (MIT) Sloan School of Management TDQM-95-08, February 1995.
- [Mil01] R. Miller, M. Hernandez, L. Haas, L. Yan, C. Ho, R. Fagin, L. Popa, "Clio: A Semi-Automatic Tool For Schema Mapping," *ACM SIGMOD Record, web edition, March 2001*. <http://www.acm.org/sigmod/record/issues/0103/index.html>
- [Rah01] E. Rahm, P. Bernstein, "On Matching Schemas Automatically," Tech. Report 1/2001, Comp. Science Dept., U. Leipzig, Feb. 2001, <http://dol.uni-leipzig.de/pub/2001-5>, to appear in VLDB Journal.
- [Red97] T. Redman, "Data Quality for the Information Age," Artech House, 1996.
- [Str94] D. Strong and R. Wang, "Beyond Accuracy: What Data Quality Means to Data Consumers," MIT Sloan School of Management, Cambridge, MA TDQM-94-10, October 1994.
- [Wang93] R. Wang, H. Kon, S. Madnick, "Data Quality Requirements Analysis and Modeling," *IEEE International Conference on Data Engineering 1993*.

## Appendix A Comparison With Other Methodologies

To understand the novelty of our work, we compare with a methodology motivated by finance and industrial databases [Wang93]. The points of difference are:

- Government organizations often have individual data values that drive important decisions, and are evaluated by humans, rather than by tools.
- Some methodologies suggest filtering out data that do not meet standards. In many government applications, one uses the best data available – but more cautiously.
- Aging is a major problem with much of our data.
- Coverage can be *very* sparse, and too costly to increase.
- Other methodologies require the data administrator to estimate quality, since data are entered by clerks. With professional information analysts, one may often get good estimates.
- Our research is investigating the improvement of an existing system, not the design of a new system from scratch.
- Our methodology has no need to treat subjective and objective metrics differently.

# Quality Mining

## A Data Mining Based Method for Data Quality Evaluation

Sabrina Vázquez Soler and Daniel Yankelevich  
Pragma Consultores and  
Departamento de Computación – FCEyN  
Universidad de Buenos Aires, Argentina  
svazquez@dc.uba.ar, dyanke@pragmaconsultores.com

**Abstract:** The value of information depends directly on the quality of the data used. Decisions are no better than the data on which they are based. How can organizations assess the quality of their information? How can they know if their data are useful?

Quality control and management have become competitive needs for most businesses today, and there is a wide experience on the topic of quality. Approaches range from technical, such as statistical process control, to managerial, such as quality circles. An analogous experience basis is needed for data quality.

In this paper we present a method for data quality evaluation based on Data Mining. We introduce QuAsAR, a mechanism for the systematic analysis of correctness based on the information itself.

In order to evaluate the performance of the method, we apply it to a real case study. This case study helps us to analyze support and confidence intervals and distribution of erroneous data.

**Keywords:** Data Quality, Data Mining, Quality Mining, Quality Control, KDD.

**Acknowledgements:** This research was partially supported by the ANPCyT under ARTE Project grant PIC 11-00000-01856 and UBACyT grant PIC TW72. The authors also want to thank Martin Patrici and Monica Bobrowski for their helpful comments.

### 1 INTRODUCTION

Having the right information at the right time is a key issue in today's organizations. The value of information directly depends on the quality of the data used. Decisions are no better than the data on which they are based [2]. However, few organizations handle information as a tangible asset. How can companies assess the quality of their information? How can they know if their data are useful?

Managers need to have the ability to verify the usefulness and correctness of the information they use, not only for decisions making, but also to allow them to learn more about the business. The information may help in restructuring areas, improving workflow, etc.

In general, inaccurate, out-of-date, or incomplete data can have a significant impact not only on the organization that generates them. Errors in credit reporting is one of the most striking examples of the social consequences of poor quality data. For instance, the credit industry not only collects financial data on individuals, but also compiles employment records [1].

On the other hand, organizations are learning that in order to provide quality products or services, they need to implement quality programs. Many corporations have devoted significant time and energy to a variety of quality initiatives such as inter functional teams, reliability engineering, and statistical quality control [2] [19].

Quality control and management have become competitive needs for most businesses today, and there is wide experience on the topic of quality. Approaches range from technical, such as statistical process control, to managerial, such as quality circles. An analogous experience basis is needed for data quality [18] [19].

Usually, it is not easy for organizations to test their data. One of the main factors could be that domain experts - people with knowledge on business domain or on the methodology involved - are not responsible for data analysis.

Managers need to count on some mechanism to be able to perform this task and achieve those aims. Providing such a mechanism is the goal of this work [2].

In this paper we present a data quality evaluation method. We introduce a mechanism for systematic analysis of correctness based on the information itself. We also present a case study to analyze the performance of this mechanism.

The method we present is based on *Data Mining*. We use the data intrinsic rules to characterize and evaluate data. There are several reasons why it is better to analyze business patterns through rules, the most important being that the analysis is based on the patterns, which are several orders of magnitude smaller than the data. Besides, domain experts normally associate business knowledge to behavior patterns. This is a common way of characterizing knowledge.

Therefore, the knowledge of rules that allows improving information quality can also have different uses. The wrong thing is to consider data quality as good without a previous check, and realizing afterwards that wrong decisions were made.

In Section 2 we discuss the basis of our research. In Section 3, we present the *QM* method. In Section 4, we introduce *QuAsAR*, a *QM* technique. Section 5 describes the case study developed and its results.

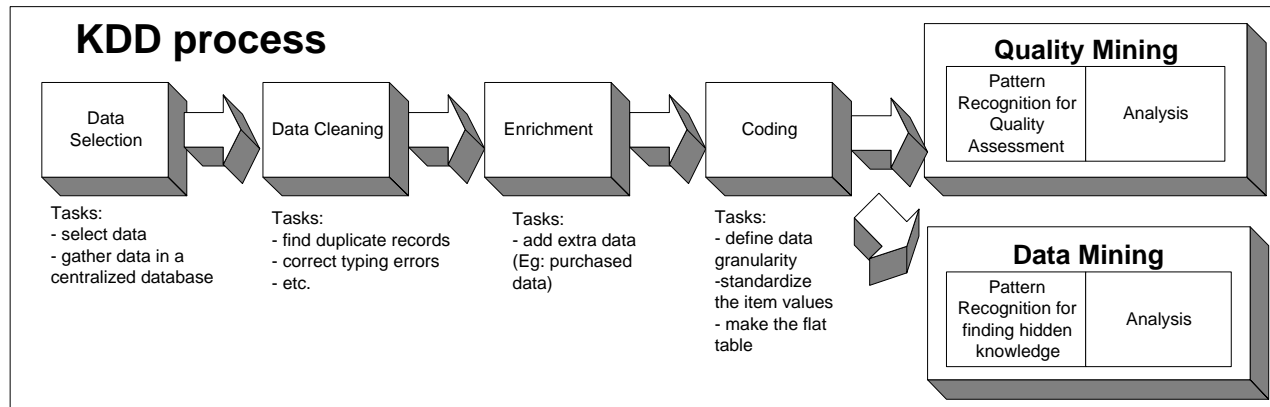
## 2 QUALITY MINING

### 2.1 Overview

We define *Quality Mining (QM)* as a method for *Data Quality (DQ)* [2] [15] evaluation inspired in *Data Mining (DM)* [1] [9] [11], that is, this method involves every kind of *DQ* techniques based on *DM*.

The main difference between both methods is the goal they attempt to reach. The aim of *DM* [1] [11] is to find new knowledge while *QM* strives for the evaluation of data quality using reliable patterns, which do not necessarily introduce new knowledge.

Gathering knowledge from data is directly related to the *Knowledge Discovery in Databases process (KDD)* [1]. *DM* represents the new knowledge discovery dimension of this process. Moreover, it assumes that the data is complete, compiled and “cleaned” to start the knowledge discovery phase. On the other hand, *QM* represents the data quality evaluation dimension. The following chart shows our view of KDD:



This method can be used to validate a group of data before they are used for decision making. In these cases, managers need correct information.

In the case of system migrations, it is possible to use this method to gather integrity rules that may be incorporated as *Metadata* to the new system, to improve data quality.

## 2.2 Related Work

### 2.2.1 DataSphere

In AT&T [16], a technique called *DataSphere* has been developed. This technique allows the detection of *data glitches*, that is changes introduced in data by external causes not related to normal noise. We understand by normal noise uncontrollable measurement errors such as imprecise instruments, subtle variations in measurement conditions (climatic conditions, software degeneration, etc.) and human factors. Otherwise, *data glitches* are systematic changes caused by mega phenomena such as unintended duplicate records, switched fields, and so on.

This technique partitions the attribute space in subsets based on two criteria: distance and direction. It is also possible to use clustering and classification to generate the subsets. Directional information is superimposed on distance using the concept of pyramids. Every layer-pyramid combination represents a class of the *DataSphere* partition. The data points in each class are summarized by a profile. The glitch detection is very fast since it is based on the profiles, which are several orders of magnitude smaller than the original data set.

### 2.2.2 Wizrule of WizSoft

This application can obtain rules based on data from the definition of three variables: minimum probability – that corresponds to confidence -, minimum accuracy and number of appearances of rules – that corresponds to the support [20]. Pattern recognition algorithm is not given; the documentation describes that statistical methods are used to determine patterns.

There are two main differences between this technique and *QuAsAR* (see below for further reference). The first one is that the concept of minimum probability does not correspond to the definition of confidence ranges that determines the rules. The second reason is the impossibility to analyze separately the concepts of support and confidence [20].

These techniques, *DataSphere* and *Wizrule*, are also included in *QM* framework, although they face data quality problem from different aspects.

### 3 QUASAR: A QUALITY MINING TECHNIQUE

#### 3.1 Introduction

The term *Data Quality* can be best defined as “fitness for use”, revealing the relativity of the concept. Fitness for use means the need to go beyond traditional concerns of data accuracy is necessary [15].

In addition, the domain experts normally associate business knowledge to behavior patterns. This is a common way of characterizing knowledge. For instance, “The Wing B of the X Hospital only deals with hepatic problems, or all the people that work there are dentists”, and so on.

If the discovery of rules shows hidden knowledge from data, their use as a mechanism for testing data quality would not be efficient. The characteristic of unknown information makes business experts analysis not easy. Although *DM* [1] [11] techniques are really adequate for finding hidden knowledge, some experiments have shown that data quality is essential to determine the reliability of the knowledge found [1].

However, having the rules does not ensure the complete solution of data quality problem: consistency does not mean/equal correctness. This method helps to find data inconsistency, it will not be possible to detect an incorrect but consistent datum. However, if it is inconsistent, we can classify it as a potential error.

#### 3.2 Overview

This technique is described as a method for data quality testing based on re-engineering [7]. It is called *Quality Assessment using Association Rules (QuAsAR)* and is based on *Association Rules (AR)* [1] [9] [11] techniques.

An association rule -a rule like  $\chi \Rightarrow I_j$ - is a representation of a relationship between variables. This technique has two main concepts used to search data: confidence and support. Formally, given a rule  $\chi \Rightarrow I_j$ , where  $\chi$  is a set of several items and  $I_j$  is an item not included in  $\chi$ , we define that:

- **Confidence** refers to the percentage of records, which  $\chi$  holds, within the group of records for which  $I_j$  and  $\chi$  hold. To find hidden information we look for rules with average percentage, because the rules with a higher percentage usually represent known information.
- **Support** refers to the percentage of records which  $\chi$  and  $I_j$  all hold. When we look for new knowledge we want to find the rules with higher support.

As in the AR techniques used as a basis for this method, it is important to have a concrete idea of what we are looking for. This is shown in the selection of the subset of information to be analyzed and in the determination of the support and confidence. The interested reader is referred to [1] [4].

In short, below are the points where *AR* and *QuAsAR* differ:

- *QuAsAR* looks for the most and least known rules.
- The rules do not necessarily represent new knowledge.
- The process is not focussed on large set of variables, but on data used for decision making.

- The confidence level is defined with two intervals: one of confidence, whose superior benchmark is 100%, and another one of mistrust, whose substandard benchmark is 0%.
- There is no direct relation between support and confidence. We do not look for rules that attain both conditions.
- We look for rules with low levels of support.

The idea is to analyze the data in order to infer the candidate rules and to calculate levels of support and confidence. For instance, the rules that appear in a very small number of records (1%) may be potential entry errors and are candidates to be evaluated. Also the opposite relationship could be observed: rules that appear in a very large number of records may be considered as a business rule. That means that all the records that do not fulfill the rule could be a potential error.

We will now describe in more detail the *QUAsAR* technique. To do that, we will explain how it works with a simple example, which deals with a hospital database. We choose two variables: Illnesses and Doctors. In this particular domain, Illnesses and Doctors are repetitive and all doctors are specialists on a specific area. We suppose that Data Selection, Data Cleaning and Coding phases were already finished and we have a flat table like the following.

	Acute Rec. Hepatitis	Dr. White	Dr. Doe	Dr. Smith	Dr. Johnson
1	Yes	Yes	No	No	No
2	Yes	No	No	Yes	No
3	Yes	No	Yes	No	No
4	Yes	No	Yes	No	No
5	Yes	Yes	No	No	No
6	Yes	No	Yes	No	No
7	Yes	No	Yes	No	No
8	Yes	Yes	No	No	No
9	Yes	Yes	No	No	No
10	Yes	Yes	No	No	No
11	Yes	No	No	No	Yes

Illnesses variable appears only once, because we chose just one value. On the other hand, in each transaction there is only one doctor. Consequently, for each record in the flat table there is only one occurrence for the Doctors variable. The columns represent the different data; each record shows the value of each transaction. “Yes” is used to indicate the appearance of the data in the transaction and “No” to show its absence.

### 3.3 Confidence and Support Definition

We handle confidence and support in a different way than *AR* does [1] [4]. This is one of the main differences between both techniques. The definition of both intervals is directly related to business domain. Therefore, a brief analysis of the data gathered is required.

**Confidence** [1] [9]: If the domain variables are repetitive, like Illnesses and Doctors in the hospital database, it is possible to find that an antecedent appears several times with the same consequent. The confidence interval should be defined with a value near the upper benchmark, to obtain rules that are practically certain. On the other hand, it is necessary to consider a larger interval for confidence, to look for the less certain rules. The benchmark of the mistrust interval can be defined as the minimum value expected for any two possible variable values.

Confidence may be delimited with a confidence interval like (100; 95] and mistrust interval like [1;0). This definition supposes that if an antecedent appears 95 % of the time with a particular consequent, then the remaining 5% deserves to be evaluated. Accordingly, those rules with a confidence below 1% are potential errors. Although it is possible to define larger intervals, this would reduce the probability of finding errors.

**Support** [1] [9]: support allows the detection of data inconsistency analyzing the number of records where both antecedent and consequent are present. The rules with an average percentage of occurrences have less probability to be potential errors. Support is focused on the lower benchmark. For example, if a certain rule, with 100 % of confidence, appears in only 2 of one million records, it represents a potential error.

When defining support, the most important issue is the determination of the minimum number of expected occurrences for any value, that is, an estimation of the minimum frequency of occurrences of a variable value; if this parameter is unknown, it is possible to define a benchmark like 1/1000 occurrences, whereas if a datum appears in more occasions it will not be considered as “irregular”. In this case, the support is set to  $((1/1000)/\# \text{ quantity of records}) * 100$ .

In short, it is possible to “juggle” with different intervals in order to adjust the search. This depends on the business domain and on the set of rules looked for.

In the example described, Illnesses and Doctors are repetitive. As a result, the confidence - relationship between the antecedent and the consequent- is established within the intervals (100, 90] and [10, 0). The support was set as 10 %, which represents a threshold 1 record.

### 3.3.1 Pattern Recognition

Once the data are gathered and cleaned, and support and confidence are defined, the next step is pattern recognition. At this point, with the flat table constructed, we can apply any AR algorithm. In this case, we will use a QM adapted version of the Dupla Matricial [16] algorithm.

To continue with the example, we will explain a high-level version of the algorithm used. The first step is the calculation of the *matrix of occurrences*, which includes the number of times that any combination of two variable values appears on a record. The right table represents the *matrix of occurrences*.

	A	B	C	D	E
A	11	5	4	1	1
B		5	0	0	0
C			4	0	0
D				1	0
E					1

Once the *matrix of occurrences* is defined, we calculate the values of support and confidence for each rule.

	Support	Confidence
AB	45	50
BA	45	100
AC	36	40
CA	36	100
AD	9	10
DA	9	100
AE	9	10
EA	9	100

The table on the left shows support and confidence values for each pair of values. The next step is selection of candidate rules.

The selected rules are the following:

- AD, DA, AE y EA because they have less than 10 % support.
- AD y AE because the confidence is in the mistrust interval.

This means that:

- Dr. Smith looks after 10% of the ACUTE HEPATITIS, with 9 % support.
- Dr. Johnson looks after 10% of the ACUTE HEPATITIS, with 9 % support.
- 100% of the illnesses attended by Dr. Smith are ACUTE HEPATITIS, with 9 % support.
- 100% of the illnesses attended by Dr. Johnson are ACUTE HEPATITIS, with 9 % support.



### 3.3.2 Analysis

The most subjective task of the *QuAsAR* technique is the analysis of the rules generated as a result of the process previously described. This subjectivity is directly related to the different qualitative value that a specific piece of data might have for different users.

At this point, it is necessary to work in association with people that have the specific knowledge of the business involved. This does not prevent some rules from being validated using another source. For instance, geographic data could be checked using maps or other sources of geographic data such as satellite images.

Some variables must be analyzed in detail. These variables correspond to values that appear by default on the input-screen of the application used to capture data. In the example, the input-screen always suggests the same doctor by default, then this doctor will probably appear related to other illnesses rather than to his/her own specialization.

If *Metadata* information is provided, this information can be used in order to check for data constraints, inconsistency in data types, and so on. All those constraints should be present in the analyzed data.

To sum up, some of the issues to be considered during the analysis of the rules gathered are:

- To work with business experts to analyze rules
- To analyze the *Metadata* of information system – if data come from a specific application
- To check default values in input-screens
- To determine the existence of other information systems, standards, laws or other elements that allow analysis automation
- To divide rules by subject areas, to simplify business experts work

To finish the former example, we were able to detect that Dr. Smith was a cardiologist and that Dr. Johnson was gynecologist. Neither of them were related to cases of acute hepatitis, thus the four rules found were wrong. The other rules did not show any error.

### 3.4 Tools: Rules Finder

This is an integrity control tool based on data re-engineering [7]. The application was developed for this research. Its development was incremental and aimed to cover the needs generated during the case study.

This application allows automatic generation and filling of flat table and the calculation of support and confidence. Also a *QM* adapted version of the *Dupla Matricial* [16] algorithm was implemented.

## 4 CASE STUDY

### 4.1 Introduction

In order to evaluate method performance, we use it in a real case study. This case study also helps us to validate the assumptions made during the method definition: Support and confidence should be

analyzed independently and confidence analysis should address both confidence and mistrust intervals.

Data for test is owned by an oil and gas company. We selected a database from an information system that stores data from operations performed in oil wells. We chose a set of data that belongs to a specific geographic area, approximately 50,000 records.

The information system was chosen because data entry application did not perform any kind of validation or integrity checking. In addition, there was no documentation of the application and the *Metadata*. Also, it was not possible to recover foreign keys, even though there was a high-level description of the physical data model. These factors indicated that the database was error-prone.

Besides, we developed the *Rules Finder* application in order to create the flat table and calculate the candidate rules based on a given support and confidence.

#### 4.2 Definitions

The goal was the analysis of stored data for wells, drilling operations, and companies contracted to perform them. This information is used for performance analysis of both companies and equipment involved. This information represents the core of the data stored in the application, thus any error in these data directly impacts in the quality of the rest of system information.

Data Selection, Cleaning, Coding and Support and Confidence definition were performed with domain experts.

#### 4.3 Results

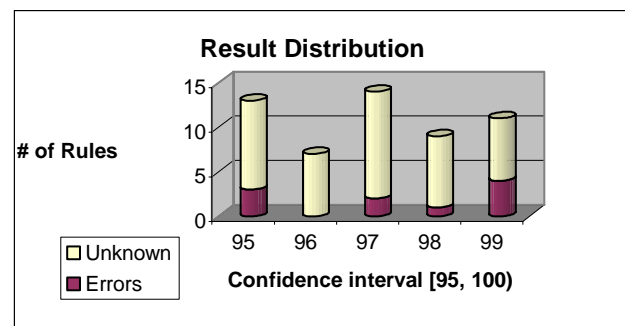
The result of this case study showed that more than 30% of the rules found were data errors. This outcome was better than we expected. From the point of view of domain experts, it was recognized as an important and productive task. There was a previous attempt to analyze data quality without a systematic method, which substantially complicated the task.

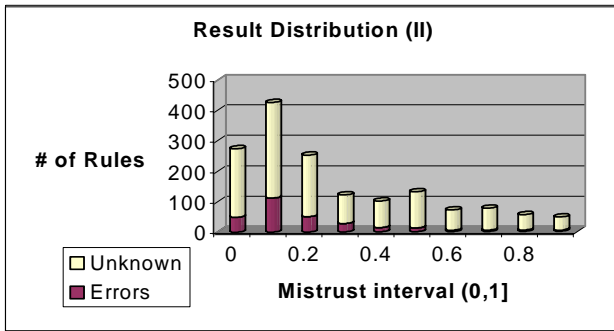
We present the results for support and confidence intervals. Each interval was analyzed separately.

From rules based on confidence intervals, it was possible to detect that 30% of them were errors. Although this number is high, it may be related to the fact that the application from which data came has no validation mechanisms. In addition, no integrity checks had been made before.

The chart on the right shows results distribution according to confidence interval. The rules are classified in two groups, the first one corresponds to rules that represent errors; the second one, named "Unknown", groups the correct and undetermined rules.

The number of rules in this interval was not significant, however we were able to detect several errors. This is a confirmation of our original assumption about the confidence interval: the analysis should not only be confined to those rules whose confidence is near the upper benchmark. Also, we did not find errors related to default values of data entry menus.

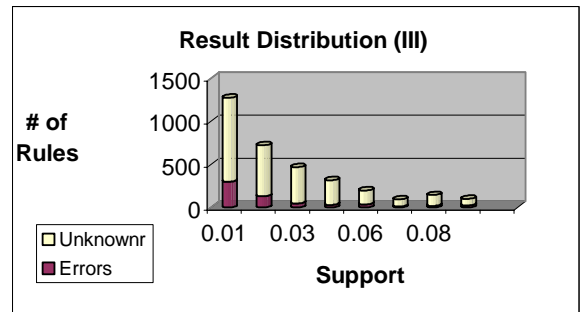




The results of mistrust interval were distributed as expected. The number of errors grows up while confidence decreases. More than 50 % of them appear in the (0, 0.1) interval. Going back to the definition of confidence, this means that in more than 99.9 % of records the antecedents were related to other consequent. That is the reason why the analysis of the mistrust interval is very

important.

The support analysis showed that more than 35% of rules found corresponded to data errors. Although this number is high, it was expected because application did not perform any validations.



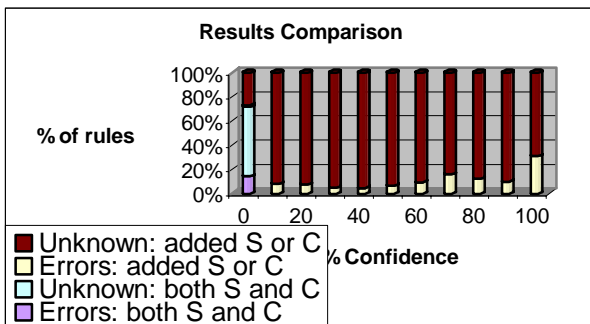
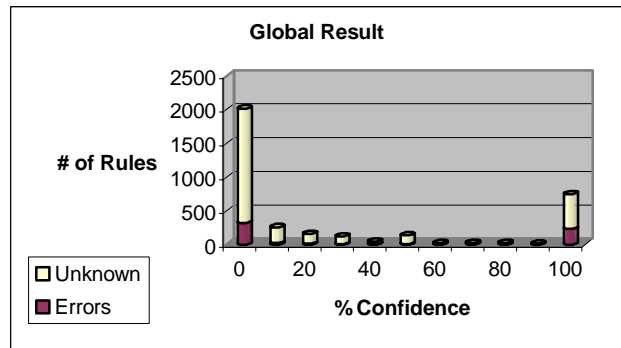
The number of errors increased while the support decreased. Also, when the quantity of occurrences grows, the probability of finding an error decreases. One of the most common errors found was the appearance of “irregular” values that did not correspond to any feasible value. The interested reader is referred to [17].

#### 4.4 Performance

When the method was defined two important assumptions were made. First, Support and Confidence should be analyzed separately, because we assume that it was possible to find more classes of errors.

The other assumption was the redefinition of confidence. We propose a two-interval approach, one of confidence and another of mistrust. Each interval focuses the search in a different way, “certain” versus “irregular”. In our case study the most efficient was the second one (0, 1]. Moreover, 52 % of the errors were found on it and 37 % on the other one [95, 100].

If support and confidence were analyzed together the result would be less significant. The right chart shows the global result distribution based on rules confidence. Although most errors were found on confidence intervals, some errors were also found in between.



The chart on the left shows the comparison between support and confidence analyzed individually and together. With our assumption we found 108% more errors, than analyzing the intervals separately. The interested reader is referred to [17].

## 5 CONCLUSION AND FURTHER WORK

We have developed a method to help organizations to verify data quality in a given context. Furthermore, this method has proved to be quite useful in evaluating data quality in general. This research also demonstrates that *DM* based techniques are useful in *DQ*. Like all the quality control mechanisms, *QM* does not resolve the problem completely. It should be used in combination with other existing techniques in order to achieve the desired results.

More theoretical work remains to be done in this area. From the *QuAsAR* standpoint, further work should focus on expanding confidence and support intervals. It is expected that the method can be extended and improved as it is applied to a wider variety of cases. Also, it is likely to find a relationship between these intervals and quality dimensions [13]. This would help to make a deeper analysis of the information, stressing it on specific dimensions.

Another possible extension of this technique is the definition of a mechanism to keep the rules updated without reprocessing or recalculating them. This may be developed using *Active Data Mining* techniques [12]. This may be useful in changing domains, where the rules generated could become outdated after a short period of time.

It is also possible to develop a preventive quality control technique [7]. For instance, making reports that show data not complying with rules, or adding automatic checks that prevent erroneous data from being entered; or implementing any external application that validates information and allows to correct invalid data, and so on.

The knowledge of rules to improve information can also have different uses. The mistake is to consider data quality as good without a previous check and realizing afterwards that wrong decisions were made.

## 6 REFERENCES

- [1] Adriaans P. and Zantinge D., *Data Mining*, Addison-Wesley, First Edition, 1996
- [2] Bobrowski M., Marré M., Yankelevich D., *A Software Engineering View of Data Quality*, European Quality Week, 1998
- [3] Dasu T., Johnson T., *Hunting of the Snark – Finding Data Glitches using Data Mining Methods*, In Proceeding of the 1999 Conference of Information Quality, Cambridge, Massachusetts, 1999
- [4] Kamal A., Stefanos M. and Ramakrishnan S., *Partial Classification using Association Rules*, American Association for Artificial Intelligence, 1997
- [5] Kismet Analytic Corporation, *Data Quality Methods*, White Paper, 1996
- [6] Little, R.J.A. and Rubin, D.B. *Statistical Analysis with Missing Data*, 1987
- [7] Melgratti H., Yankelevich D., *Tools for Data Quality*, Technical Report, 1999
- [8] Meyer P., *Introductory Probability and Statistical Applications*, Addison-Wesley, Spanish Revised Edition, 1992
- [9] Rakesh A., Tomas I. and Arun S., *Mining Association Rules between Sets of Items in Large Databases*, In Proceeding of the ACM SIGMOD Conference on Management of Data, Washington D.C., May 1993
- [10] Rakesh A., Ramakrishnan S., *Mining Generalized Association Rules*, In Proc. of

- the Very Large Database Conference, Zurich, Switzerland, 1995
- [11] Rakesh A., *Data Mining*, In Proceeding of the Very Large Database Conference, Buenos Aires, Argentina, 1998
  - [12] Rakesh A., Psaila G., *Active Data Mining*, In Proceeding of the *KDD* Conference, Portland, Oregon, 1995
  - [13] Redman T., *Data Quality for the Information Age*, Artech House, 1996
  - [14] Strong D., Lee Y., Wang Y., *Data Quality in Context*, Communications of the ACM, Vol. 40 No. 5, May 1997
  - [15] Tayi G., Ballou D., *Examining Data Quality*, Communications of the ACM, Vol. 41No. 2, February 1998
  - [16] Vázquez Soler S., Wilkinson H., *Mining Association Rules*, Technical Report, 1997
  - [17] Vázquez Soler S., *Data Mining para evaluación de Calidad de Datos*, Undergraduate thesis UBA, 2000
  - [18] Wang F., *Total Data Quality Management*, Communications of the ACM, Vol. 41No. 2, February 1998
  - [19] Wang R., Kon H., *Towards Total Data Quality Management*, MIT Working Paper Series, 1992
  - [20] Wizsoft, *Wizrules for Windows'95 User's Guide*, Version 3, 1997

# **MEDD: An Approximate Matching Technology for Database Searching, Linking, and De-Duplicating**

Arthur Goldberg and Andrew Borthwick

Practice-Oriented Paper

## **Executive Summary**

When you need to combine multiple, error-filled data feeds into a single, highly accurate database, the hardest problem is matching corresponding records. How do you match, for instance, "Thomas J. Hanks" with "Tom Hank" or "International Business Machines" with "Intl. Bus. Mach."? We present an innovative, accurate system that employs a powerful, patent-pending, machine learning technique to determine the probability that two database records correspond to the same person or company.

We start by showing why record matching is such a difficult problem and describe the basics of the record matching process. As an example, we discuss the New York City Department of Health, where we removed 300,000 duplicate records from a 2.1 million record children's health database.

MEDD is built around "comparison functions". Comparison functions check whether a pair of records has a certain matching or non-matching characteristic. Examples include "First names match", "First names match using the 'Soundex' phoneticization technique", or "Birthday does not match".

MEDD uses a training process called "maximum entropy modeling" to infer the relative importance of the different comparison functions from a small set of record-pairs which have been hand-marked as "same" or "different". Out of this process comes a "weight" which is assigned to each feature.

At runtime, MEDD operates as a function which takes a set of fields (a "search record") as an input and returns a list of database ID's which might match the search record. The ID's are ranked by a probability of match which is computed by MEDD's weighted comparison functions.

CHOICEMAKER

**MEDD**  
Maximum Entropy De-Duper

An Approximate Matching Technology for  
Database Searching,  
Linking and De-Duplicating

---

Prof. Arthur Goldberg, VP Marketing and Strategy  
Dr. Andrew Borthwick, President

### Approximate Record Matching

- Record matching tasks
  - Remove duplicates from a database
  - Link multiple databases
  - Search a database for a record
- Matching difficulties
  - No unique IDs
    - Some databases prohibit SSNs
  - Incorrectly entered data
    - Borthwick vs. Borthwick
  - Time-varying data
    - Address changes
  - Inconsistently used identifying data
    - Andrew vs. Andy

### Matching Catastrophes

NYC Department of Health Child DB	1.4M children duplicated into 2.1M records
Removing felons from Florida's voter roles	Some counties purged non-felons. Some counties did no purge because of list's inaccuracies
Wall street business data	Two clerks work full time matching by hand

### MEDD Matches Healthcare Data

- Client: NYC Department of Health
- Projects
  1. Remove immunization database duplicates
    - Prevent over and under immunization
  2. Link immunization and lead-exposure test databases
    - Enable caseworkers to address both under-immunization and lead exposure when visiting clients

### NYC Immunization Database

- Parameters
  - NYC birth cohort 122,000
  - Over 2M records
  - Monthly updates from 1,100+ institutions and providers
    - Up to 100,000 patients
    - Up to 200,000 immunization events
- Before MEDD: 3 records for every 2 kids
  - Strict criteria for automatic merging
- In 1998 clerks manually de-duplicated
  - 260,000 record pairs
  - 1,700 person-hours

### MEDD De-Duplicates NYC Immunization Database

- Work in 1999-2000

Birth year	Records	Dupes removed
<b>1996</b>	203,389	25,553
<b>1997</b>	216,336	34,773
<b>1998</b>	208,315	47,830
<b>1999</b>	157,946	42,228
<b>TOTAL</b>	785,986	150,384

### MEDD Links Two Databases

- Databases
  - Immunization
  - Lead exposure
- Synergy between the two programs
  - The same kids can be under-immunized and missing a lead screening test
  - Both databases cover all NYC children
- Finish in early 2001

7

### NYC MEDD/MCI System

- Information about every child in either database is stored in a MEDD-based Master Child Index (MEDD/MCI)
- Each system can retrieve data from the other by finding corresponding IDs in the MEDD/MCI

```

    graph TD
      LD[Lead Database] <-->|Data Exchange| ID[Immunization Database]
      LD -- Correlation --> MCI[MEDD Master Child Index]
      ID -- Correlation --> MCI
    
```

8

### MEDD/MCI Record Matching

- Remove duplicates
- Connect immunization and lead exposure children
- Determine whether incoming records are already in MCI
- Periodically scan MCI for residual duplicates

9

### NYC DOH's Benefits from MEDD

**Savings**

- Automatically removed 200,000 records in '99-'00
  - Original process would have required hand-examining at least 600,000 record-pairs
  - Cost of 2 person-years
- To summer '01, almost 600,000 records removed

**Improvements**

- Matching incoming records prevents creation of duplicates
- Enabled linkage of immunization and lead databases
- Old process was much less accurate
  - Error rate of a typical clerk is over 1%
  - Clerks only reviewed very similar records. Many "tricky" matches were never reviewed
- DOH accepting "noisy" data feeds (billing feeds from HMO's, forms filled out in doctor offices)

10

### Production Matching Basics

**Input** Search record

**Blocking**

- Find thousands of possible matches

**Match decision making**

- For each possible match
  - Evaluate many comparison functions against search record
  - Combine comparison functions by weight to produce match probability

**Output** IDs and probabilities of likely matches

11

### Production Matching

```

    graph TD
      SR[Search Record] --> B[Blocking]
      B --> MPM[Many Possible Matches]
      MPM --> MEM[Maximum Entropy Matching]
      MEM --> MPLM[Match Probabilities of Likely Matches]
      MPLM --> MP{Match Probability}
      MP -- Low --> NM{{Non-Match}}
      MP -- High --> M{{Match}}
      MP -- Intermediate --> HR{{Human Review}}
    
```

12



### Comparison Function Examples

#### Database of Children

- Do first names match?
- Do first names match approximately using "phonetic matches" such as Soundex, edit-distance, NYSIIS, or Jaro-Winkler?
- Do uncommon first names match?
- Do we have an indicator that the child is part of a multiple birth?
- Do Medicaid numbers match or mismatch?
- Do birthdays match?

13

### Comparison Function Examples

#### Database of Businesses

- How many words in the name match?
- Can the names be converted to the same abbreviation?
- Are the names the same after translating foreign words to English?
- Do country, phone number, or street address match?

14

### Complex Comparison Functions

#### Adapt to database quirks

**Child medical database example**

**HMO XYZ sends Day of Birth = "1"**

Birthday = July 1, 1998 not July 15, 1998

**A special comparison function**

IF Provider = "HMO XYZ"  
 AND Day of Birth = 1  
 AND dates differs only on day of birth  
 THEN **Match**

15

### Customized with Java

#### Java-based Comparison Functions

- Simple first-name Soundex comparison function:

```
feature firstNameSoundexMatch {
    match equals(soundex(FIRST_NAME));
}
```

- Comparison function for the HMO example on the previous slide:

```
feature HMOXYZandFirstOfMonth {
    match ((q.FACILITY_ID == "XYZ" && q.DOB.getDay() == 1) ||
           (m.FACILITY_ID == "XYZ" && m.DOB.getDay() == 1)) &&
           q.DOB.getMonth() == m.DOB.getMonth() &&
           q.DOB.getYear() == m.DOB.getYear();
}
```

16

### Maximum Entropy Matching Math

- The probability a pair of records match

$$\frac{\text{MatchProduct}}{\text{MatchProduct} + \text{No-MatchProduct}}$$

MatchProduct = product of weights of all comparison functions predicting **Match** for the pair

No-MatchProduct = product of weights of all comparison functions predicting **No-Match** for the pair

17

### MEDD Decides Match

#### 99.5% Confidence

Field Name	Record		Match?	Weight
	1	2		
Last name	Smith	Smith	Match	1.153
First name	Emily	Emely	No-match	1.350
Soundex First name	EML	EML	Match	4.708
DOB	4/28/97	4/28/97	Match	1.138
Street	4528 3 <sup>rd</sup> Ave	4528 3 <sup>rd</sup> Ave	Match	4.342
City	Bronx	Bronx	Match	1.103
State	NY	NY		
Zip	10462	10462	Match	3.013
Phone	718-123-4567	718-123-6789	No-match	2.130
Med Rec Number	11856437503	11856437503	Match	6.587

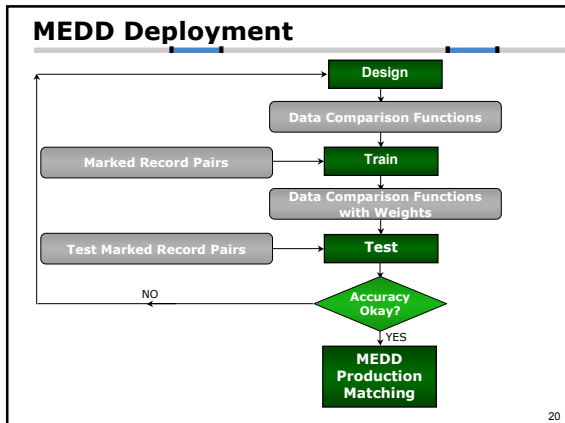
Match product = 587.2       $\frac{587.2}{587.2 + 2.9} = 0.995$   
 No-Match product = 2.9

18

### MEDD Decides No-match 97.9% Confidence

Field Name	Record		Comparison	Weight
	1	2		
Last name	Lopez	Lopez	Match	1.153
First name	Girl	Susan	No data	
Soundex First name				
DOB	1/11/97	1/2/97	No-match	28.949
Street	987 Cornelia	456 Park	No-match	2.937
City	Brooklyn	Brooklyn	Match	1.103
State	NY	NY		
Zip	11211	11211	Match	3.013
Phone	718-123-4567	718-234-5678	No-match	2.130
Med Rec Number	1001002	567435		

$MatchProduct = 3.8$   
 $No-MatchProduct = 181.1$   
 $\frac{3.8}{181.1 + 3.8} = 0.021$



### Principles of Maximum Entropy

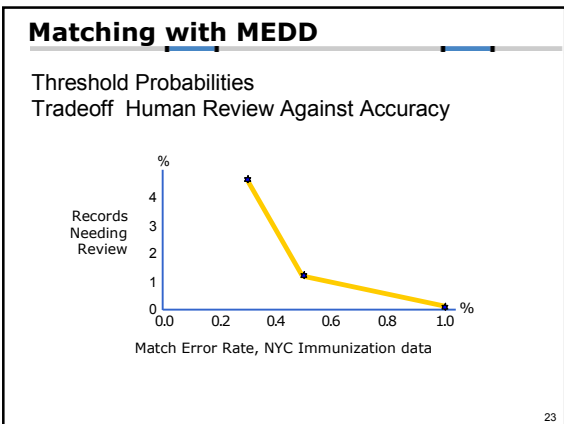
**How are weights determined?**

- Input record pairs marked **Match** or **No-match**
- Weights selected so model predicts average probability of match for each comparison function equal to average probability for that comparison function in training data

**Name**  
Probability records match given that name matches = 2/3

**Phone**  
Probability records match given that Phone matches = 7/9

# Demo



### Technical Information

**Platforms**

- Win32, Linux, Solaris, and other UNIX

**Modes of operations**

- Online as a CORBA/EJB/RMI/COM Module
- Batch mode with a flat file input
  - For one-time runs

**Available for Oracle, other DB's to follow**

**System is delivered fully customized for the client's database by ChoiceMaker staff**

## ChoiceMaker

### Management

- Andrew Borthwick, President
  - Designed and implemented MEDD
  - NYU CS PhD 1999
  - Expert on maximum entropy modeling
- Arthur Goldberg, VP Strategy and Marketing
  - NYU CS Professor, co-director MSIS graduate program
  - Expert on network performance
  - Five years at IBM Research
- Staff includes three other Ph.D. computer scientists

### Funding

- NSF Small Business Innovation Research Grant
- Investment from CCS, a \$120M Japanese software firm

25

## MEDD Features

### Easy to Understand

- MEDD outputs a match probability, unlike other systems which output a "score"

### Highly Customizable

- Powerful Java-based environment for creating custom comparison functions
- Advanced machine learning technology learns the human intuition for computing overall probability that a record-pair matches

### Highly Accurate

- NYC DOH measured it as equivalent to two clerks working together

26

CHOICEMAKER

## Questions

[Arthur.Goldberg@choicemaker.com](mailto:Arthur.Goldberg@choicemaker.com)  
[Andrew.Borthwick@choicemaker.com](mailto:Andrew.Borthwick@choicemaker.com)  
212 905-6031  
ChoiceMaker Technologies, Inc.  
41 East 11th Street, 11th Floor  
New York, NY 10003  
[www.ChoiceMaker.com](http://www.ChoiceMaker.com)

## **Cleaning up Very Large Databases and Keeping Them Clean**

**Priscilla Broberg**

U.S. Caden (a division of Manpower)  
currently working at Agilent Technologies

### **Executive Summary**

This presentation shows a real-world example of how a very large Customer database was cleansed and de-duplicated to shrink it down to a manageable size. The techniques used to do this are shown, as well as the processes that were implemented to maintain the new level of data cleanliness. The tricks and techniques are applicable to customer files or databases of any size in any business. Actual before and after data examples are shown.

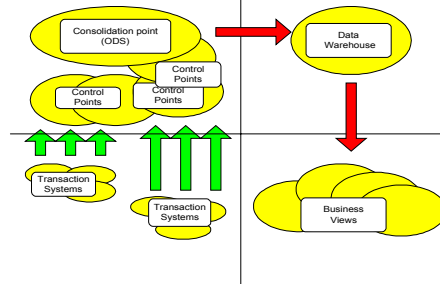
Topics covered include:

- 
- Typical customer data flows, from data entry to reporting
- Proper placement of data cleansing and merging in the data flow
- Techniques to maximize effectiveness of merge/purge (de-duplication)
- Ideas for maintaining a higher level of data cleanliness, and minimizing data duplication.

## Cleaning up Very Large Databases and Keeping Them Clean

The story of how a customer database got very large and very messy, then got small and clean again.

## Data Warehousing Architecture



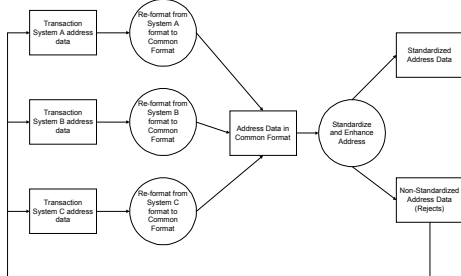
## The Consolidation Point Provides Clean, De-Dupped Data to the Warehouse

- Cleanses data
- Standardizes data
- Enhances data (e.g. zip+4)
- Eliminates duplicates (merge/purge)
- Communicates back to transaction systems
  - rejected transactions
  - successfully loaded transactions

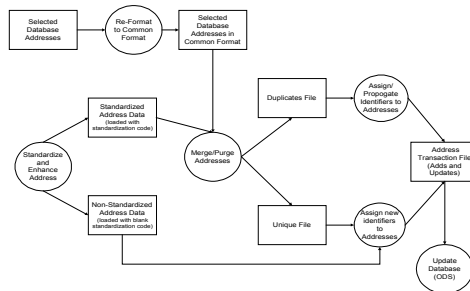
## WHY DO WE NEED TO MERGE/PURGE CUSTOMER DATA?

- Data from separate transaction systems is entered and identified differently
- Need for company-wide view of customers ("Master list")
- Need to consolidate customer information Worldwide
  - avoid double counting
  - save on database storage
  - able to identify one customer with one unique identifier (cross-referenced to source systems)

## Standardization Process Flow



## MERGE/PURGE PROCESS FLOW



## How Did We Get Into This Mess?

- ODS Database designed in late 1980's to cleanse and load a single type of customer data - Order Processing Customers. Data only went to one application for reporting. ALL records were required to be loaded, regardless of data quality!
- Later, additional sources of data, as well as receiving applications were added for Direct Marketing. These were allowed to be rejected, if they did not meet data quality standards.
- Merge/Purge rules changed.
- Moved from Mainframe to Unix platform, and changed cleansing and merge/purge tools.
- ODS Database had no delete capability. All data was added or updated, then remained there forever!
- Only incoming transactions were cleansed and merge/purged against the database.
- Once data was loaded, it was never re-cleansed or re-merge/purged.

## Other Contributing Factors

- Many records were coded with the wrong country code. Only those with US country codes (US, and affiliates such as Puerto Rico, Guam, etc) went through standardization!
- We have no edit for verifying the country code against the address. We just accepted what was input.
- Once a record is loaded as non-standard, it NEVER participates in the merge/purge.
- Non-standardized records contributed to a lot of data duplication.
- Split from Hewlett-Packard caused us to inherit a database full of HP customers as well as Agilent customers. There was no attribute of the customer data to tell them apart.

## Preparing for The BIG CLEAN-UP

### Step #1: Pre Clean-up

- Removed data associated with Direct Marketing
  - Identified by Source Number
  - Had to make sure data was not also associated with active sources.
- Documented current Merge/Purge rules and reviewed with users
- Using a Marketing Reporting Tool (the Data Warehouse recipient of our customer data), we were able to identify which customers belonged to Agilent by reporting customer numbers on orders with Agilent product lines.
- Identified customers who had been active in the past two years, and deleted all others.
- Number of site (address records) after clean-up went from approximately 11 million rows to 1.1 million rows.
- This became the starting point for our re-standardization and re-merge/purge.

## Preparing for the BIG CLEAN-UP

### Step #2: Analyze remaining data

- Determine how much data is US, how much Canada, and how much non-US.
- Country code not reliable. However, we used ACE to discover this, and locate the incorrectly coded records!
- Perform test merge/purge runs on non-US/non-Canada data, using line1, line2, etc. method.
- Adjust merge/purge parameters based on results of test runs.

## Clean-up Steps

### Data-Cleansing

- Country Code clean-up must be done first. Since this is part of the match-key, re-calculate match-key.
- Re-standardize US and Canada. After re-standardization, re-calculate match key again. (Postal code is also a component of match-key)
- Update database with new country codes and match-keys, as well as newly-standardized addresses.
- Our match-key algorithm: First letter of Business Name, followed by first four numbers of address, followed by first 3 bytes of postal code, followed by 3-byte country code. '@' used as filler where no data exists.
- Example: IBM 123 Main Street, Anytown, Anystate, 99999 would be coded as: !123@999000 ('000' is our country code of US).

## Example -Before Standardization

3009 NW 75 AVENUE	MIAMI FLORIDA 33122	ATTH: LILIANA P. VELAZQUEZ 351	000000333351
7200 NW 7 STREET 2ND FLOOR	MIAMI FLORIDA 33126	351	000000333351
C/O FARMCOX 105255	7051 NW 37 STREET	MIAMI, FL. 33166-6559	301 1331000001
14413 IMPORT DR.	LARDO, TX 78041 USA	201	07804144201
C/O MIAMI PANALPINA INT	3505 N.W. 107TH AVE.	MIAMI, FL 33178	355 C311733355
2100 BLUE LAGOON DRIVE SUITE 1050	MIAMI FLORIDA 33126	355	00000105255
10777 WESTHEIMER STE 625	HOUSTON TX 77042	351	00000107351
1900 CONCOURSE DRIVE	SAN JOSE, CA 95131	223	00000119023
420-B FARMERICHON DRIVE	EL PASO, TX 79907	201	07910739201
2429 TERMINAL BLVD.	MOUNTAIN VIEW, CA 94043	357	00000242357
400 REIMANN AVENUE	SANDWICH IL 60548-0900	412	C0548000412
1310 MEMOREX DRIVE	SANTA CLARA, CA 95050	583	00000133583
SUITE 118	950 RICHARD AVENUE	SANTA CLARA CA 95050	583 0050505083
2712 EAST MERALOMA AVE	690 MAIN STREET	STRATFORD, CT 06615-0129	489 05500000489
ACCOUNTS PAYABLE	ANARHEIM, CA 92806	549	10000273549
1101 CYPRESS CREEK ROAD	2712 E MERALOMA AVENUE	ANARHEIM CA 92806	549 1273273549
6781-R SIERRA COURT	CEDAR PARK TX 78613	405	07861278405
6780-R SIERRA COURT	DUBLIN, CA 94568	427	00000945427
ATTH: TIMOTHY BOB SMITH	DUBLIN CA 94568	427	00000945427
ACCOUNTS PAYABLE	1305 E ALGONQUIN ROAD	SCHAUMBURG IL 60196	428 06105130428
	1301 E ALGONQUIN ROAD	SCHAUMBURG IL 60196	428 060105130428



## Clean-up Steps Eliminating Duplicate Data

- If required to maintain record of eliminated data, use the dup groups to create elimination transactions. An elimination transaction is basically like a “change of address” transaction. All that is needed is the old address identifier and the new (surviving) address identifier.
- If this is NOT required, delete any addresses not in your “mail” file, and you are done!

## Clean-up Steps Eliminating Duplicate Data (cont)

- Steps to performing eliminations:
  - 1) Create elimination transactions from dup groups
  - 2) Apply eliminations to all tables in which the address identifier is used. For example, our database uses this identifier in X tables. Change old identifier to new identifier, based on transaction.
  - 3) Once all tables have been updated, create a row in an “elimination table” to keep track of this change (old ID --> new ID)
  - 4) Finally, delete old (eliminated) address record

## Sample Elimination Transactions (Created from Sample Dups File)

Old ID	New ID
017588008	017508378
018679713	018660122
014707877	014268700
017626543	017037130
017784498	017037130
017818072	017037130
017907210	017037130
007083284	003488990
017023274	010279043
017095742	010279043
017819768	010279043
018408512	018180551
007083034	002223280
015338690	002223280
017140834	017038804

## Lessons Learned

- It is important to understand your current data flow and processing. If you haven't documented it thoroughly, start now!
- Make sure your users understand the data-cleansing and merge/purge rules. They own these!
- Know what data you have control over, and what data you do not. For example, we can clean-up data, but we cannot force the source systems to send us clean data.
- For best results, re-standardize all addresses in database whenever you get a new zip+4 update file from Firstlogic.
- Re-merge/purge entire database at least 4 times a year.
- Use Firstlogic tools to analyze your data, as well as to cleanse it in production.
- Don't assume you cannot merge/purge non-US data. It can be done quite effectively using the user-definable fields (Merg\_Purg1, Merg\_Purg2, etc).
- Read the Firstlogic Software Update Bulletins and Customer Care Bulletins that come with your upgrades. There may be new features you can take advantage of!

## Improvements/Benefits

- Reduced address rows in database from 11 million to < 2 million
  - Benefits:
    - Less disk space usage
    - Easier database administration
    - Faster processing times, as data merge/purges against fewer rows
    - Improved data quality, as duplicates are eliminated
    - Better decision making, as user confidence in data improved
    - Improved processing times on downstream systems, as less data is passed to them

## Cost Savings

- Support went from 3 full-time programmers rotating on-call duty (24/7), to Call-center, with 1 on-call “deep support” programmer.
- Call-center support much less expensive
- Support programmers became available to work on new projects.
- Went from one full-time DBA to one part-time DBA.
- Lowered disk space costs
- Lowered processing (machine time) costs
- Estimated total annual savings: \$500,000



## **A Framework for Information Quality in a Data Warehouse: IQ in the context of Data Marts and Data Warehouses**

Jonathan Wu

BASE Consulting Group, Inc.

475 14<sup>th</sup> Street, Suite 600

Oakland, California 94612-1900 USA

(510) 628-3300 Ext. 224, (510) 628-3311

[jwu@baseconsulting.com](mailto:jwu@baseconsulting.com)

Practice-Oriented Paper

### **Executive Summary**

Data warehousing technology provides integrated data from a multitude of sources that is non-volatile and transformed into meaningful information for decision-making purposes. As organizations embrace data warehousing technology as a means of accessing information, the need for quality information within a data warehouse is imperative to the sustained success and use of this technology. There have been several instances where poor quality of the data within a warehouse has led directly to the abandonment of this technology. By understanding the process flow of data from its source of origin through the various stages of manipulation and into the data warehouse, the potential for data errors can be mitigated.

While the quality of information within transactional systems must be addressed because it directly impacts the quality within a data warehouse, the process flow of data from source to target is of greater concern for information quality due to the various stages of data movement and manipulation. By developing a framework of information quality within a data warehouse, issues with data quality can be identified and addressed in a timely manner. The benefits of an established framework include: 1) establishing confidence that the data warehouse contains quality information, 2) identifying data issues from the source systems, 3) discovering changes in business or system processes that have not been reflected in the data transformation process, and 4) providing the group responsible for maintaining the data warehouse with the means of addressing user questions concerning data integrity.

The 6<sup>th</sup> International Conference on

## A Framework for Information Quality in a Data Warehouse

Jonathan Wu  
Co-Founder  
BASE Consulting Group, Inc.

Version 2.10 29 August 2001

Information Quality

## Presentation Abstract

Data warehousing technology provides integrated data from a multitude of sources that is non-volatile and transformed into meaningful information for decision-making purposes. As organizations embrace data warehousing technology as a means of accessing information, the need for quality information within a data warehouse is imperative to the sustained success and use of this technology. There have been several instances where poor quality of the data within a warehouse has led directly to the abandonment of this technology. By understanding the process flow of data from its source of origin through the various stages of manipulation and into the data warehouse, the potential for data errors can be mitigated.

While the quality of information within transactional systems must be addressed because it directly impacts the quality within a data warehouse, the process flow of data from source to target is of greater concern for information quality due to the various stages of data movement and manipulation. By developing a framework of information quality within a data warehouse, issues with data quality can be identified and addressed in a timely manner. The benefits of an established framework include: 1) establishing confidence that the data warehouse contains quality information, 2) identifying data issues from the source systems, 3) discovering changes in business or system processes that have not been reflected in the data transformation process, and 4) providing the group responsible for maintaining the data warehouse with the means of addressing user questions concerning data integrity.

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 2

## Presentation Information

- ◆ **Author:** Jonathan Wu
- ◆ **Organization:** BASE Consulting Group
- ◆ **Presentation Title:** A Framework for Information Quality in a Data Warehouse
- ◆ **Contact Information**
  - E-mail: [jwu@baseconsulting.com](mailto:jwu@baseconsulting.com)
  - Phone: (510) 628-3300 x224

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 3

## Agenda

- ◆ **Overview of Data Warehouse Process Flow**
  - Data from Source Systems
  - Data Migration
  - Data Cleansing
  - Data Transformation
  - Loading the Data Warehouse
  - Reconciling the Data Warehouse
- ◆ **Data Control Points as a Framework for Information Quality**
  - Prevent Controls
  - Detect Controls
- ◆ **Summary**

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 4

## Overview of Data Warehouse Process Flow

The diagram illustrates the data flow from Source Systems (represented by various colored cylinders) to a Staging Area (2) and (3), then to Data Warehouse Tables, and finally to the Data Warehouse. The process is numbered 1 through 5: 1. Migrating the data, 2. Cleansing the data, 3. Transforming the data, 4. Loading the data warehouse, and 5. Reconciling the data warehouse.

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 5

## Data from Source Systems

The diagram shows five types of source systems: Legacy Data Store (purple cylinder), ERP Application (red cylinder), Custom Applications (cyan cylinder), Legacy Systems (light blue cylinder), and Flat Files from External Sources (orange folder icon).

Source Systems

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 6

## Data Migration

Data Migration

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 7

## Data Cleansing

### Customer Information

FIRST NAME	LAST NAME	COMPANY NAME	AREA CODE	PHONE
JIM	Kirk	ibm	212	5551212
Jim	Kirk	ibm	212	5551212
James	Kirk	IBM	212	5551212
JAMES	KIRK	ENTERPRISE	212	5551212
↓	↓	↓	↓	↓
Martin	Zweig	Zweig Funds	415	5551212

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 8

## Data Transformation

**CUSTOMER\_CONTACTS**

FIRST\_NAME  
LAST\_NAME  
AREA\_CODE  
PHONE  
ADDRESS1  
ADDRESS2  
ADDRESS3  
CITY  
STATE  
ZIP\_CODE

**CONTACT\_INFORMATION**

FULL\_NAME  
PHONE  
ADDRESS  
CITY  
STATE  
ZIP\_CODE

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 9

## Loading the Data Warehouse

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 10

## Reconciliation Process

Company XYZ						
Statement of Comparison between the Data Warehouse (DW) and Source Systems (SS)						
As of August 29, 2001 8:00:58						
TABLE NAME / COLUMN NAME	NUMBER OF ROWS			CONTROL TOTALS		
	DW	SS	DIFFERENCE	DW	SS	DIFFERENCE
GL ACTUAL BALANCES PERIOD_YEAR	980,029	980,029	-	1,856,732,942.00	1,856,732,942.00	-
GL BUDGET BALANCES PERIOD_YEAR	1,053,906.00	1,053,906.00	-	2,104,023,966.00	2,104,023,966.00	-
GL JOURNALS CREDIT	8,350,661.00	8,358,714.00	(8,053.00)	304,016,413,952.71	309,236,482,066.29	(5,220,068,113.58)
GL_CODE_PERIODS CHART_OF_ACCOUNT_NUM	140,271.00	140,271.00	-	14,167,371.00	14,167,371.00	-

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 11

## Agenda

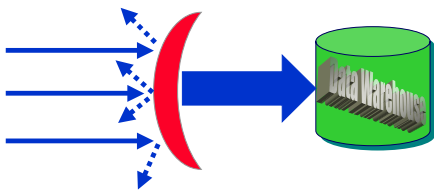
- ◆ Overview of Data Warehouse Process Flow
  - Data from Source Systems
  - Data Migration
  - Data Cleansing
  - Data Transformation
  - Loading the Data Warehouse
  - Reconciling the Data Warehouse
- ◆ Data Control Points as a Framework for Information Quality
  - Prevent Controls
  - Detect Controls
- ◆ Summary

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 12

### Data Control Points

◆ **Prevent Controls**

- Controls over the accuracy and completeness of data **before** it is loaded into the data warehouse.




Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 13

### Data Control Points

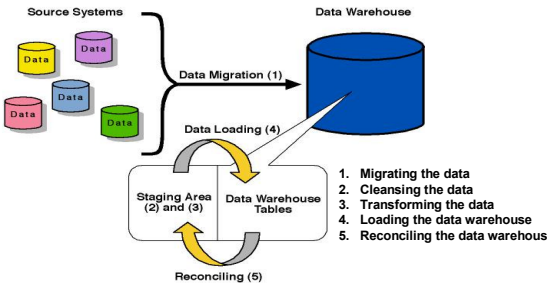
◆ **Detect Controls**

- Controls over the accuracy and completeness of data at the completion of each stage or **after** it is loaded into the data warehouse.



Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 14

### Overview of Data Warehouse Process Flow



1. Migrating the data
2. Cleansing the data
3. Transforming the data
4. Loading the data warehouse
5. Reconciling the data warehouse

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 15

### Data Control Points

**1. Migrating the data [Prevent Control]**

Goal - Prevent meaningless information by not moving it.

Motto - *"When in doubt, leave it out."*

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 16

### Data Control Points

**2. Cleansing the data [Prevent Control]**

Goal - Prevent unwanted redundant data by comparing and incorrect data by validating.

Motto - *"Cleanliness is next to godliness."*

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 17

### Data Control Points

**3. Transforming the data [Prevent Control]**

Goal - Prevent meaningless data by transforming it.

Motto - *"Business rules."*

Copyright © 2001 BASE Consulting Group, Inc. - All Rights Reserved 18

## Data Control Points

### 4. Loading the data warehouse [Prevent Control]

Goal - Prevent unwanted data through conditions and filters.

Motto - *"If the data does not fit, you must omit."*

## Data Control Points

### 5. Reconciling the data warehouse [Detect Control]

Goal - Detect data quantity and quality exceptions by reconciling.

Motto - *"If at first you don't prevent, reconcile, reconcile, reconcile."*

## Agenda

### ◆ Overview of Data Warehouse Process Flow

- Data from Source Systems
- Data Migration
- Data Cleansing
- Data Transformation
- Loading the Data Warehouse
- Reconciling the Data Warehouse

### ◆ Data Control Points as a Framework for Information Quality

- Prevent Controls
- Detect Controls

### ◆ Summary

## Summary

The success of a data warehouse rests with the users' perceptions of it.

If the data is incorrect or incomplete, user confidence and use of the data warehouse will diminish.

# **Monitoring and Data Quality Control of Financial Databases from a Process Control Perspective**

(Practice-Oriented Paper)

Janusz Milek<sup>†‡</sup>, (janusz.milek@predict.ch)

Martin Reigrotzki<sup>†</sup>, (martin.reigrotzki@predict.ch)

Holger Bosch<sup>†</sup>, (holger.bosch@predict.ch)

Frank Block<sup>†</sup>, (frank.block@predict.ch)

<sup>†</sup>PREDICT AG, Reinach BL, Switzerland, (www.predict.ch)

<sup>‡</sup>Automatic Control Laboratory, ETH Zürich, Switzerland, (www.aut.ee.ethz.ch)

## **Abstract**

The paper presents the application of several process control-related methods to the monitoring and control of data quality in financial databases. The quality control process itself can be seen as a classical control loop. Measurement of the data quality is conducted via application of quality tests, which exploit data redundancy defined by meta-information or extracted from data by statistical models. Appropriate processing and visualization of the test results enable human or automatic diagnosis of possible data quality problems. Selected model-based process monitoring methods are shown to be useful for detection, diagnosis, and, in some cases, also compensation of data quality problems. The test results are of interest not only for data quality control but also for business-relevant information extraction and monitoring. The presented methods are incorporated into our DQontrol product [1], and have been applied in the monitoring of a productive financial database at a customer site.

## **1. Introduction**

Information quality is one of the most important factors determining quality of conclusions drawn using consolidated data. Hence, it is necessary to continuously measure and improve the information quality. Huge financial databases, containing terabytes of data and invaluable information amounts, particularly need detailed and efficient monitoring approaches to extract useful information and distinguish it from artifacts and errors, which have to be eliminated.

Usually, the databases contain data from diverse sources which are loaded on a periodic and partially manual basis. Systematic data quality monitoring of such databases requires automated quality testing, result visualization, diagnosis, and compensation of data quality problems. Due to the truly industrial scale, high relevance, and hierarchical structure, huge databases can be treated similarly to large industrial processes. This analogy can be helpful to arrive at useful monitoring

approaches. Moreover, it can be expected that modern statistical process monitoring methods can be particularly useful to monitor databases. These methods exploit spatial and temporal redundancy of the data.

Note that modern process monitoring systems are not just limited to (i) fault detection, but may also include the following further stages: (ii) fault isolation, (iii) diagnosis, and (iv) compensation, so that their application gives rise to fault-tolerant measurement and control systems, see [11]. Counterparts of the mentioned stages in database quality monitoring systems may improve not only the quality but also fault-tolerance.

## 2. Some Elements of the Quality Control Loop

Left picture in Figure 1 shows the information flow in an example database containing financial data, e.g., of a bank, telecommunication, or insurance company. The data which usually come from heterogeneous sources, are extracted, transformed, and loaded into the database, before being delivered to the users. Data quality can decrease at each stage if the corresponding operations are disturbed in some way. Such a disturbance will be called a *fault*. The process of ensuring high data quality can be treated as a control system (Figure 1, right picture), composed of the measurement elements, controller, and actuators. The system must cover the whole information processing chain, from the data capture to the end user delivery.

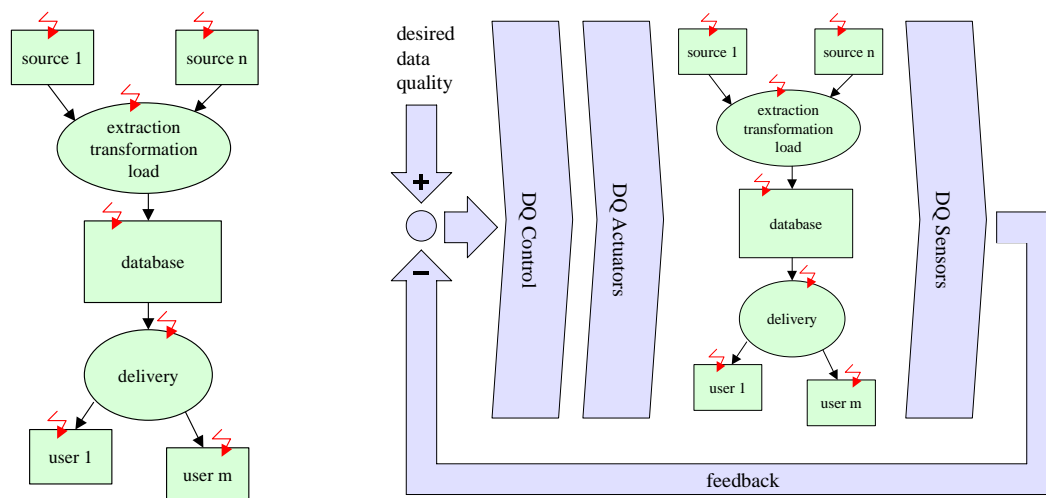


Figure 1: Example data flow (left) and information quality control system (right)

DQ Sensors measure data quality by running a number of quality tests. DQ controller analyses the test results, performs diagnosis and schedules appropriate data quality improvement actions. DQ actuators implement these actions. The goal of the overall feedback is to enforce the desired data quality. Some elements of the quality improvement process, for example data quality measurement, visualization, documentation, fault diagnosis, or compensation of simple faults can be conducted autonomously in a fully automatic way. Other, more complex elements of the process, like analysis and compensation of complex data quality problems cannot be automated and will not be considered here.

## 2.1. Factors Influencing Data

The data in a financial database can be influenced by the following factors: (i) individual customer behavior, (ii) market-related variations, (iii) seasonal variations, (iv) data quality issues. One goal of the monitoring can be to detect and distinguish all these types of factors. Basic data quality deficiencies (like missing values, data formats, and code tables) are most visible in the fault-related dimensions and can be easily identified using simple formal tests. Seasonal and market variations can be modeled using time series analysis, where individual customer variations can be analyzed e.g., using data mining methods [13]. The individual variations are suppressed using the later described data aggregation technique.

## 2.2. Quality Measurement and Classification of Tests

Data quality can be measured by performing appropriate tests. The tested piece of information is compared to the reference information. The generic test process is shown in Figure 2.

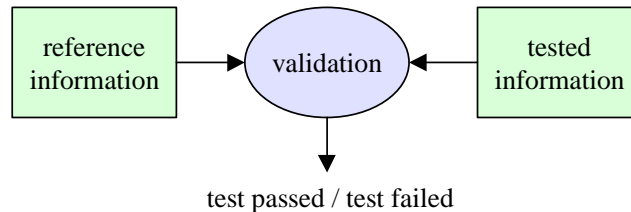


Figure 2: Principle of testing

There are two basic types of reference information:

- Meta-information, independent from the data and having the form of *strict* relations. Examples of reference information are: field validity (missing), field formats, code tables, keys, reference relations, strict business relations (like  $a = b + c$ , or  $a$  stays constant). The test related to meta-information can be called *technical*.
- Statistical models, obtained from reference fault-free data and having the form of *approximate* relations. Tests using statistical information are called *statistical* or *model-based*. Examples of models include mean, histogram, correlation, time series analysis model, as well as approximate business relations ( $a$  similar to  $b$ ,  $a$  changes slowly).

The tests can be also classified according to the number of involved (1) variables (univariate/multivariate tests), (2) records (single/multi-record tests), and (3) tables (single/multi-table tests).

## 2.3. Fault Signatures

Fault signature depicts the *absolute* or *relative* number of cases when a given test failed (due to some data quality problem), aggregated in the *aggregation dimensions* and presented in the *presentation dimensions*. The latter should be discriminative with respect to faults and informative to business/user-related applications. Example presentation dimensions may include time, partition (if, for technical reasons, the data are partitioned), customer segment, product, and



subsidiary. Figure 3 shows a color-coded example of absolute test signature in customer segment/time coordinates.

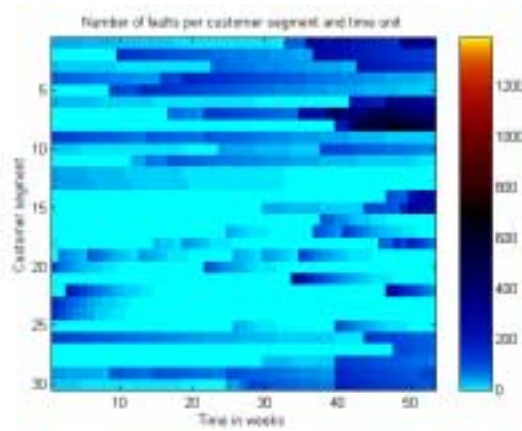


Figure 3: Example fault signature

Fault signatures enable simple fault visualization, assessment of data quality, and as such can be useful for fault elimination/correction, since it is simple to isolate data subsets having low data quality, or analyze data quality as a function of time.

#### 2.4. Classification of Meta-Information and Technical Tests

Meta-information describes strict relations which must be satisfied by the data. Examples of such relations are given in Table 1.

Level	Test type
Field level	missing value, format, code table, ranges
Referential integrity	keys, duplicates
Business-related	univariate and multivariate strict business-related dependencies for one customer or account, for example: variable is equal (less than ) to sum (concatenation) of other variables, a variable is equal to number of corresponding entries in another table, a variable is constant in time, a variable equals another variable shifted in time, etc.

Table 1: Meta-information and the related tests

Unfortunately, obtaining complete meta-information can be difficult and time consuming, especially if data are loaded from several sources. Related data quality problems are common in such a case; the only solution is to continuously collect and update the meta-information.

## 2.5. Classification of Statistical Models and Statistical Tests

Statistical monitoring methods comprise ideas belonging to econometrics, process monitoring, and data mining [12]. The main assumption of the statistical monitoring is that certain statistical data properties do not depend on time (*continuity assumption* [8]). Usually, the monitoring procedure comprises two steps. First, the selected statistical properties are estimated from available reference (fault-free) data. These properties constitute the model. Then, the model is used to validate new data and reject those data samples, which are not model-conform.

The most general statistical data description can be given in the form of multivariate probability density functions (pdf). A priori known multivariate pdf can be used to classify data samples as correct (probable) or incorrect (improbable). However, estimation of the pdf for raw record-level data is very difficult due to the *curse of dimensionality* and weak relations amongst the raw data samples. (See [4] for a recently proposed algorithm.) The aforementioned problems can be avoided using the following “sub-optimal” methods.

- The first approach is to decrease the number of variables in the estimated pdf. The dimensionality reduction makes pdf estimation task more feasible but is offset by an accuracy loss, since variable correlation cannot be fully exploited. Simple pdf-related statistical tests for univariate pdf may involve its estimation, analysis of peaks in the estimated distributions (see Figure 4), or outlier detection. Advanced tests may utilize logistic regression-type tests [13], hidden Markov models [5], clustering techniques, or modeling of multivariate histograms as slowly changing time functions.
- The second approach is to use aggregated (summed) data. Such data usually exhibit stronger redundancy than the original data, even if the individual customer-related variations are suppressed. Two types of redundancy exist: spatial (between variables, segments, partitions, etc.) and temporal (in time). Redundancy is extracted by statistical models which can be identified directly from the data [7] and used for monitoring purposes. See [6], [8], and [3] for application examples.

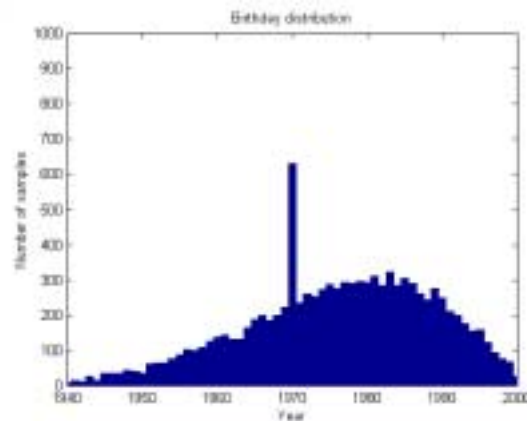


Figure 4: Univariate histogram of birthdays reveals clear outlier (default value)

In this paper only the second approach is exploited, the one common in the process monitoring methods. The following table summarizes redundancy types and the corresponding models:

<b>Redundancy type</b>	<b>Redundancy meaning</b>	<b>Appropriate models</b>
Temporal	relates values of one variable for different time instants	time series or lagged-variable models like AR, ARIMA
Spatial	relates values of several variables for one time instant	multivariate static models like linear regression, PCA, nonlinear models like NNPCA, hypersurface
Spatial and temporal	relates values of several variables for different time instants	multivariate time series models like VAR, VARMA, transfer function models like ARX, ARMAX, BJ

Table 2: Redundancy types and the corresponding models

## 2.6. Data Aggregation

Analysis of the aggregated data (like total amounts of assets, customers counts, and service sales) may give a valuable insight into the contents and data quality of a financial database. These time series have a generally understood meaning and can be compared to the usually available reference controlling data. Additionally, most of them can be treated as global indicators, useful for decision makers (e.g., to perform market monitoring and prediction). Moreover, it is often fair to suppose that such time series are slow changing and the relations between particular variables are almost constant over time (consider average assets per customer group, which also have clear intuitive meaning).

As previously noted, data aggregation suppresses individual customer variations. Hence, minor faults related to single customers may go unnoticed. The aggregated data are influenced by seasonality and business conditions as well as data quality issues. The aggregation dimensions should, if possible, coincide with the already mentioned fault-related dimensions.

The aggregated data form compact multivariate time series and can be stored for reference for long periods. Examples of aggregated data are: (1) number of accounts per customer segment, product type, partition, and time unit, and (2) sum of transactions per customer segment, product, partition, and time unit. Sensible aggregation operations include sum, count, mean, variance, minimum, maximum, histograms; such aggregates can be further aggregated in other dimensions without need for re-computations [4].

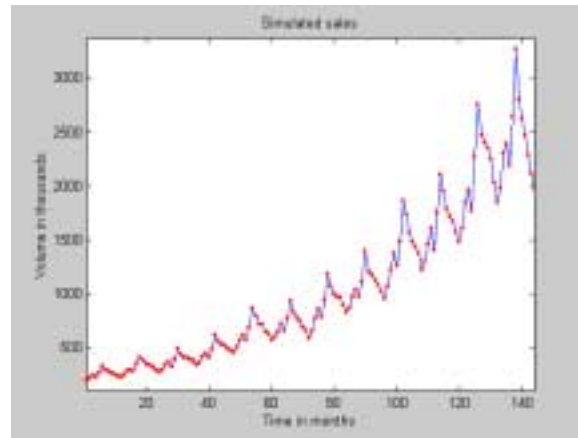


Figure 5: Example aggregate (simulated sales) forms a time series and exhibits visible temporal redundancy. The example follows [2].

### 3. Selected Process Control-Related Tests

This section demonstrates the application of selected process monitoring methods with respect to financial databases. Generally, such methods may include the following stages [9]-[10]:

- *Fault detection*, i.e., obtaining evidence that a group of variables/data samples is influenced by a fault. The detection principle is to test if data satisfy the model equation.
- *Fault diagnosis*, i.e., determination of which variable(s) is/are influenced by the fault. The isolation is performed via model-based variable elimination and by testing the consistent data subsets (via application of the so called structured residuals, which are described later).
- *Fault compensation*, i.e., reconstruction of the proper value of given variable. The reconstruction is possible using the model and fault-free variables and requires a high degree of redundancy.

There exist two basic process monitoring approaches [6]. In the first approach (called *parity space* method) a constant residual generator is estimated from fault-free data. Then, the generator is used to test new data. An alarm is raised if the residuals exceed given thresholds. In the second approach there is no fixed model. Instead, a parametric model is constantly estimated from the data. The monitoring is performed by testing variations of the model parameters. Block diagrams of both approaches are shown in Figure 6.

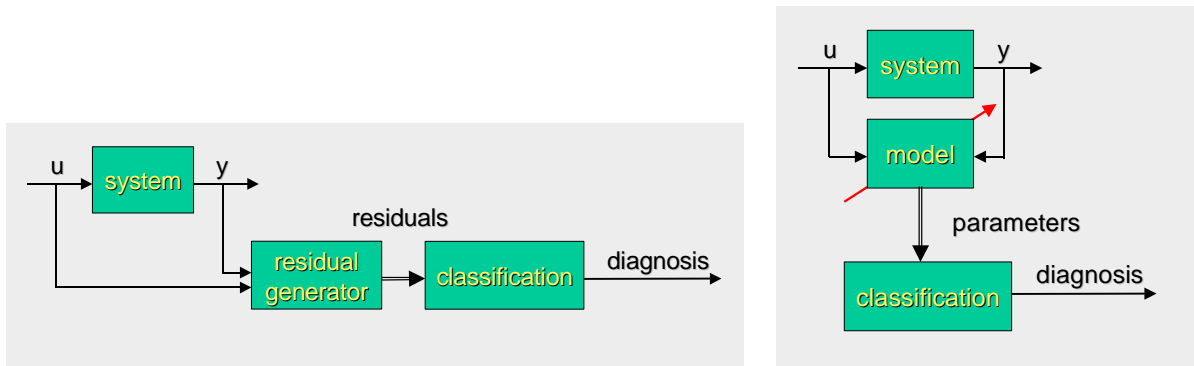


Figure 6: Model-based monitoring: by testing residuals (left) and model parameters (right)

Table 3 contains examples of model-based tests, which belong to general model classes from Table 2, and are described in detail in the forthcoming sections.

Method	Redundancy	Model	Objects	Tested values
Differencing in time	temporal	$x(t)=x(t-1)+\varepsilon(t)$	record numbers, aggregated assets	residuals
RLS	spatial	$x_t^i = \hat{\alpha}_t^i \bar{x}_t$	record numbers, aggregated assets	parameters
Ellipsoidal bounding	spatial	$x(t)^T F^{-1} x(t) < \alpha$	aggregated transactions	residuals
PCA	spatial	$\Theta^T x(t) = 0$	aggregated transactions	residuals

Table 3: Example model-based tests

### 3.1. Monitoring Almost Constant Variables via Differencing in Time

This method is appropriate for exploiting temporal redundancy and testing if the aggregated variables change slowly, e.g., in time/partition or time/customer segment coordinates. Examples of such variables are record counts, asset sums, number of customers, accounts, etc.. The underlying model is  $x(t)=x(t-1)+\varepsilon(t)$ , where  $\varepsilon(t)$  is small compared to  $x(t)$ .

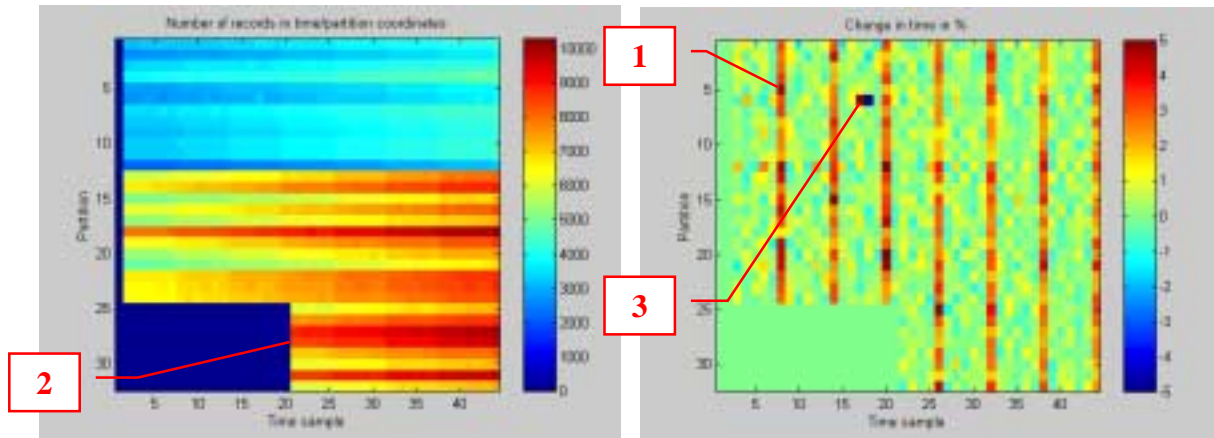


Figure 7: Example of data load monitoring via differencing number of records: left – original number of records, right – relative difference in %

Simulated results shown in Figure 7 depict the original, color-coded record count (left picture) and its relative increase in percent (right picture), generated via the following residual generator:  $e(t) = 100*(x(t)-x(t-1))/x(t)$ .

The monitoring principle is to raise the alarm if bounds on  $e(t)$  are violated. (Note that monitoring complex seasonal variations like the one shown in Figure 5 requires more advanced time series models like ARIMA [3], but the monitoring principle is still the same.) The residuals  $e(t)$  are color-coded and graphically presented in partition/time coordinates. The following effects are visible in Figure 7: (1) increase in number of records every 6 time units, (2) the appearance of new partitions #25-32, and (3) small single outlier, not visible in the record counts.

The relative increase in the number of customers/assets can be also graphically presented for the most recent month in customer segment/product coordinates. Such an analysis (called *BusinessBarometer* in DQontrol [1]) is probably even more interesting for database users than for data quality specialists, since it may deliver direct business-related benefits.

### 3.2. Monitoring Slowly-Varying Relations via Recursive Least Squares (RLS) Method

This method can test if, for a given time instant, the record count for a given partition and time sample is proportional to the total number of records in the table. Periodically, e.g., quarterly loaded tables for which the previously discussed differencing method is not appropriate, have the aforementioned property; an example is shown in Figure 8.

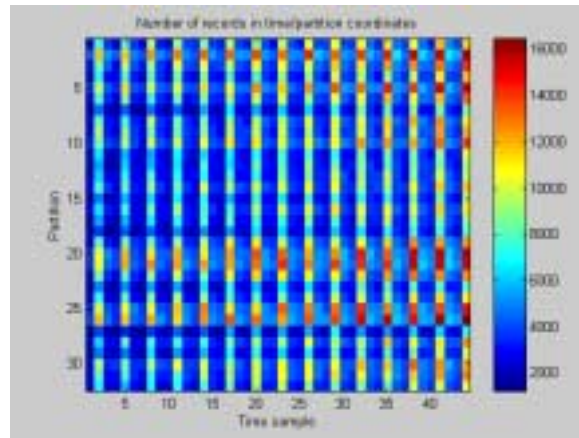


Figure 8: Number of records in an example table with periodic loads

The number of records in each partition  $x_t^i$  is modeled as a function  $x_t^i = \hat{\alpha}_t^i \bar{x}_t$  of mean for all partitions, denoted  $\bar{x}_t$ . The coefficient  $\hat{\alpha}_t^i$  is estimated using RLS algorithm [7]. Note that the model utilizes spatial redundancy, i.e., proportions between the record counts in all partitions. The algorithm is defined by two equations: the parameter update and covariance update

$$\hat{\theta}(t) = \hat{\theta}(t-1) + \frac{P(t-1)\varphi(t)}{1 + \varphi^T(t)P(t-1)\varphi(t)}(y(t) - \varphi^T(t)\hat{\theta}(t)),$$

$$P(t) = \left( P(t-1) - \frac{P(t-1)\varphi(t)\varphi^T(t)P(t-1)}{1 + \varphi^T(t)P(t-1)\varphi(t)} \right) \frac{1}{\lambda},$$

where  $\theta(t)$  denotes the parameter vector ( $\hat{\theta}(t) \equiv \hat{\alpha}_t^i$ ),  $P(t)$  covariance matrix,  $\varphi(t)$  regression vector ( $\varphi(t) \equiv \bar{x}_t$ ),  $y(t)$  modeled variable ( $y(t) \equiv x_t^i$ ), and  $\lambda$  forgetting factor.

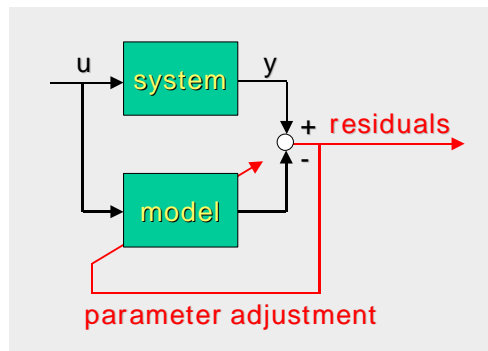


Figure 9: Principle of adaptive modeling used by RLS algorithm

The results, delivered by the RLS algorithm for the fault-free data from Figure 8 are shown in Figure 10. The plots include the modeled number of records in the partition #1, RLS model output, residuals  $y(t) - \varphi^T(t)\hat{\theta}(t)$  and model parameter  $\hat{\alpha}_t^i$ .

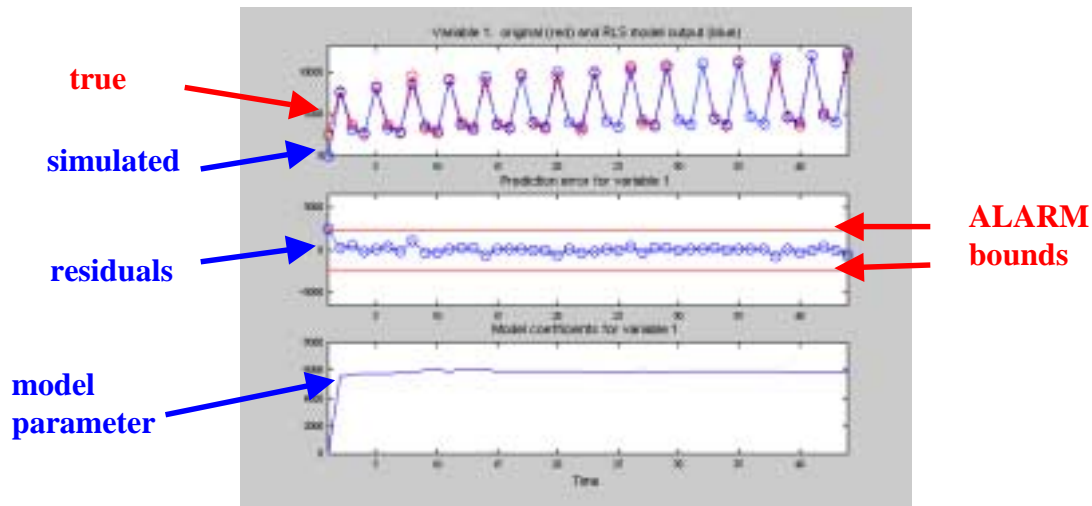


Figure 10: RLS monitoring of the record counts in the first partition

The model parameter remains almost constant, and the residuals stay within the bounds.

### 3.3. Monitoring Correlation via Ellipsoidal Bounding

This method is appropriate for the monitoring of selected aggregated variables for specified product and customer segment with deterministic relations between the aggregated variables. The data related to different partitions and time instants are assumed to stay within hyperellipsoidal bounds. The hyperellipsoid origin is at  $\bar{x}$ , and it is defined by the quadratic form  $(x - \bar{x})^T F^{-1} (x - \bar{x}) = \alpha$ , where

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i, \quad F = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(x_i - \bar{x})'$$

The data can be validated using the following condition  $(x - \bar{x})^T F^{-1} (x - \bar{x}) < \alpha$  and visualized together with the bounding hyperellipsoids for varying values of the parameter  $\alpha$ .

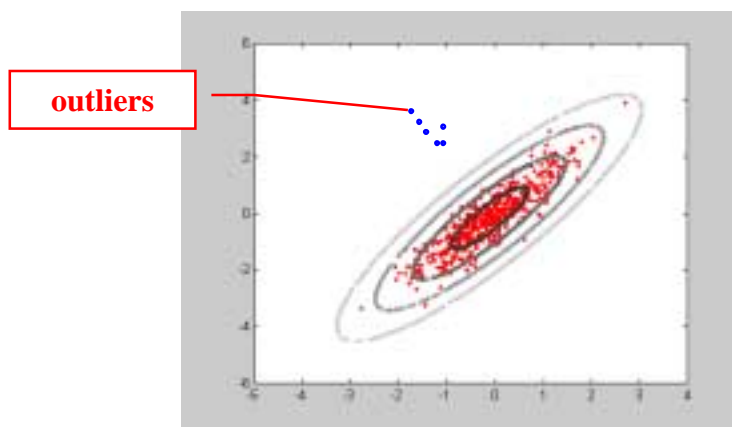


Figure 11: Example dependence between the centered number of accounts and sum of assets



### 3.4. Monitoring of Aggregated Variables via Principal Component Analysis (PCA) Algorithm

This method can be used for monitoring an arbitrary number of variables, e.g., aggregated for specified product and customer segments. The aggregated data related to different partitions and time instants are assumed to be approximately located in some hyperspace, i.e., to exhibit a high degree of redundancy.

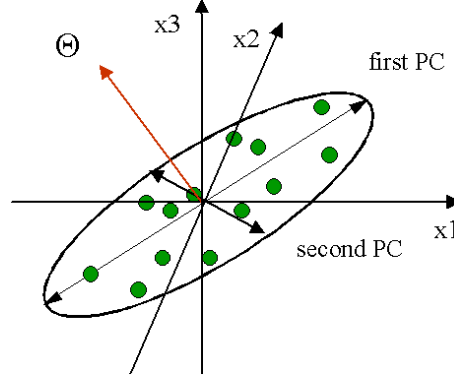


Figure 12: Principle of PCA modeling: data define a hyperellipsoid, the least hyperellipsoid axes  $\Theta$  are almost orthogonal to the data

#### The Principal Component Analysis (PCA) Model

The aggregated, centered, and normalized data are stored in the matrix  $X \in \mathfrak{R}^{K \times N}$ , such that its columns correspond to  $N$  variables and rows  $x(k)^T \in \mathfrak{R}^N$  to  $K$  time samples:  $X := [x(0) \ x(1) \ \dots \ x(K-1)]^T$ . The Singular Value Decomposition  $X = U\Sigma V^T$ , where  $U \in \mathfrak{R}^{K \times K}$  is an orthogonal matrix,  $\Sigma \in \mathfrak{R}^{K \times N}$  is a diagonal matrix  $\Sigma = [diag(\sigma_i) | 0_{(K-N) \times N}]$ ,  $i = 1..K$  with  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_K \geq 0$ , and  $V \in \mathfrak{R}^{N \times N}$  is an orthogonal matrix, enables determination of the model. First, the elbow at the plot of singular values (Figure 13) has to be found to fix the model order  $L$ . The coefficient matrix  $\Theta$  is given as the last  $N-L$  column vectors  $v$  of  $V$   $\Theta = [v_{L+1} \ v_{L+2} \ \dots \ v_N]$ , while the orthonormal vectors spanning the model hyperspace are the first  $L$  column vectors of  $V$ ,  $P = [v_1 \ v_2 \ \dots \ v_L]$ . Note that the model utilizes spatial redundancy (multivariate proportions between the aggregated variables).

#### PCA Fault Diagnosis Algorithm

Fault diagnosis is performed in the following steps

- Fault detection via evaluation and assessment of the norm of the primary residuals

$$res = \|\Theta^T x\|_2 \quad (1)$$

- Fault isolation via evaluation and assessment of the norm of the structured residuals

$$res_i := \sqrt{x^T \Pi_i [I - P(P^T \Pi_i P)^{-1} P^T] \Pi_i x} \quad (2)$$

where  $\Pi_i$  is a diagonal matrix with ones for retained variables and zeros for eliminated variables. Note that structured residuals are sensitive to faults in the retained variables and completely insensitive to faults in the eliminated variables.

- Fault-free reconstruction of the data

$$\tilde{x}_i = -(\tilde{\Theta}\tilde{\Theta}^T)^{-1} \tilde{\Theta}\hat{\Theta}^T \cdot \hat{x}_{-i}, \quad (3)$$

where  $\tilde{x}_i$  is a vector containing the reconstructed variable(s).  $\hat{x}_{-i}$  is a vector containing all variables except the variable(s)  $i$  to be reconstructed.  $\tilde{\Theta}$  contains only the column(s)  $i$  of  $\Theta$  and  $\hat{\Theta}$  contains the remaining columns.

### Application Example

The presented example utilizes simulated data but follows a real financial application. The processed data contain various customer transactions for a given product and customer segment. The data are aggregated within partitions and time periods. In the test example there are 19 variables in 5 partitions, collected during 30 time samples. An additional 5 variables contain the numbers of summed entries in each partition. (Altogether there are 100 variables.)

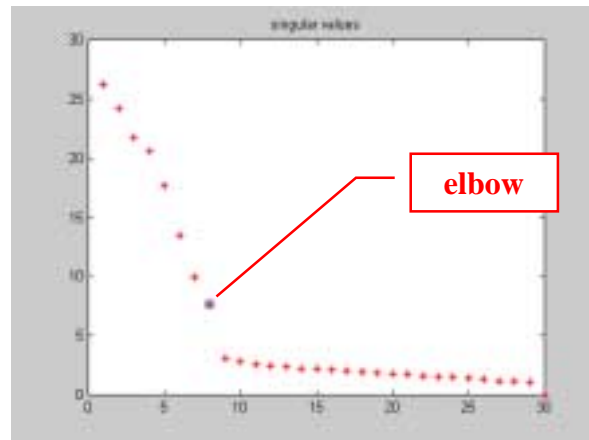


Figure 13: Singular values of the data matrix suggest model with 8 degrees of freedom

Figure 13 depicts singular values of the normalized and centered fault-free reference data matrix. The suggested model order  $L = 30 - 8 = 22$ .

Following possible real situations it is assumed that a simulated fault may corrupt: (i) single variable in one partition, or (ii) single variable in all partitions, or (iii) all variables in one partition. Note that for each case the  $\Pi_i$  matrices in (2) are different. Figure 14 shows the primary residuals  $\tilde{\Theta}^T x$  and the norm of the structured residuals (2) for an example fault corrupting all variables from the partition 2 (#21-39) for the time sample 9.

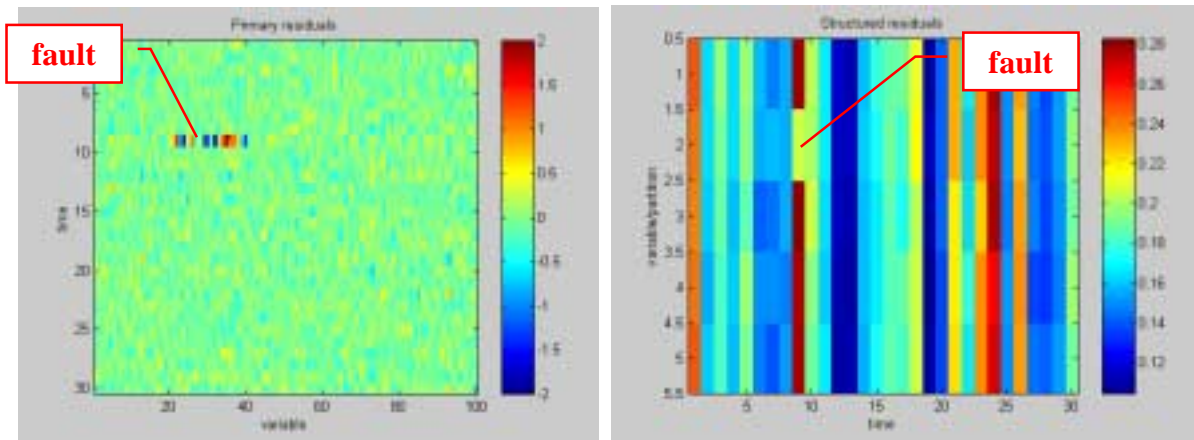


Figure 14: Primary residuals (left) and structured residuals (right)

The fault causes an increase of the norm of the primary residuals  $\Theta^T x$  (left picture), and decrease of the structured residuals (2) related to partition 2 (right picture), and is correctly isolated. Hence, the fault-free reconstruction of the corrupted data via (3) is possible. Figure 15 shows such a reconstruction of the variable #22 from partition 2, which enables an assessment of the fault's magnitude and sign, at least in the terms of the aggregated quantities.

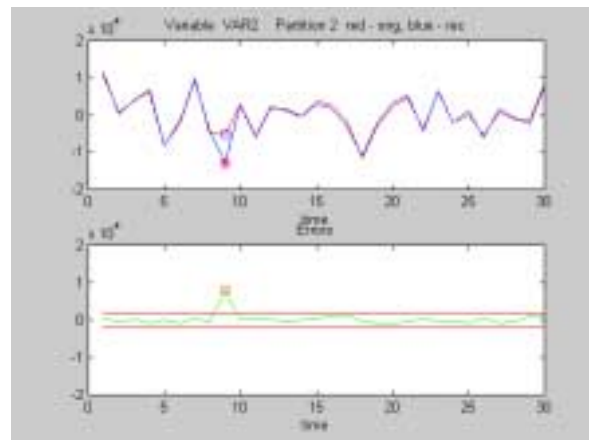


Figure 15: Fault-free reconstruction (3) of the corrupted data (above) and the residuals (below)

#### 4. Fault Visualization, Diagnosis, and Compensation

Automated measurement of data quality produces large amounts of findings within a possibly short time period. In an example database with 2000 variables and 5 tests per variable, a total of 10000 aggregated test results must be evaluated. In order to cope with such amounts of findings, the subsequent processing must be highly automated and exploit appropriate visualization methods.

### 4.1. Drill-Down Visualization and Alarming

Aggregated test results, presented in the form of signatures, can be further compressed to enable quick assessment of the overall database quality. Figure 16 shows an example of a drill-down traffic-light like assessment of variables and tables. Depending on number and distribution of the test failures their result can be assessed as satisfactory (green), warning (yellow), or alarm (red). The results of all quality tests for one variable can be combined into a joint quality assessment. At the next stage quality assessments of all variables in one table are used to build overall quality assessment of the table.

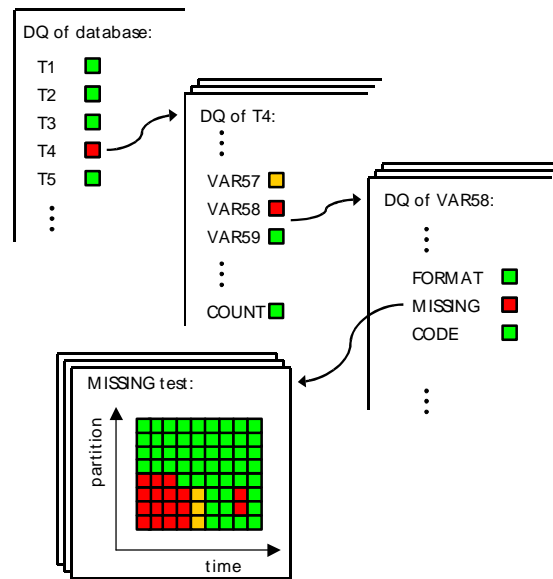


Figure 16: Example of a hierarchically structured data quality assessment

Additional alarms can be associated with a decrease of the quality at any stage, using the previously described differencing of the fault signatures. A good practice is to require that each major alarm is acknowledged by data quality personnel. Note that color-coded signaling, drill-down visualization of the monitored system’s state, and alarm acknowledgements are state-of-the-art tools used in modern process monitoring systems.

### 4.2. Fault Diagnosis via Signature Clustering

Correlation or clustering of test signatures can be exploited to relate a detected data quality symptom to other similar symptoms, and, finally, to their original cause. All symptoms belonging to the same cluster can then be corrected simultaneously, accelerating thus the analysis and correction of the data quality findings. Furthermore, by identifying and correcting the underlying cause of the findings, future data quality problems can be eliminated. Similarity of signatures can be assessed using clustering or correlation techniques. The numerical burden can be kept low, e.g., by reducing number of dimensions by additional aggregation prior to the signature clustering.

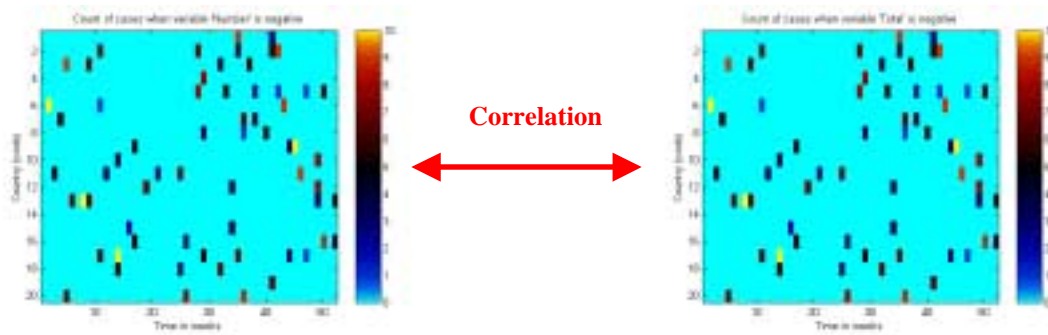


Figure 17: High correlation of the signatures suggests that both symptoms have a common cause

### 4.3. Data Cleansing

Automatic data correction means elimination or, at least, reduction of data quality problems by explicit modification of the original data. The correction can be performed on the variable level (modification of value) or record level (record elimination). Data correction must be implemented with care since the original data are manipulated. In certain situations like existence of an alternative data format or completely redundant variables full data correction is possible (there is no information loss). More often, however, the fault recovery information is not available, so full fault correction is not possible. Often it is better to replace the faulty (misleading) value by the missing value (what prevents numerical use of the incorrect value).

Data cleansing can be useful also when performed for the aggregated data, as in the previous PCA modeling example: comparing real aggregated values to the ones reproduced by the model enable assessment of the data loss and can be very helpful for the fault diagnosis. The discussed PCA fault isolation method can be also useful to select better information source if one of two highly correlated variables is corrupted, see [9] for a detailed algorithm.

## 5. Conclusions

This paper discusses several process control-related methods applied in the context of monitoring and control of data quality in financial databases. It shows that the control process can be considered a classical control feedback process, and that application of the model-based methods is useful to detect, diagnose, and eliminate data quality problems. Moreover, the model-based methods give an insight into business-related information contained in the data. The methods constitute part of DQontrol product [1], and have been applied to the data quality monitoring of a real financial database at a customer site, delivering business benefits, such as improvements of the modeling quality, a reduction in the number of the modeling cycles, and better data understanding. These benefits in turn lead to financial savings and better utilization of highly skilled data analysts.

## References

- [1] F. Block, J. Milek, M. Reigrotzki. (2000). Datenqualität als Basis künftiger Business Intelligence Applikationen. SAS Warehousing 2000, Zürich.
- [2] G.E.P. Box, G.M. Jenkins. (1970). Time Series Analysis: Forecasting and Control. Holden-Day.
- [3] R. Busatto. (2000). Using Time Series to Assess Data Quality in Telecommunications Data Warehouses. Proceedings of the 2000 Conference on Information Quality, Cambridge, MA, pp. 129-136.
- [4] T. Dasu, T. Johnson, E. Koutsofios. (2000). Hunting Data Glitches in Massive Time Series Data. Proceedings of the 2000 Conference on Information Quality, Cambridge, MA, pp. 190-199.
- [5] R. Elliott, L. Aggoun, J. Moore. (1995). Hidden Markov Models. Estimation and Control. Springer Verlag.
- [6] J. Gertler. (1998). Fault Detection and Diagnosis in Engineering Systems, Marcel Dekker.
- [7] L. Ljung. (1998). System Identification: Theory for the User. 2<sup>nd</sup> edition, Prentice Hall.
- [8] S. Makridakis, S. C. Wheelright, R. Hyndman. (1998). Forecasting: Methods and Applications. 3<sup>rd</sup> edition, John Wiley and Sons Inc..
- [9] J. Milek, F. Kraus. (2000). Use of Analytic Redundancy in Fault-Tolerant Sensor Systems. IMEKO 2000 International Measurement Confederation, XVI IMEKO World Congress, Vienna, Austria, Proceedings Volume V, pp. 121-127.
- [10] J. Milek, O. Hermann, F. Kraus. (2000). Use of Hypersurfaces for Fault Detection, Isolation, and Reconstruction. IFAC 4th Symposium on Fault Detection Supervision and Safety for Technical Processes. SAFEPROCESS 2000, Budapest, pp. 1199-1204.
- [11] J. Milek. (2002). Diagnose in der Messtechnik. Chapter in Handbuch der industriellen Messtechnik, in preparation, 7th edition by Pfeifer, Ruhm, and Modigell, Oldenbourg Verlag.
- [12] X. Z. Wang. (1999). Data Mining and Knowledge Discovery for Process Monitoring and Control. Springer Verlag.
- [13] S. M. Weiss, N. Indurkha. (1998). Predictive Data Mining. Morgan Kaufmann Publishers, Inc..

## **THE IMPLEMENTATION OF INFORMATION QUALITY FOR THE AUTOMATED INFORMATION SYSTEMS IN THE TDQM PROCESS: A CASE STUDY IN TEXTILE AND GARMENT COMPANY IN THAILAND**

**Athakorn Kengpol**

Department of Industrial Engineering, Faculty of Engineering  
King Mongkut's Institute of Technology North Bangkok, Thailand  
1518 Piboolsongkram Rd., Bangkok 10800  
Tel./Fax + 66 2 5874842  
Email: athakorn@kmitnb.ac.th

**Abstract:** Nowadays, industries are looking forward to obtain a precise data in both internal database and external database across their organisations. This paper attempts to clarify and recommend a way to share precise data environment. The contribution of this paper is to propose a framework to improve components of Information Quality (IQ) for Automated Information Systems (AIS) in a practical way via a case study firm in Thailand. The core set of information quality and Benefits/Costs/Risks (BCR) analysis of IQ have been presented through a case study which is a medium size textile and garment industry in Thailand. Several designed procedures in AIS have been upgraded to minimise its mismatched data. The major result indicates that, with the enhancement of IQ project, the projected risk cost of losing current customer has been dramatically reduced compared with the projected risk cost of losing potential new customer. The other results, limitations and recommendations are also presented.

**Keywords:** Information Quality, Automated Information Systems, BCR Analysis

### **1. Introduction**

As information is one of the most powerful tools in business nowadays, in particular Automated Information Systems (AIS), which formally did not communicate, are increasingly required to share the information environment. Focused on achieving the sharing environment are often categorised as playing a pivotal role in facilitating and/or inhibiting system integration.

The contribution of this paper is to propose a way to improve AIS and information quality in practice through a case study. As known that textile and garment is a foundation industry in Thailand and it is even stronger during the devaluation of the currency due to lower wages and more choices of company. A number of information regarding to designed requirements and specifications from abroad are quickly coming simultaneously with orders. The textile and garment industry is in needs to implement Database Management System (DBMS) to cope with a high demand and requirements for further development. The main hurdle in implementing of DBMS in this industry is how to achieve the precise information between transaction of DBMS and users. That is the reason that this paper aim to apply the AIS to improve Information Quality (IQ).

The company in this case is a medium scale textile & garment company who previously used manual but currently using AIS in management information systems. As the cost of AIS have become reachable if the economy of scale is achievable, the management of this firm targets to improve AIS whilst the level of Information Quality (IQ) must be high for every user. It is

because different data users impose different quality requirements and that was of acceptable quality for one system might not be so in another. In addition, data that was sufficient accuracy and timeliness for local users may not be acceptable at another site, particularly other continents. Cost of inaccurate or inadequate data can become sky high. Problems with information quality can result in tangible and intangible damage varying from loss of customers/users confidence to loss of orders. This company found that there are several current customers turn down the order because they found many mismatched data between their DBMS and company's DBMS.

In the following sections, it clearly explains series of step beginning from background of TDQM such as requirements and evaluation of IQ. Then management process of IQ which has 4 major steps in detail, for example, establishing the TDQM environment, identification of IQ projects, implementation of IQ projects and benefits/costs/risks (BCR) Analysis. After that the next section will explore a developed model within a case study textile and garment firm which will explain how to improve AIS and TDQM in a practical way.

## 2. Requirements and Evaluating of IQ

It has been known that Information Quality (IQ) is not a binary attribute [12]. We should not declare that a set of information is (or is not) of high quality but we should evaluate information in the context of each specific intended to use [13]. Then we should also feed the result of the evaluation back to improve the IQ. This implies that multiple evaluations must be applied and the results must be recorded for using in the future.

Table 1: Core Set of IQ Requirements (Adapted from DoD Guidelines [2])

Data Quality	Characteristics Description	Conformance Measures
Accuracy	A quality of that which is free of error. A qualitative assessment of freedom from error, with a high assessment corresponding to a small error [3].	Percent of values that are correct when compared to the actual value. For example, M=Male when the subject is Male.
Completeness	Completeness is the degree to which values are presented in the attributes that require them. [1].	Percent of data fields having values entered into them.
Consistency	Consistency is a measure of the degree to which a set of data satisfies a set of constraints. [11].	Percent of matching values across tables/files/records.
Timeliness	As a synonym for currency, timeliness represents the degree to which specified data values are up to date [11].	Percent of data available within a specified threshold time frame (e.g., days, hours, minutes).
Uniqueness	The state of being the only one of its kind. Being without an equal or equivalent.	Percent of records having a unique primary key.



--	--	--

Remark: DoD represents Department of Defense of USA.

Requirements and Evaluating of IQ have two distinct aspects one involving the “correctness” objective such as in Table 1 at IQ Characteristics Description column. The another one concerns the “appropriateness” of data for some specific purpose [9,11,13]. The data users usually assure that the purpose of IQ assurance (for the Total Quality Management or TQM purpose) is to provide the best data possible [7,8].

If the data users obtain such IQ assurance data which they believe it is the best data possible [7,8], it means this obscures the need to evaluate data. In other words, if the information is the best available or as good as it can be produced, then there is no other alternatives but to use it, in this case there is no point to worry about how good it is. The flaw is that saying that the information is as good as it can be produced does not inform us *how* good it is. Therefore, we need an explicit evaluation as in Table 1 at Conformance Measures column. It describes how to perform an evaluation which those activities are done to ensure that data are correct and appropriate for their specific purpose. After we define requirements and evaluation of IQ which is the background of this paper, the next section describe four total data quality management process.

### 3. Management Process of Information Quality

The management process of IQ applies Total Data Quality Management (TDQM) approach to support database migration and improve database in conformance to business rule. The TDQM actually borrow the Total Quality Management (TQM) methodologies to apply human resources and quantitative method to improve products and/or services. The TDQM approach integrates functional management techniques and former improvement efforts to create or sustain of the continuous improvement process.

As illustrated in Figure 1, the TDQM consists of four major processes which has been described within the Defense Information Systems Agency (DISA) [2] and simultaneously enhanced by the author. Firstly, Establishing the TDQM environment by designing of management and infrastructure support. Secondly, Identification of information quality projects. Thirdly is the Implementation of IQ projects. Finally, are the Benefits, Costs and Risks Analysis of IQ projects.

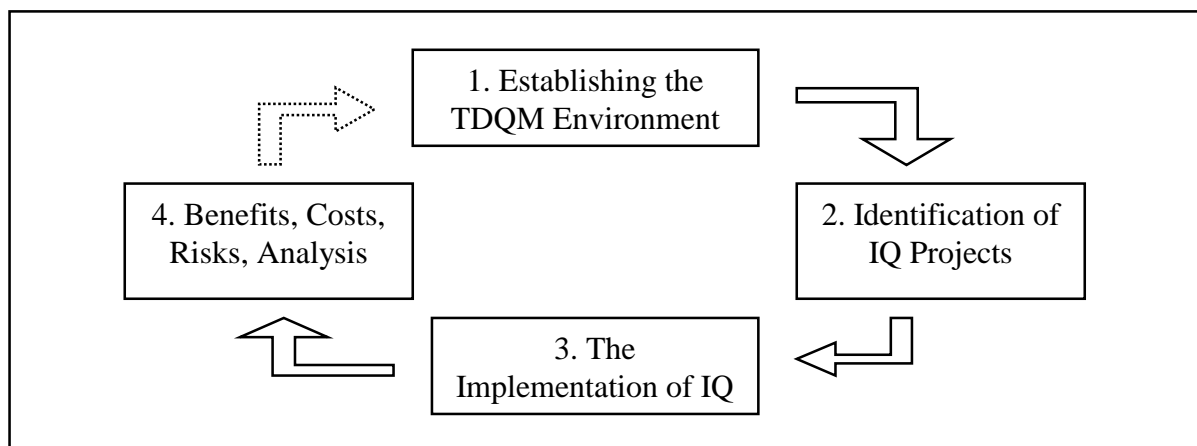


Figure 1: Total Data Quality Management Process

Adapted from the Defense Information Systems Agency (DISA) [2] and simultaneously enhanced by the author

### **3.1 Establishing the TDQM Environment**

This is the first step in TDQM process and probably one of the most difficult step in the process, as similar to the introduction of Concurrent Engineering into the industry [4]. The existing culture which is characterised by an attitude, for example users know the IQ problems but do not want it to be fixed or AIS staff knows how to identify the IQ problems but do not want to say that they do not know how to change functional requirements that drives the data. It is because they worry about job cut. Given these existing barriers, one of the most important solutions is the involvement of the top management to support all obstacles as their pace of commitment is the only way to gradually change the existing culture in organisation. The action plan should be developed as following:

- Overall goals and objectives to the achievement of IQ.
- Strategies to obtain goals and objectives.
- Infrastructure needed such as responsibilities of members in the team, training programs etc.

### **3.2 Identification of IQ Projects**

How the TDQM project philosophy is promoted within the organisation will affect its success. If the TDQM is exaggerated, then under-delivery will result in a loss of creditability and hence slow down the introduction process. On the other hand, if the TDQM is under-publicised, then senior management may lose sight of the TDQM and therefore, it will not be fully exploited or supported its capabilities. Generally, either users or AIS administrators or both should select IQ projects. It is a critical to listen and understand in mutual. For example, users usually report complains with record errors. These inaccuracies happening in queries, reports and data correlation problems are good indicators of IQ issues. Alternatively, system administrators may suggest recommendations based upon problems with data collection, processing errors and procedures. Therefore, The recommendations for the selection of IQ projects are as follow:

- Choose the project that has a highly opportunity for project success. It is very important to make the project success as the first time that it is firstly implemented so that the organisation can adopt the idea more easily in the future. The steps are: (1) select the project that has a high chance of success, (2) that has the highest impact lost cost to the firm, (3) where the significant improvement can be made. Certainly, project that can be solved with minimum effort but results are obvious will increase the attractiveness of TDQM project to the top management.
- Doing a prototype project. If there is no support from the top management, performing a pilot project is another alternative. Choose the project with low risk but high visibility which is critical to the organisation success. Focus also on a project that has a huge chance of success. Importantly, choose the effort that is neither too large that is doomed for failure from the beginning, nor too small those improvements can be negligible.

### 3.3 The Implementation of IQ Projects

The implementation of IQ projects can be separated into two distinct steps: Developing the IQ Implementation Plan and Implementing the IQ Projects.

#### 3.3.1 Developing the IQ Implementation Plan

As recommended by DoD [2], the project management will be applied in developing the IQ implementation plan in which it provides information on:

- Task Summary: list project goals, objectives, scope and synopsis of anticipated benefits.
- Task Description: describe data quality tasks.
- Project Approach: summarise tasks and tools to be used as a baseline in developing BCR Analysis.
- Report Analysis: List reports that conform to the BCR Analysis.

#### 3.3.2 Implementing the IQ Projects

In implementing the IQ projects, based upon the TDQM process, it is defined into four activities:

1. Define: as in Table 1, to identify the IQ requirements.
2. Measure: to measure in conformance with the requirements as in Table 1.
3. Analyse: to verify, validate and assess the causes for poor IQ and seek for the improvement.
4. Improve: to improve the IQ, it may have to change data entry procedures, enhancement of data validation rules or using a uniform standard using throughout the organisation.

### 3.4 Benefits, Costs and Risks Analysis of IQ Projects

One of the most crucial tasks in IQ improvement is the identification of benefits, costs and risks which are connected to direct root causes of IQ problems and the indirect root causes that damage information.

Table 2: Major Types of Benefits, Costs and Risks of IQ

Benefits	Costs		Risks
	Direct IQ Costs	Indirect IQ Costs	
1. Customer Loyalty	1. Controllable Costs - Prevention Costs - Appraisal Costs - Correction Costs	1. Customer Incurred Costs	1. Correction Costs Too High
2. New Customer	2. Resultant Costs - Internal Error Costs - External Error Costs	2. Customer Dissatisfaction Costs	2. Schedule Too Long
3. Reputation Improvement	3. Equipment and Training Costs	3. Creditability Lost Costs	3. Costs of Buy-In Programmer for Correction

Source: DoD Guidelines [2] and Simultaneously Enhanced by the Author

The major types of Benefits, Costs and Risks of IQ are illustrated in Table 2. At Costs generally, Direct IQ Costs which consist of controllable costs, resultant costs and equipment and

training costs are quantitative by estimating based upon labour hour devoted to prevention, appraisal, correction activities, poor data quality cause internal&external error which need modifications, also equipment and training costs. In Indirect Costs are normally qualitative but it could be estimated whenever possible to adequately assess the impacts of poor IQ [2]. For example, inability to match payroll records with the initial employment records can become overpayment to employees.

In terms of Benefits and Risks, they are more difficult to quantify than costs. In particular, customer loyalty, reputation improvement (for Benefits) and too long schedule (for risks) are even harder quantifiable. If several alternative projects are available, it is advisable to apply a holistic decision making approach called “Analytic Network Process” or ANP which is a discrete multi-criteria decision making (MCDM) appropriate for complex model [5] or “Analytic Hierarchy Process” (AHP) [6] for simple case study. Alternative decision making models can be seen in [10]. The more detail and application of the approach will be discussed in the case study.

### **3.4 Summary**

As recommended from the DoD guidance on data quality management [2], the goal is to ensure that:

- (1) Users are involved in the IQ improvement process.
- (2) Measurable data characterised are predetermined.
- (3) The information acquired are conformed to the requirement in (2)

According to the Figure 1, the approach to achieve the goal comprises of four steps. Firstly, the TDQM environment must be established where from the top management to the users is responsible to seek for an action plan. Secondly, is the identification of IQ projects which should provide the high chance of success. Then the implementation of IQ projects which consist of two steps: developing of the IQ implementation plan and implementation of the IQ projects. Finally, the BCR Analysis of IQ projects. The case study in the next section will discuss deep into details of the application of the TDQM concept into a real IQ project.

## **4. Information Quality Case Study**

### **4.1 Introduction to the Problem at a Textile and Garment Company in Thailand**

The firm is a SME textile and garment company who produces several textile and garment products for domestic and export market. Currently, they are using Local Area Network (LAN) system to connect five departments together e.g. Sales and Marketing, Finance and Accounting, Operations and Manufacturing, Human Resource, and Health and Safety. This firm has had a fully integrated system for a year, however, mismatched information from its Database Management System (DBMS) to each department is frequently occurred. This problem has been addressed and the management of the firm has fully supported to implement an improvement of their IQ in AIS. The example of problems are, there are frequently mismatched information of orders and specific requirements between (Sales and Marketing department) and (Operations and Manufacturing department) which almost cause lost of order and penalty charged. Another example is that there are some mismatched information between (Human Resource department) and (Finance and Accounting department) in regarding some unauthorised information can be accidentally shown at Finance and Accounting department in which they asked just only to know enrolment date etc. This company is in needs of improving in AIS by TDQM philosophy.

## **4.2 The AIS Improvement by IQ Project Using TDQM Philosophy**

The TDQM philosophy does not appear on the scene unexpectedly, most is result of many years of development process. The patterns of appearance and arrival of TDQM have common themes. The senior management in this company should have initiated it but in fact, TDQM philosophy is firstly introduced to the company by a group of middle management. Such appearance makes some difficulties in making awareness to the organisation culture. According to the seniority culture in Asian countries, senior management needs to be considerably convinced and explained about the benefit of TDQM philosophy and IQ. Fortunately, senior management agrees that TDQM and IQ would able to enhance operations within the firm and since then fully commitment have been drawn. As illustrated in Figure 1 and based upon an endorsement from the management, the first step Establishing the TDQM Environment can be achieved.

## **4.3 The Identification of IQ Project**

As this project is the first in its kind within this firm, the management decided that they should have done it in a small scale, therefore, there is only one project in which the goal is to minimise or eliminate the mismatched information between users and DBMS on the AIS. This is a critical hurdle to the company in regarding the creditability to its customer and reliability of management in-house.

## **4.4 The Implementation of IQ Project**

To implement the IQ improvement of this project, there are two steps: developing the IQ implementation plan and implementing the IQ improvement project.

### **4.4.1 Developing the IQ Implementation Plan**

The project goal is to minimise the mismatched data between users and DBMS in AIS. The objective of this project is to increase the accuracy and the consistency of data using on the AIS which will enhance the reputation of the firm ultimately. The scope of this project is in operation between DBMS and 5 departments namely Sales and Marketing, Finance and Accounting, Operations and Manufacturing, Human Resource, and Health and Safety department. The approach of the project will be done in both way: enhancement of the DBMS program and minimise the human error in process. The anticipated cost and benefits will be discussed in section 4.5.

### **4.4.2 Implementation of the IQ Project**

After the core set of data quality have been identified, the IQ team, who consists of a project manager, programmers and department representatives, have come to develop an implementation plan and seek for hurdles, then try to overcome them. The requirements of the core set of data quality and developed procedures are illustrated in Table 3 which is an enhancement from Table 1. The difference between them are in Table 3, they add Validity analysis instead of Timeliness and Operations because the IQ team considers validity has much more important role than timeliness and uniqueness. The Developed Procedures or “Filter” has been developed based upon Accuracy, Completeness, Consistency and Validity Analysis.

The result from the analysis, causes and solution of mismatched data that conform to the analysis are illustrated in Table 4. At table 3; from the operations point of view, Filter does indeed

contribute to the success of the improvement of IQ project. Several designed programs and procedures, as far as concerned, eliminate the mismatched data of AIS. The Benefits/Costs/Risks Analysis of IQ is discussed next.

Table 3: Core Set of Data Quality and Developed Procedures for the Firm (Adapted from Table 1, [2] and simultaneously enhanced by the author)

<b>Data Quality</b>	<b>Developed Procedures (Filter)</b>	<b>Conformance Measures</b>
Accuracy Analysis	Computational verification data between sources and end users.	Percent of values that are correct when compared to the actual value. For example, there are frequently happen that, loading some data from a terminal, it shows mismatched data. In particular, obtained incorrect data bundle with the correct data is the most concerned.
Completeness Analysis	Computational verification between data.	Percent of data fields having values entered into them. For example, downloading data from its DB to column for analysis but some fields are missing or null.
Consistency Analysis	Computational verification flow between point to point comparison.	Percent of matching values across tables/files/records. This is the most concerned due to the credit of its business can become grim if the inconsistency occur frequent when they are communicating with its customer.
Validity Analysis	Valid integration of values within data set.	Percent of data available within a specified threshold time frame (e.g., days, hours, minutes, seconds). In this case, it means the original data are considered to be valid or not. If yes, for how many percent.

Table 4: Results, Causes and Solutions of the Project

<b>Result Occurred</b>	<b>Causes</b>	<b>Solutions for AIS</b>
------------------------	---------------	--------------------------

Mismatched Data	Input into the wrong field	Design program and procedure to element; design an authorisation level for the users.
	Input incorrect data	Design program and procedure to eliminate such as selected for string, number or others.
	No data but thought as it has had	Design a Data Map program and procedure to explain a perspective view of data in Database.

#### 4.5 The Benefits, Costs and Risk (BCR) Analysis of IQ Project

The automated data quality, as recommended in Table 3 and 4, validates several principles of Benefits and Costs. However, evaluating Risks is one of the critical parts in the improvement process. Particularly, risks of losing current customer and potential new customer move to other more reliable business partners.

Table 5 illustrates projected savings by BCR analysis using current costs (both direct and indirect costs, as recommended in Table 2, are included) as a baseline. For example at Controllable Cost, Database (DB) Programmers Operations is the cost of employing programmer staffs to look after the DB. The DB Users Operations means cost of employing user staff to key inputs and draw output data. Both costs are major expenses on IT to the firm and after establishing of IQ project, the new improved cost become 4.5 MB or 30% savings (1.98MB) in total based upon current costs at 6.48 MB in replacing staffs by using new enhanced programs as in Table 3 at Development Procedures (Filter). In terms of Risks in particular, there are two possibilities estimated, lost of current customer and lost of potential new customer. The meaning of lost of current customer and potential customer have been clarified in Table 5. The comparison between both of them indicated that lost of current customer has more impact to the firm (before 20 MB compared with after 7 MB, therefore, save lost risks 13 MB) than lost of potential new customer (before 8 MB compared with 3 MB, therefore, save lost risks 5 MB).

Table 5: Projected Savings by BCR Analysis

Description of Benefits/Costs/Risks Estimated	DB Programmers Operations	DB Users Operations	Total Operations
<b>Current Costs</b>	3.6 MB	2.88 MB	6.48 MB
<b>New Improved Costs</b>	2.4 MB	2.10 MB	4.50 MB
<b>Benefits</b>	1.2 MB or 33 % Savings	0.78 MB or 27 % Savings	1.98 MB or 30 % Savings

Remarks: Savings are based upon replacing staffs by using new enhanced programs as in Table 3 at Development Procedures (Filter).

<b>Risks</b>	<b>Before</b>	<b>After</b>	<b>Total Preventive Lost Risks</b>
- Lost of Current Customer	20 MB	7 MB	13 MB
- Lost of Potential New Customer	8 MB	3 MB	5 MB

Remark: Lost of Current Customer is calculated based upon in-depth interviews current customer about reason for cancellation of order which can be converted to amount in currency.

Lost of Potential Customer is calculated based upon data from Marketing Department in contacting with potential new customer.

MB means Million Thai Baht.

## 5. Conclusions and Recommendations

The central contribution of this paper is in applying IQ concept to improve the accuracy, completeness, consistency and validity of AIS in order to minimise risks of current customer lost and potential customer lost. The concept of IQ has been described and applied through a case study of a Textile and Garment Company in Thailand.

The associated contribution is to present a Benefits, Costs and Risks analysis to identify trade-off between current costs and new improved costs through IQ concept. The BCR analysis helps to reveal a general theme that lost of current customer risks play more important role to the firm than lost of potential new customer. It is because major stake of incomes are from current customer, therefore, major cash flow of this firm are based upon the relationship between them. Running a business in Thailand, reputation usually has an equal important with market prices, therefore current customer deserve the first priority and also supported by data in Table 5. Moreover, if they lose orders from their current customer, it is not only financially but also reputation lose which they have been built up for years.

The data come from a textile&garment company, therefore, although the results are only appropriate for the textile&garment business, they provide a useful test of the methodology for improving IQ. Engineering firms or firms in other industries may implement this concept so that they can customise their data analysis in seeking for a new product strategy.

The Total Data Quality Management Process should be continuously reassessed in which either positive or negative feedback results should be recorded for further enhancements. The current and potential new customer will obtain the benefit and the firm will be better off ultimately.

The author hopes that findings from the research can benefit any firm in developed and developing countries to share their *precise* data. IQ are increasingly important across global industries, simultaneously with economic turmoil in countries in Asia. Therefore, a system of comprehensive improvement in new ways of IQ is desperately needed.

## 6. References

- [1] *Data Quality Foundation* (Select Code No. 500-149), AT&T, 1992.
- [2] DoD Guidelines in Data Quality Management (Summary).
- [3] FIPS PUB 11-3, *American National Dictionary for Information Systems*, 1991, February.
- [4] Kengpol, A., "Transferring Concurrent Engineering into the Developing Country", *Proceedings of The International Conference on Production Research (ICPR 2000)*, 2000, August, Asian Institute of Technology, Thailand.



- [5] Kengpol, A. and O'Brien, C., "An Analytic Network Process for the Evaluation of Investment in Time Compression Technology", *Proceedings of The Eleventh Annual Meeting of the Production and Operations Management Society (POMS-2000)*, 2000, April, San Antonio, Texas, USA.
- [6] Kengpol, A. and O'Brien, C., "The Development of a Decision Support Tool for the Selection of Advanced Technology to Achieve Rapid Product Development", *International Journal of Production Economics*, 2001, Vol.69, pp. 177-191.
- [7] Kon, H.B., Lee, J. and Wang, R.Y., "A Process View of Data Quality", TDQM Research Program, MIT, 1993, March.
- [8] Lee, J. and Wang, R., "On Validation Approaches in Data Production", TDQM Research Program, MIT, 1993, August.
- [9] Liepins, G.E. and Uppuluri, V.R.R. (eds.), *Data Quality Control: Theory and Pragmatics*, Marcel Dekker, Inc., New York, 1990.
- [10] Pandey, P. C. and Kengpol, A. "Selection of An Automated Inspection System Using Multiattribute Decision Analysis", *International Journal of Production Economics*, 1995, Vol.39, pp. 289-298.
- [11] Redman, T.C. *Data Quality Management and Technology*, Bantam Books, 1992 (ISBN 0-553-09149-2).
- [12] Rothenberg, J "A Discussion of Data Quality for Verification, Validation and Certification (VV&C) of Data to be Used in Modeling, Rand Draft DRR-1025-DMSO (forthcoming).
- [13] Rothenberg, J and Kameny, I. "Data Verification, Validation and Certification to Improve the Quality of Data Used in Modeling" *Proceedings of the 1994 Summer Computer Simulation Conference (SCSC'94)*, 1994, July, La Jolla, CA, pp. 639-644, Society for Computer Simulation (SCS), (ISBN 1-56555-029-3).

# A Methodological Approach to Data Quality Management Supported by Data Mining

Udo Grimmer

DaimlerChrysler AG  
Research & Technology, FT3/AD  
Ulm, Germany  
udo.grimmer@daimlerchrysler.com

Holger Hinrichs

Oldenburg Research and Development  
Institute for Computer Science Tools  
and Systems (OFFIS)  
Oldenburg, Germany  
holger.hinrichs@offis.de

**Abstract:** In this paper, we use the example of the car manufacturing domain to illustrate how data quality problems are addressed in practice today. We then propose a process model for data quality management (DQM) which meets the requirements of the current ISO 9001 standard and thus introduces a methodological, process-oriented approach to DQM. Data mining methods that are typically applied to find interesting and previously unknown patterns in large amounts of data are being used to support several phases of this process model. The main idea behind the application of data mining methods is to deem data anomalies deviations from a ‘normal’ quality state. The primary advantage of our approach is an increased degree of automation and enhanced thoroughness and flexibility of DQM.

## 1. Data Quality Challenges in Automotive Manufacturing

While product quality has always been a central focus at DaimlerChrysler, data quality has not yet received the attention it deserves. This does not mean that data quality has been completely neglected thus far, but most data quality initiatives have had solely a strong local focus. With the growing demand for the integration of distributed, heterogeneous databases into corporate warehouse applications, data deficiencies have become obvious, necessitating corporate-wide DQM to address data quality issues across system boundaries. This global data quality view implicates additional data quality perspectives and presents challenges regarding the related tools and methodologies.

As correcting data already stored in a database system is much more expensive than setting up appropriate measures to prevent substandard-quality data to be entered into the systems, precautions should be given top priority. However, as real environments are complex, there will be always a need for measuring, monitoring, and improving the quality of data after it has initially been stored. This was one of the motivations for initiating a research project which focuses on the application of data mining technologies in the context of DQM. We have introduced the term *Data Quality Mining*, which we believe to have great potentials for both future research work and a successful transfer of data mining technology into daily work processes.

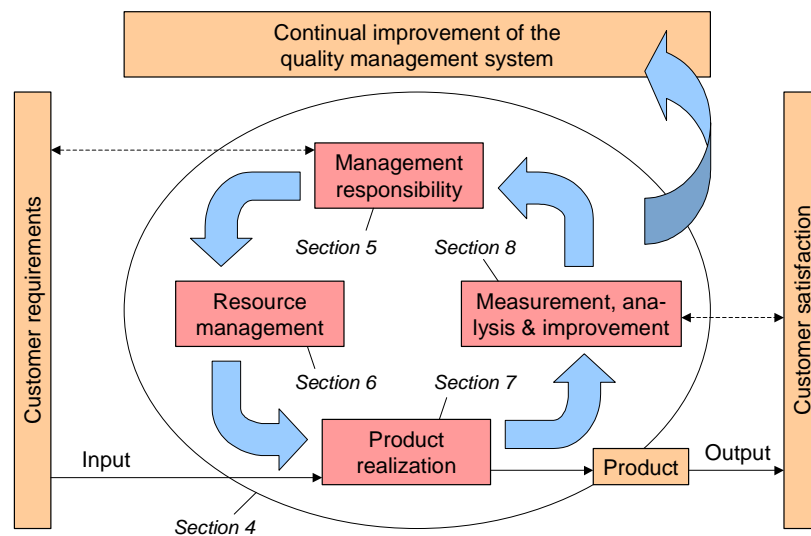
In section 2, we propose a quality management system for data integration that meets the requirements of the current ISO 9001 standard. Section 3 presents a case study where a subset of these concepts has been applied to the QUIS (QQuality Information System) database of the Global Services and Parts division of the Group using data mining techniques. Finally, we give an overview of related work and further research issues in section 4.

## 2. ISO 9001 Compliant Data Quality Management

In [16], the term ‘quality’ is defined as the ‘degree to which a set of inherent characteristics fulfills requirements’. Quality characteristics form the backbone of *quality management (QM)*, defined as ‘coordinated activities to direct and control an organization with regard to quality’. The system within which QM is performed is called *quality management system (QMS)*.

### 2.1. The ISO 9001 Standard

The ISO 9000 family of standards was developed by the International Organization for Standardization (ISO) to assist organizations in implementing and operating effective quality management systems. The current ISO 9000:2000 revision was published in December 2000. In this section, we give an introduction to the ISO 9001 standard, i.e. that part of the series that specifies requirements for a QMS.



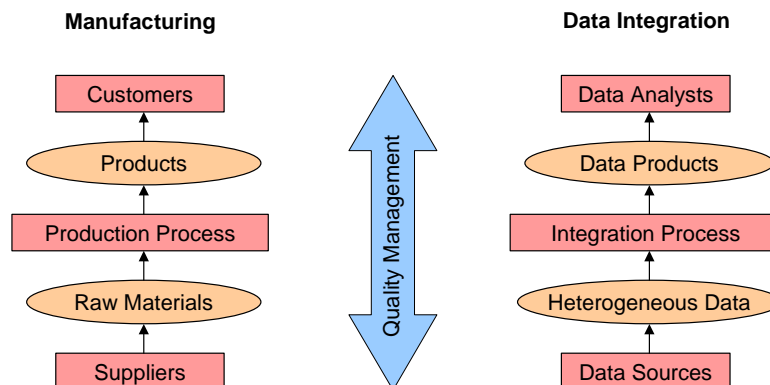
**Figure 1: Model of a Process-based QMS [17]**

ISO 9001 promotes the adoption of a process approach when developing, implementing, and improving a QMS to enhance customer satisfaction by meeting customer requirements [17]. Within this process, an organization has to identify various activities, then link them and assign resources to them, thus building up a system of communicating processes. Figure 1 depicts such a process-based QMS. Customers play a key role in this model since their requirements are used as input for the product creation process and customer satisfaction is continually subjected to analysis.

ISO 9001 is made up of eight sections. The first three contain general information about the scope of the standard, normative references, and terms. Sections 4 to 8 describe requirements for a QMS and its components, as indicated in Figure 1.

## 2.2. Data Quality Management

In the following section, we sketch a QMS for the process of data integration from heterogeneous sources as an exemplary data processing activity that is especially important in data warehouse applications like customer relationship management or supply chain management. As Figure 2 illustrates, the data integration process can be viewed as a kind of production process. Following this analogy, we can adapt the well-established QM concepts known from the manufacturing/service domain to the context of data integration, hereafter called *data quality management (DQM)*.



**Figure 2: Analogy Between Manufacturing and Data Integration**

An essential aspect of DQM is data quality measurement. As DeMarco so rightly states: ‘You cannot control what you cannot measure’ [3]. We need metrics to be able to calculate the degree of conformance with given requirements. In the manufacturing domain, we have to measure characteristics like lengths, weights, speeds, etc. For databases, on the other hand, we need to measure characteristics like consistency, timeliness, and completeness of data products (see also section 3.2). Yet, metrics for data quality characteristics are still a matter of research [1], [13], [22], [26].

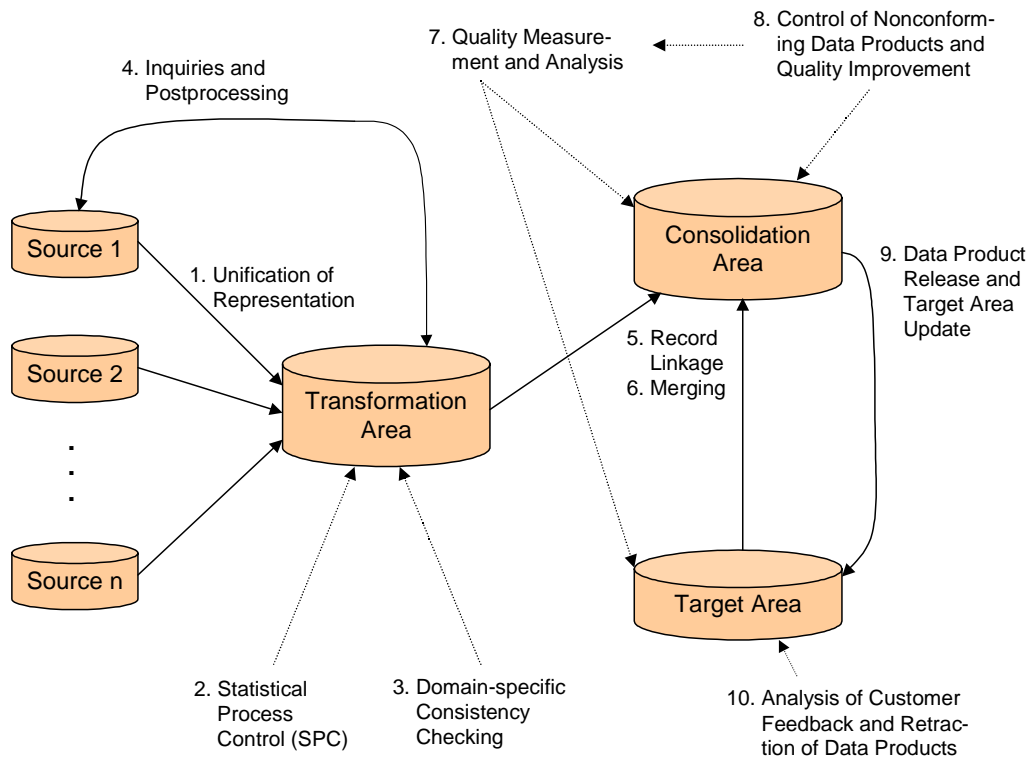
## 2.3. A Data Quality Management System for Data Integration

In this section, we present a process model for data integration which defines the exact integration steps to be executed. Based on ISO 9001, this process model is enriched with DQM steps to ensure that customer requirements are fulfilled. Integration steps plus DQM steps along with organizational DQM activities form a QMS for data integration called *data quality management system (DQMS)*.

The DQMS should be viewed as an integral part of an organization. It is closely coupled to the organization’s management, its human and technical resources, and – of course – its (data) suppliers and (data) customers. Customers specify quality-related requirements and provide feedback concerning their satisfaction with the data products supplied.

In this paper, due to the space constraints, we concentrate on ISO 9001 sections 7 (product realization) and 8 (measurement, analysis, and improvement). For details on the remaining ISO 9001 sections that concern organizational aspects such as documentation, resource management, etc. and their influence on DQM, see [9]. Furthermore, we will only describe the operational

and their influence on DQM, see [9]. Furthermore, we will only describe the operational phases of data integration organized as a 10-step process model, which forms the core of our DQMS (see Figure 3).



**Figure 3: Operational Steps of Data Integration**

*Step 1: Unification of Representation*

In this step, source data are moved into a temporary storage called the *transformation area*. The transformation area is assumed to have a global schema that is covered (in terms of content) by the source schemata. The main task of this step is to map the heterogeneous source data structures to the global data structures of the transformation area. The following mapping tasks are especially important:

- Generating unique keys which refer to source system identifiers.
- Unifying character strings syntactically (with regard to umlauts, white spaces, etc.).
- Unifying character strings semantically (in case of synonyms).
- Decomposing complex attribute values into atomic ones (e. g. addresses).
- Aggregating atomic attribute values into complex ones (e. g. date values).
- Converting codings (e. g. gender values m/f to 1/2).
- Converting scales (e. g. inches to cm).

### *Step 2: Statistical Process Control*

After unification, *statistical process control (SPC)* is performed on the transformation area data as per classical SPC theory [27]. The idea is to compute attribute-specific statistical figures (mean, variance, etc.) and log them over time. The newly computed figures are then compared to previously logged key figures. This allows data deficiencies (e. g. transfer errors) to be detected at a very early stage and appropriate actions such as initiating a new data transfer to be taken.

### *Step 3: Domain-specific Consistency Checking*

In this step, the transformation area records are checked with regard to consistency using domain-specific knowledge. The latter should be represented in such a way that it can be processed automatically. Different types of representation are possible:

- *Rules*, e. g. ‘IF RepairDate < ManufacturingDate THEN Error (Severity 0.9)’.
- *Lookup tables*, e. g. engine types.
- *Regular expressions*, e. g. for special equipment codes.
- Arbitrary domain-specific functions.

In addition, business rules could also be discovered at runtime by means of data mining methods (see section 3) and then be applied to transformation area data. If an inconsistency is detected, an appropriate action has to be executed, e. g. by generating an error message or warning.

### *Step 4: Inquiries and Postprocessing*

If appropriate domain knowledge is available (or data mining methods are being applied), the major proportion of inconsistencies can be *detected* automatically. However, very few inconsistencies can be *corrected* automatically. Consequently, if an inconsistency is not tolerable, an inquiry has to be sent to the data source affected (generated automatically from an error message, if possible). Corrected records then have to be integrated into the transformation area appropriately.

### *Step 5: Record Linkage*

The goal of this step is to detect duplicate records, i. e. records that describe the same real-world entity, both within the transformation area and between the transformation area and the so-called *target area* where the consolidated data are to be stored in the end.

Because of the heterogeneity and potential internal redundancy of data sources, records have to be linked by means of non-key attributes like name, city, gender, delivery date, for example. Several methods suitable for automating this task have been proposed in literature. The most prominent ones are (i) Probabilistic Record Linkage [18], (ii) Duplicate Elimination Sorted-Neighborhood Method [14], and (iii) Neighborhood Hash Joins [8]. All these methods result in a set of record pairs which potentially belong together. These pairs, more precisely their transitive closures, now have to be analyzed with respect to whether or not the linkage is correct. In marginal cases, an interactive review is inevitable.

### *Step 6: Merging*

Records that describe the same real-world object must now be merged to a single record in order to avoid unintentional redundancy. By applying certain criteria (information content, attribute-specific priorities on data sources, timeliness, etc.), the best pieces of information have to be extracted from the records in question and written into a target record. In our process model, this target record is *not* written to the target area as one could expect, but into another temporary storage, the so-called *consolidation area*, instead. The target area is not updated until the very last step of the process model, thus ensuring that only data which have passed all the conformance tests (some of which will follow in the subsequent steps) are written to the target area and thus made available for analysis tasks.

The records that merged in a consolidation area record are then deleted from the transformation area. The remaining transformation area records are moved to the consolidation area without any modification. The consolidation area now serves as the starting point for the following DQM activities.

### *Step 7: Quality Measurement and Analysis*

In this step, a check must be done to ensure that the data in the target area meet the specified customer requirements (in compliance with ISO section 8) even after they have been updated with the current consolidation area data. To do this, the actual quality of data must be measured, using appropriate metrics and measuring software according to ISO section 7. These measurements<sup>1</sup> must span both the (present) target area data and the consolidation area data. (Note that the target area has *not* been updated yet!)

In a subsequent analysis phase, the results of the quality measurements have to be compared to a priori-specified quality requirements. If data do not meet a given requirement, appropriate action has to be taken (see step 8). Conflicts resulting from contradicting requirements (e. g. high timeliness vs. high consistency) have to be resolved, e. g. by data replication and different treatment of the replicas.

Measurement results concerning the effectiveness of processes have to be recorded and analyzed (ISO section 8), leading to process improving activities if necessary (see below).

### *Step 8: Control of Nonconforming Data Products and Quality Improvement*

In this last step before the target area update, data products which do not conform to the given requirements must be treated appropriately in accordance with ISO section 8. The following options may be taken:

- Sort out and re-request data.
- Restrict the use of data to specific scenarios, e. g. by flagging.
- Eliminate detected nonconformities and then continue with step 7.

---

<sup>1</sup> Including consistency checks as in step 3 (and, if required, postprocessing as in step 4), since a merging of records may introduce new combinations of attribute values and thus new inconsistencies.

While all these activities tackle only the symptoms of a problem, further (cause-oriented) measures may be taken to increase the system's ability to fulfill quality requirements in the future according to ISO section 8:

- Improve the integration process, especially by tuning process parameters such as attribute mappings, SPC parameters, consistency rules, record linkage parameters, merging criteria, for example.
- Improve quality planning and quality control processes, e. g. by finding better means to capture user requirements, by optimizing measurement methods, or by improving feedback methods.

#### *Step 9: Data Product Release and Target Area Update*

Depending on the analysis results of step 7, the approved proportion of data is now released, i. e. the consolidation area records affected are flagged as 'passed'. The passed proportion of data is then moved from the consolidation area to the target area, replacing obsolete target area data if necessary. With this step, the newly integrated data are made available for customer use.

#### *Step 10: Analysis of Customer Feedback and Retraction of Data Products*

The organization concerned has to record customer feedback and evaluate it as a measure of the DQMS performance (ISO section 8). If a deficiency of a released data product is detected during current use (i. e. by a customer), and this deficiency significantly impairs the usability of the data product, the organization has to retract the product from the target area and 'repair' it if possible (see step 8) before re-releasing it. If necessary, cause-oriented measures should be taken into account (see step 8).

### 3. Case Study: DQM for QUIS

In the following, we describe our experiences with the implementation of selected steps from the DQMS described in section 2.3 to the QUIS (Quality Information System) database that is running on a 20-processor HP-UX machine with 16 GB main memory and 720 GB disk space. QUIS is a central database in the Global Services and Parts division of DaimlerChrysler. It contains technical and commercial data of passenger cars and trucks from the warranty and goodwill periods. It is used for different tasks such as product quality monitoring, early error detection and analysis, or reporting. An application which is targeted at deviation detection of warranty and goodwill costs is presented in detail in [12]. Current data quality problems like inconsistent or incomplete data are generally related to the complex system environment, which consists of various operational source systems with partially non-conforming data models and sophisticated data transfer processes. In this case study, data mining methods have been used for data quality assessment.

#### **3.1. Data Quality Mining**

According to Fayyad et al., the typical task of data mining is the investigation of large amounts of data to discover '*valid, novel, potentially useful, and ultimately understandable patterns*' [7].



Although the discovery itself can be automated, the subsequent interpretation of these patterns (rules, decision trees, etc.) always requires human interaction. Let's take a look at the following rule, which was found by a rule learning algorithm during the analyses of the QUIS *axles* table:

```
"IF Model = 210 AND Plant = 050 AND PDate <= 1999/09/22
  THEN PID = B (86514 cases, 100.00%)"
```

This rule states that in all known (86514) cases, the value for the attribute *PID* is 'B', if all three conditions from the IF part hold. Hence, we could apply this rule for consistency checks for any new data, for example. If we find a record where the conditions from the IF part hold, but the value of the attribute *PID* does not equal 'B', we need to check whether this is a new, valid finding or whether one or more of the values for *Model*, *Plant*, or *PDate* are incorrect.

As this example leads us to conclude, there are only marginal differences between the application of a data mining algorithm for discovering new patterns, and the application of the same algorithm for discovering data anomalies. What makes the difference is merely the way patterns are applied and results interpreted. The term *data quality mining* is used to indicate this differentiation. In [10] we define data quality mining as the deliberate application of data mining techniques for the purpose of data quality measurement and improvement. The goal of data quality mining is to detect, quantify, explain and correct data quality deficiencies in very large databases.

Naturally, there are some data mining algorithms that are more appropriate for data quality mining tasks than others. In particular, algorithms from the fields of deviation detection and dependency analysis bear the largest potentials for data quality mining.

### 3.2. Selection of Relevant Data Quality Aspects

In the initial phase of the QUIS data quality project, we conducted a series of workshops with QUIS data customers (business end users, knowledge engineers, management, and database administrators) to identify the key data quality characteristics. Starting from a list of potential data quality aspects as found in [29], we applied the Quality Function Deployment Matrix approach in the context of DQM as published in [25]. The goal was to map the subjective data quality requirements to objective, quantifiable criteria. As a prerequisite for the application of data mining methods, we had to focus on aspects that could directly be derived from the QUIS data. In accordance with the above constraints, the following five data quality aspects have been identified:

- *Completeness* (with respect to both records and attributes): for records, the percentage of data objects stored in QUIS compared to the number of real entities; for attributes, the percentage of missing and/or blank values.
- *Timeliness*: the period between the date any data have been entered in the operational source systems and the date they become available in QUIS.
- *Consistency*: the number of anomalous records compared to the whole number of records. Consistency rules do not need to be defined manually, as they are discovered from historic data by data mining methods.
- *Accuracy*: the degree of conformance of the data objects in QUIS with the real-world objects.

- *Validity*: the percentage of valid (=known) attribute values for different data fields compared to the number of values stored in certain QUIS reference tables.

Regarding appropriate data mining techniques, we have chosen the following for an initial application and evaluation (the data quality aspects addressed are listed below):

- *Descriptive statistics*: completeness (w. r. t. attributes), validity.
- *Statistical outlier detection*: completeness (w. r. t. records), timeliness, accuracy.
- *Decision rules* and, initially, *association rules*: consistency.

### 3.3. Data Quality Mining to Support Operational Steps of Data Integration

Following the methodology presented above, we partially implemented some of the steps for QUIS. Initially, we had to investigate the historic data stored in the QUIS tables (this would have been unnecessary if the DQMS proposed in section 2.3 had been applied from the very beginning). For this investigation we developed a prototype that will be described in the section entitled Descriptive Statistics for Relational Database Content Summarization.

Now we discuss the relationship of the individual operational steps 1 through 10 from the data integration process model to the QUIS application, including sample applications of data mining methods.

*Step 1: Unification of Representation* was met by operational procedures, i.e. records from the source systems are uniquely identified by key attributes such as the vehicle or part identification number.

*Step 2: Statistical Process Control* was of special importance, since data collecting and transferring processes are susceptible to interference. The corresponding actions are set out in the section on Statistical Approach to Deviation Detection.

*Step 3: Domain-specific Consistency Checking* and *Step 4: Inquiries and Postprocessing* were already operationalized for a number of business rules. Here, the problem was that new or as yet unknown inconsistencies are not covered by the existing rules because changes in the operational systems are sometimes not communicated to the QUIS administrators quickly enough. The application of a rule generating method to identify potential data anomalies (including new, but still unknown patterns) is described in Decision Rules to Discover Potential Data Anomalies.

*Step 5: Record Linkage* and *Step 6: Merging* are currently irrelevant for QUIS since the procedures in place are sufficient.

*Step 7: Quality Measurement and Analysis*, as described in section 2.3, can only be applied to the target area (QUIS) as currently there is no physical consolidation area installed. Finding general (in the sense of 'can be agreed by all data customers') measures to automatically decide on whether the data quality of the system after the update is still satisfactory or not remains a great challenge for the operationalization of this step.

Different approaches are taken for *Step 8: Control of Nonconforming Data Products*, depending on the error type and the source system. For some of the source systems, incorrect data is rejected and re-loaded after the errors have been corrected in the source system. Alternatively, erroneous data is corrected and loaded immediately, with a note to that effect being passed to the source system's administrator. However, not all errors are caused by substandard-quality source

system data. Also, data might be corrupted or lost during transfer processes. In these cases, root causes have to be analyzed, and the particular process has to be fixed and re-executed. As such procedures were already in place, they were therefore not included within the scope of the current project.

*Step 9: Data Product Release and Target Area Update* are operationalized for QUIS and not directly affected by our research activities.

Finally, *Step 10: Analysis of Customer Feedback and Retraction of Data Products*, is viewed as a mid-term activity. We are, of course, interested in monitoring how our results influence QUIS data quality. We have already started collecting certain data quality measures regarding completeness that will allow for comparisons over the next few months. This part, however, still needs further refinement and completion.

Following this overview, we now discuss some sample applications of data mining methods which directly concern steps 2 to 4 of the process model described in section 2.3.

### *Descriptive Statistics for Relational Database Content Summarization*

One of the key tasks for any data analysis, and for data mining projects in particular, is the so-called explanatory phase. According to the CRISP-DM process model (Cross Industry Standard Process for Data Mining) [2], this phase consists of two subphases: data understanding and data preparation. Experiences from many sources have shown that up to 80 percent of all efforts expended in data mining projects is related to these phases, and that the quality of the results of the analyses carried out depends more on thorough preparation during these phases than on the optimization of any parameter taken from data mining methods in the subsequent modeling phase. Main efforts during the explanatory phase are related to the collection and selection of the right – i.e. the most relevant – data and to statistical, and visual analysis in order to achieve a clear understanding of data contents and representations and data transformations in preparation for the subsequent modeling phase.

**For this reason, we took great care to obtain a sound, univariate description of the QUIS data in a first step before applying further, more sophisticated methods. We developed a prototype that automatically generates detailed documentation of the entire QUIS database at attribute level either in Postscript, PDF, or HTML format,. Compared to the functionality of state-of-the-art data mining tool suites, we believe this prototype to have enhanced scalability and the capability to provide a more sophisticated output as it contains a description of each field, together with content and index references.**

Figure 4 shows the output of the prototype for the attribute *AXLE\_TYPE* from the QUIS table AXLE. Additional textual descriptions extracted from modeling tools, or simple text files can be included for each attribute, too. This is a key prerequisite if database contents need to be discussed with business customers, who normally do not know the mapping between their specific business terms and database attribute names.

2.1.3 AXLE.AXLE\_TYPE

Data type: VARCHAR(1)  
 NOT NULL  
 Number of different values: 3 (0.146987%)  
 Average occurrences of a value: 680.333  
 Standard deviation: 1017.36  
 Values:  
 - 1850 'S' (90.6418%)  
 - 190 'D' (9.30916%)  
 - 1 '0' (0.0489956%)

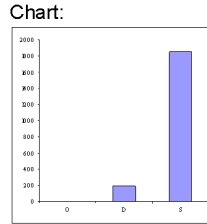


Figure 4: Database Content Documentation

General scenarios the prototype can beneficially contribute to include (i) data and system understanding (e. g. source system analysis), (ii) system design (e.g. entity relationship modeling), or (iii) process optimization.

Statistical Approach to Deviation Detection

Within this part of the project, we developed a statistical prototype for outlier detection. Outlier detection is a burning issue for most analytical applications, and a great deal of work has already been devoted to it in the past. For large real-world applications, it is impossible to provide comprehensive background knowledge specifying what exactly is to be considered an outlier. Data mining methods can be useful here, as they are able to process huge amounts of data autonomously and derive the corresponding knowledge. [4] describes such an approach for the analysis of a very large amount of time series data.

We use the same approach – i.e. deriving models or normative parameters in the simplest case – from historic data. The data used for learning must be guaranteed to be of sufficient quality regarding the parameters to be derived, e. g. mean and variance. Once models have been derived and stored in the model tables for later reuse, corresponding values for new data are computed each time new data are input. Then, a test on deviation is applied. If statistically significant deviations are discovered, suspicious records are flagged and warning messages generated. Figure 5 provides an overview of the steps involved in the model generation and model application processes.

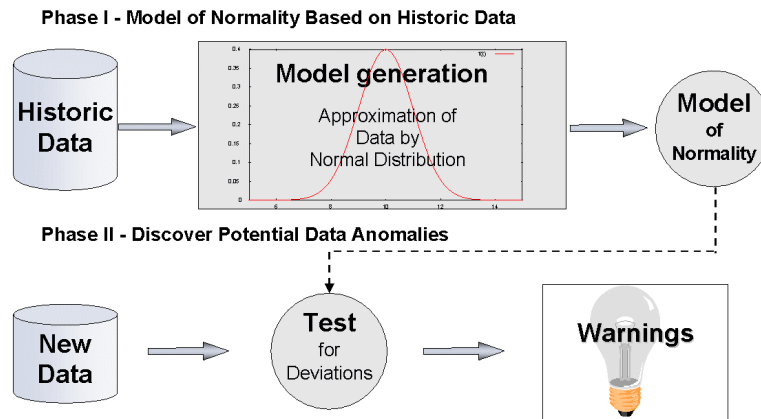


Figure 5: DataQualityMiner-Prototype for Deviation Detection

Example findings from this prototype are, for example, irregular (that is below or above configurable action limits) numbers of warranty claims on a daily or weekly basis for single repair shops as shown in Figure 6.

```

Claims from garage 199, Period: 01.01.2001-30.06.2001

Weekends (Saturday + Sunday)
13./14.01.2001  claims:    73
                expected:  31.23 test statistic: 0.0052496

Weekdays (Monday through Friday)
16.-20.04.2001  claims:    692
                expected: 1097.02 test statistic: 0.0045372
    
```

### Figure 6: DataQualityMiner-Prototype: Sample Results

The prototype offers great flexibility as it covers different queries that can be specified by plain SQL statements. Different parameters such as the model type or an aggregation level regarding a second dimension such as time can be specified as well. Models are updated either as per a fixed schedule or ad-hoc on request. The application of this prototype contributes to (i) automated, early error detection and (ii) permanent process monitoring.

#### Decision Rules to Discover Potential Data Anomalies

To discover consistency problems in the database, we applied the commercial tool GritBot [23] to the individual QUIS database tables as well as to several joined tables (views). Unfortunately, there is no technical description available on the methods GritBot uses. What we have, however, been able to guess from the results and the works published by Quinlan is that GritBot performs several rule generating runs, each time considering another attribute from a subset of  $n$  (default  $n=5$ ) attributes as target (dependent) attribute. Next, each record that violates a certain rule is assigned a significance value indicating the probability that the anomalous value might occur by chance rather than by error. Numeric attributes are discretized automatically, and groupings of values for nominal attributes are generated where appropriate. There are only few parameters that the user can change, but this seems to be somewhat of an advantage, as we were able to achieve good results using only the default settings.

```

GritBot [Release 1.02] Mon Jul 3 13:30:52 2000  Options: Application `AXLE1'
Read 5219164 cases (14 attributes) from AXLE1.data
1995 possible anomalies identified

case 1324925: [0.000]
  AXLE_MDAT = 1919/02/28 (55911 cases, mean 1996/03/31, 100.00% >= 1995/09/01)
  AXLE_TYPE in {311, 919, 305, 306, 061, 018, 460, 313, 470, 054, ..., 075} [480]
  AXLE_PDAT > 1995/09/11 and <= 1997/10/29 [1995/11/15]
case 3773780: [0.000]
  AXLE_EPOS = 2 (118966 cases, 100.00% `1')
  AXLE_TYPE = 510
case 2746691: [0.000]
  AXLE_MODEL = 739 (32459 cases, 100.00% `730')
  AXLE_TYPE = 018
  AXLE_EPOS = 1
  AXLE_ASSID = D
....
    
```

### Figure 7 : Sample GritBot Results

An example from the analysis of the QUIS table *axles* is depicted in Figure 7: for case 1324925, for example, the value for the field *AXLE\_MDAT* was identified to be anomalous, whereas for case 2746691 the value of the field *AXLE\_MODEL* seems to be incorrect. However, for the latter case there is no evidence that it is the value of the attribute *AXLE\_MODEL*, which is wrong. This might also be a correct value, and one or more of the values for *AXLE\_TYPE*, *AXLE\_EPOS*, or *AXLE\_ASSID* might be incorrect.

GritBot proved to be highly scalable in terms of the number of records per table, and the results generated were greatly promising. Without requiring specification of any business knowledge in advance, business rules known by QUIS data customers were proved, and, to top it off, new potential data anomalies were discovered. For our purposes, GritBot's major drawback is the missing database interface: we were continually forced to manually extract two flat files – one file containing the data and a corresponding names file describing the types and possible values for each attribute. As GritBot generated long output listings (some 270,000 lines for all single tables), additional postprocessing steps were implemented to aggregate analytical results.

The application of association rule methods as proposed in [21] seems to be another promising approach. There are, however, two main challenges related to the application of association rules: (i) the special data representation required by most association rule algorithms, and (ii) processing the large number of rules generated. We have already addressed the former issue and are now able to apply association rule algorithms directly on a relational database [11]. Further work is needed to take up the latter challenge. We are about to implement a system which automatically computes quality scores for each record depending on the percentage of its conformance with the rule set generated by the algorithms [10]. This, we hope, will make human interpretation of the rule set to a large extent dispensable.

Both approaches are generally applicable for (i) unsupervised consistency checks and (ii) detection of random (non-frequent) errors.

## 4. Related Work and Future Challenges

In 1992, DeLone and McLean [5] set up a model of information system success that included the quality of data as a key success factor. In the following years, two major research projects emerged, viz. MIT's *Total Data Quality Management (TDQM)* program [28], active since 1992, and the ESPRIT project *Foundations of Data Warehouse Quality (DWQ)* [19], which ran from 1996 to 1999. Apart from TDQM and DWQ, several minor research activities have been launched during the last two to three years, reflecting the rising awareness of the importance of DQM. Among these are CARAVEL [8], IntelliClean [20], HiQiQ [22], and Potter's Wheel A-B-C [24]. [21], [4], and [15] use data mining methods to detect errors automatically.

All in all, although there are some projects which deal with data quality aspects, several critical research issues remain. The overall challenge will be to promote corporate-wide data quality awareness and thus establish DQM as a primary success factor in organizations. DQM should make it evident to all the people in an organization that data quality is not just a local issue, but a global challenge of strategic importance. In the long run, we need to strive for certification of data and data processing systems, as the methodological approach proposed in section 2 suggests. To reach this goal, a number of research issues need to be investigated further:

- Commonly accepted metrics are necessary to enable organizations to assess the quality of their data. For data quality characteristics such as consistency, completeness, and absence of redundancy, metrics have already been defined (see [13]). However, ‘soft’ characteristics like relevance and understandability are very difficult to handle, because they are exposed to subjective influences and require extensive domain and context knowledge. Hence, methods that allow partial automation of measurement by appropriately integrating interactive components must be developed to accommodate these characteristics.
- A methodology for introducing DQM into an organization is needed. Costs related to substandard-quality data and appropriate data quality assurance steps need to be analyzed in more detail (see [6] for an initial discussion). If such costs can be identified and proved using real figures, this would greatly facilitate argumentation for data quality project funding.
- A comprehensive, flexible, and standardized metadata management, plus scalable, domain-independent, and automatable software tools for data quality measurement and improvement have to be developed. To ensure efficiency even with very large data volumes, DQM operators should be implemented as close to the database management system as possible. Furthermore, tools should offer user-oriented means to maintain data quality figures and monitor them over time.
- Automated correction of inconsistencies also requires a great deal of further research. Data mining-based approaches seem to be especially promising in this field (cf. section 3). Future research work should include the integration of domain-specific knowledge and the optimization of efficiency, precision, and recall of such methods.

Although we foresee the need for extensive research work, we believe that our combination of an ISO 9001-compliant process model and automated, data mining-based quality measures will yield a promising foundation for corporate-wide, scalable data quality management.

## 5. References

- [1] Ballou, D. P., Tayi, G. K.: Enhancing Data Quality in Data Warehouse Environments, *Communications of the ACM*, **42** (1): 73-78, 1999.
- [2] CRISP-DM Consortium: *CRISP-DM 1.0 – Step-by-Step Data Mining Guide*, <http://www.crisp-dm.org>, 2000.
- [3] DeMarco, T.: *Controlling Software Projects*, Yourdon Press, New York, 1982.
- [4] Dasu, T., Johnson, T., Koutsofios, E.: Hunting Data Glitches in Massive Time Series Data, *Proc. of the 2000 Conference on Information Quality*, 2000.
- [5] DeLone, W. H., McLean, E. R.: Information Systems Success: The Quest for the Dependent Variable, *Inf. Systems Research*, **3** (1): 60-95, 1992.
- [6] English, L.P., *Improving Data Warehouse and Business Information Quality*, New York: John Wiley & Sons, 1999
- [7] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery: An Overview, in: Fayyad U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (eds.), *Advances in Knowledge Discovery and Data Mining*, AAAI-Press, pp. 1-37, 1996.
- [8] Galhardas, H., Florescu, D., Shasha, D., Simon, E.: Declaratively Cleaning your Data using AJAX, *Journ. Bases de Données Avancées*, Oct. 2000.

- [9] Hinrichs, H., Aden, T.: An ISO 9001:2000 Compliant Quality Management System for Data Integration in Data Warehouse Systems, in: Theodoratos, D., Hammer, J., Jeusfeld, M., Staudt, M. (eds.): *Proc. Intl. Workshop on Design & Management of Data Warehouses (DMDW), Interlaken, Switzerland, 2001*.
- [10] Hipp, J., Güntzer, U., Grimmer, U.: Data Quality Mining – Making a Virtue of Necessity, *Proceedings of the ACM SIGMOD/PODS 2001 Conference (Workshop DMKD'01)*, 2001.
- [11] Hipp, J., Güntzer, U., Grimmer, U.: Integrating Association Rule Mining Algorithms with Relational Database Systems, *International Conference on Enterprise Information Systems* (forthcoming), 2001.
- [12] Hotz, E., Grimmer, U., Heuser, W., Nakhaeizadeh, R., Wieczorek, M.: REVI-MINER, a KDD-Environment for Deviation Detection and Analysis of Warranty and Goodwill Cost Statements in Automotive Industry, *Proceedings of the ACM SIGKDD* (forthcoming), 2001.
- [13] Hinrichs, H.: Datenqualitätsmanagement in Data Warehouse-Umgebungen, *Datenbanksysteme in Buero, Technik und Wissenschaft, 9. GI-Fachtagung BTW 2001, Oldenburg*, pp. 187-206, Springer, Berlin, 2001 (in German).
- [14] Hernandez, M. A., Stolfo, S. J.: The Merge/Purge Problem for Large Databases, *Proc. of the 1995 ACM SIGMOD Conference*, 1995.
- [15] Hinrichs, H., Wilkens, T.: Metadata-Based Data Auditing, *Data Mining II (Proc. of the 2<sup>nd</sup> Intl. Conf. on Data Mining, Cambridge, UK)*, pp. 141-150, WIT Press, Southampton, 2000.
- [16] International Organization for Standardization: *ISO 9000:2000: Quality Management Systems – Fundamentals and Vocabulary*, Beuth, Berlin, 2000.
- [17] International Organization for Standardization: *ISO 9000:2000: Quality Management Systems – Requirements*, Beuth, Berlin, 2000.
- [18] Jaro, M. A.: Advances in Record Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida, *Journal of the American Statistical Association*, **84**: 414-420, 1989.
- [19] Jarke, M., Jeusfeld, M. A., Quix, C., Vassiliadis, P.: Architecture and Quality in Data Warehouses, *Proc. of the 10<sup>th</sup> Intl. Conf. CAiSE\*98, Pisa, Italy*, pp. 93-113, Springer, Berlin, 1998.
- [20] Lee, M. L., Ling, T. W., Low W. L.: IntelliClean – A Knowledge-Based Intelligent Data Cleaner, *Proc. of the 6<sup>th</sup> ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, Boston, MA*, 2000.
- [21] Maletic, J. I., Marcus, A.: Data Cleansing – Beyond Integrity Analysis, *Proc. Conf. on Information Quality IQ2000, MIT, Boston, MA*, pp. 200-209, 2000.
- [22] Naumann, F., Leser, U., Freytag, J. C.: Quality-Driven Integration of Heterogeneous Information Sources, *Proc. of the 1999 Intl. Conf. on Very Large Databases (VLDB '99), Edinburgh, UK*, 1999.
- [23] Quinlan, R.: *GritBot – An informal tutorial*, <http://www.rulequest.com>, 2000.
- [24] Raman, V., Chou, A., Hellerstein, J. M.: Scalable Spreadsheets for Interactive Data Analysis, *Proc. of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD), Philadelphia*, 1999.
- [25] Redman, T. C.: *Data Quality for the Information Age*. Boston/London: Artech House, 1996.
- [26] Shanks, G.: *Semiotic Approach to Understanding Representation in Information Systems*,



- Proc. of the IS Foundations Workshop, ICS Macquarie University, Sydney, 1999.*
- [27] Shewhart, W. A.: *Economic Control of Quality of Manufactured Product*, D. Van Nostrand, New York, 1931.
  - [28] Wang, R. Y.: A Product Perspective on Total Data Quality Management, *Communications of the ACM*, 41 (2): 58-65, 1998.
  - [29] Wang, R. Y., Strong, D. M.: Beyond Accuracy: What Data Quality Means to Data Consumers, *Journal of Management Information Systems (JMIS)*, 12(4), 5-34, 1996.

## Bellman : A Data Quality Browser

Theodore Johnson and Tamraparni Dasu

AT&T Labs – Research

[johnsont@research.att.com](mailto:johnsont@research.att.com)

[tamr@research.att.com](mailto:tamr@research.att.com)

**Abstract:** When a data analyst starts a new project, she is often presented with one or more very large databases (containing hundreds or thousands of tables). Extracting useful information from the databases can be a difficult problem: documentation is usually minimal, the data is poorly structured and difficult to join, and the quality of the data is often poor. As an aid in exploratory analysis, we are developing a data quality browser that allows the analyst to quickly gain an understanding of the contents of the tables and their relationships. In addition, the browser serves as a platform for issuing data mining queries targeted towards a further understanding of data quality problems. We illustrate the utility of the data quality browser with several examples.

### Introduction

Starting a new and large data analysis project can be a disheartening prospect, because the first task is to understand a new data set. Often, the analyst is given access to a large production database (or a snapshot of it) and asked to produce results. The database can be very large, containing hundreds or thousands of tables. Documentation is usually minimal. The data is often of poor quality – missing data, mistyped data entries, square pegs in round holes, and so on. The task of extracting information from the database is frustrated by both the complexity of the data and data quality problems in the data.

There are many reasons why production databases become complex and disordered. New applications are fit into old data models, new functions require new tables for support, applications are merged, etc. (see [1] for more discussion). While it would be desirable to prevent the disorder from occurring in the first place, the analyst is (usually) not in a position to dictate how the production database evolves. Instead the analyst must work with the data as it is provided (and perhaps make recommendations about database restructuring for data quality improvement).

We are developing Bellman, a *data quality browser* to help the analyst explore and understand a large, complex, and dirty data set. The key idea is to issue *data profiling* queries, store the results (inferred metadata), and present them to the user. Data profiling results are stored in the database itself, allowing the user to issue ad-hoc queries on the inferred metadata. In this paper, we show how even the availability of simple data profiling queries, stored in the database and available for ad-hoc analysis, can simplify the task of understanding the structure of a database and the quality of the data.

Several companies offer a data profiling product (Evoke Software [2], Metagenix Inc [3], Knowledge Driver [4]), for use in database migration or re-engineering. Another related tool is the Integrity Analyzer [10]. While these tools can be used to support the data analyst, we are developing the data quality browser because:

- Our thesis is that a database browser which is enhanced with data profiling and special purpose data mining features is an effective tool for the data analyst carrying out exploratory and data quality analyses.
- We need to address special features of AT&T data.
- We plan to use the data quality browser as a test bed for research in data quality mining.

## Data Profiling

Data profiling refers to the process of querying a database to extract information about the properties of its data. An Evoke Software white paper [5] suggests three kinds of data profiling:

- *Column profiling* refers to the analysis of a particular field of a table. For example, statistics such as the number of NULL values, the number of unique values, the distribution of values, the length or range of the values in the field.
- *Dependency profiling* refers to the analysis of the correlation of fields in the same row of a table. This type of analysis includes the automatic determination of keys (collections of fields whose values are unique in every tuple) and functional dependences (relationships in which one collection of fields uniquely determine the value of another field e.g., a zip code functionally determines a state). See [6, 7, 8] for a discussion of academic research on this topic.
- *Redundancy profiling* refers to the analysis of the correlation of fields across tables. For example, this type of analysis can be used to suggest which fields can be used to join two tables.

It is clear that the column, dependency, and redundancy profiling discussed above is valuable for understanding the structure of a database, a pre-requisite not only for data migration but also for data cleaning. However, it is also clear that additional types of profiling are needed for a data quality browser. For example, additional types of information include summaries of temporal changes in the database, the structure of the database (i.e., a collection of tables) and the inter-relationships between databases.

In its current state, Bellman collects only a few types of statistics – the number of rows in a table, the number of unique values of a field, the number of times a field is NULL, the most frequent values of a field, dependencies between attributes, and “signatures” of field values. However, we have found that a browser that incorporates even this limited amount of information and presents it in an interactive manner greatly improves the database exploration process.

## Data Browser Architecture

The architecture of Bellman is shown in Figure 1. The browser consists of three components: a GUI, an analysis engine, and a profile repository. The analysis engine issues queries to the data tables (and perhaps to previously profiled information) and interprets the results. The browser profile repository stores profile information that the analysis engine has computed, and also retrieves it for display by the GUI or for use by the analysis engine. The GUI allows the user to issue commands to the analysis engine and to view results. The data quality browser is written in Java, and accesses the target database using JDBC. The analysis engine is written in C++ and accesses the target database using ODBC or OCI.

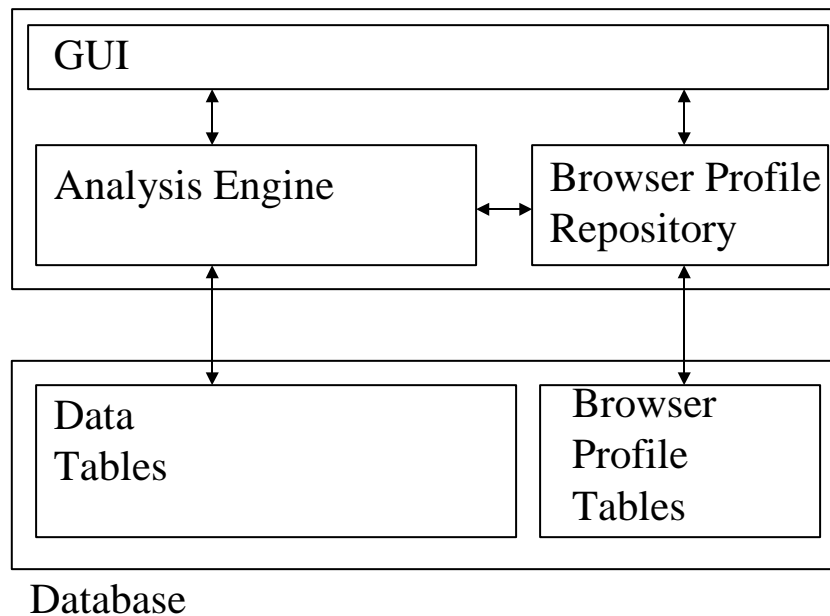


Figure 1 : Architecture of the Data Quality Browser

## Example Application

We illustrate the use of Bellman through a series of examples. Figure 2 shows the first screen displayed to the user after connecting to the database (we have censored tablespace names in order to protect AT&T data). The number in parentheses to the right of the tablespace name is the number of tables in the tablespace. Placing this information next to the tablespace name immediately informs the analyst of some tablespace properties. For example, the DB\* tablespace has not been loaded and the S\* tablespace is very large. From this window, the user can invoke some types of profiling analyses on all tables in a table space (using the “Do Table Counts” and “Do All Counts” buttons).

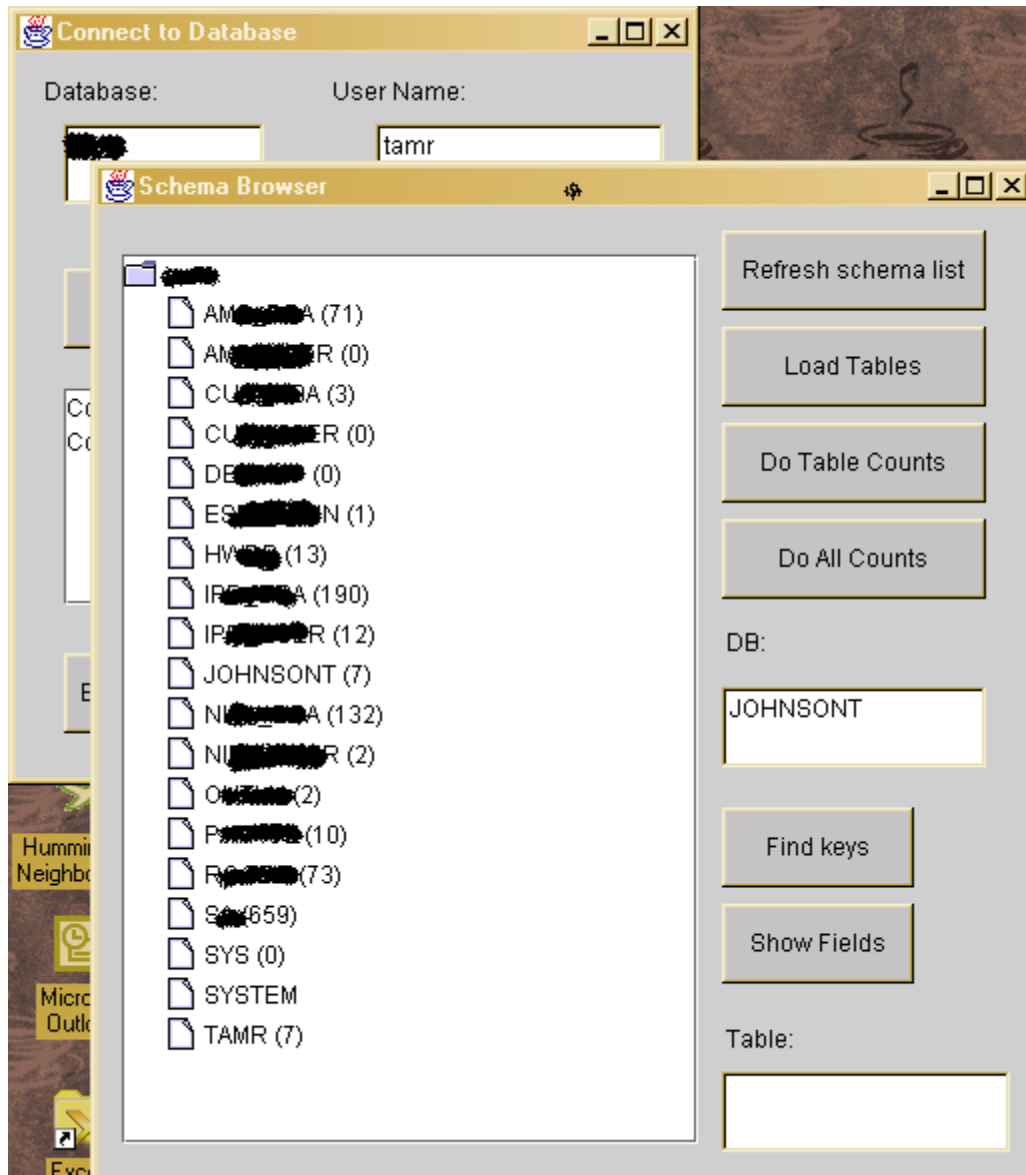


Figure 2 Tablespaces in the databases (annotated with the number of tables in the tablespace).

By clicking on the “Load Tables” button, the user can load all of the tables in a tablespace into the window, as is shown in Figure 3. The number to the right of the table name is the number of rows in the table. This information is useful when browsing a new database as one can see which tables are likely to contain detail data (e.g., very large tables) as opposed to dimensional or special purpose data. We can see that some of the tables are very small or even empty (i.e., the AUD\*X tables).

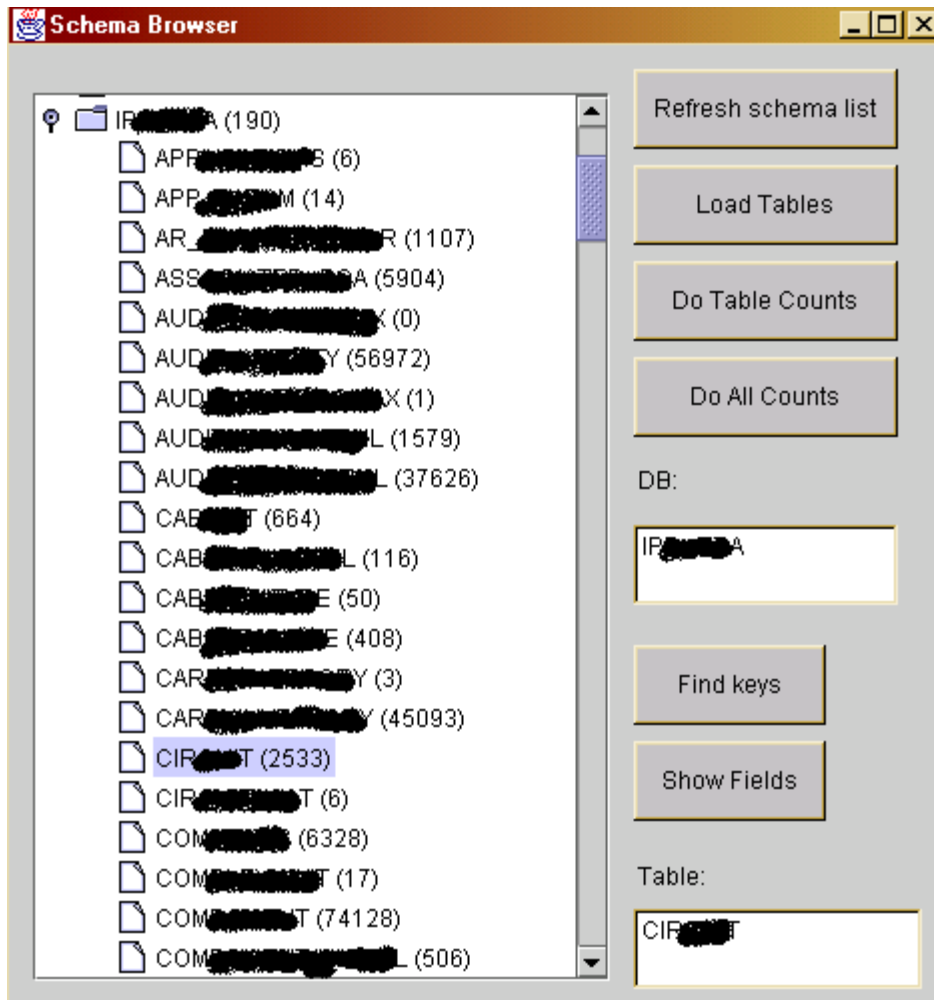


Figure 3: The tables in the IP\*A tablespace. The number of rows in the table is to the right of the table name.

We can examine the format of the CIR\*T table by clicking on the “Find Keys” button when the CIR\*T table is highlighted. A new window, shown in Figure 4, displays the fields of the table along with the field data type, the number of unique values of the field, and the number of null values of the field. The number of rows in the table is displayed in the upper right hand corner of the window. The combination of the number of rows in the table and the number of unique / null values of the field indicates the nature of the information in the field. In this table, it is clear that the 1<sup>st</sup>, 10<sup>th</sup>, and 11<sup>th</sup> fields are all keys. Several of the other fields are likely to contain significant amounts of information, but most of the fields are completely null. Developers often create null fields (and empty tables) as placeholders for the development of new system functions. However these placeholders are confusing to the analyst, hence the value of identifying them.

Use	Field	Type	Size	Unique	Null
X	CK...	DECIMAL	10	2533	0
X	CK...	CH	25	1196	1324
X	CL...	CH	8	3	0
X	TR...	CH	12	0	2533
X	TR...	CH	10	4	2240
X	A_...	DECIMAL	10	110	0
X	A_...	DECIMAL	10	576	0
X	Z_...	DECIMAL	10	325	0
X	Z_...	DECIMAL	10	442	0
X	AC...	DECIMAL	10	2533	0
X	ZC...	DECIMAL	10	2533	0
X	CK...	CH	25	0	2533
X	CK...	CH	25	0	2533
X	CK...	CH	25	0	2533
X	TR...	CH	10	0	2533
X	TR...	CH	10	0	2533
X	TR...	CH	10	0	2533
X	AS...	CH	6	1	2530
X	ZS...	CH	6	1	2530
X	Z_...	DECIMAL	10	0	2533
X	A_...	CH	8	0	2533
X	SE...	CH	5	1	2531
X	FA...	CH	5	1	2532
X	AU...	DECIMAL	10	701	0

Figure 4: Fields in a table.

An option in the “Field Chooser” window will open a new window (shown in Figure 5) that displays the most common values of the highlighted field. Although this is a very simple function, understanding the data in a field is critical for understanding the contents of the database. Making this information available at the click of a button greatly simplifies exploratory tasks.

The chart on the left hand side of the window is a chart of the counts of the most common values. In this case the chart area is almost empty, indicating the extreme skewness of this field. Because default values usually occur frequently, examining the most common values will often expose hidden default values. In this case, the “1995-08-30” date occurs suspiciously often, and is likely to be a default value. The other dates in the window occur suspiciously close together, so they might also be hidden default values.

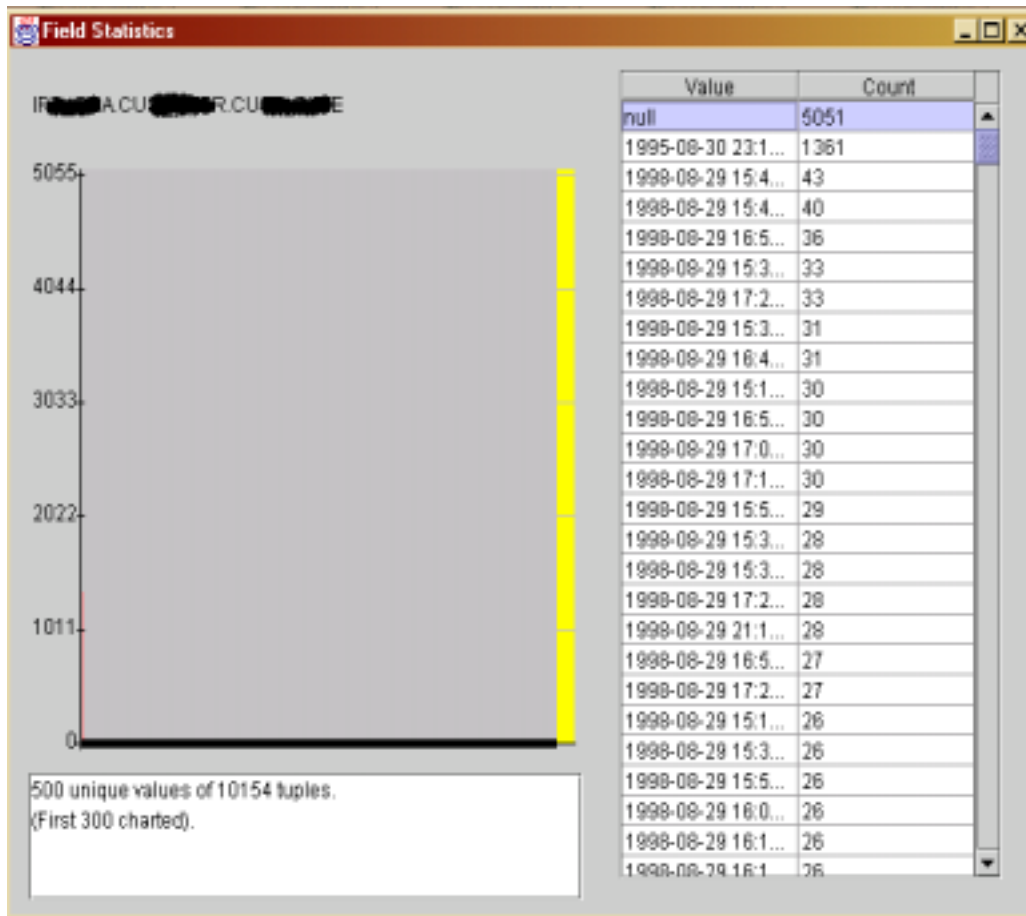


Figure 5: Identifying hidden default values

As Figure 2 shows, the database we set out to analyze is very large and complex. It is also poorly documented. In the process of understanding the structure of the databases, we need to know which fields (or collections of fields) are the keys of the table (that is, they are unique in every row). Because the data is likely to be dirty, we will accept “approximate” keys (i.e., they have few duplicate values).

A window for finding approximate keys is accessible from the “Field Chooser” window shown in Figure 4. Fields marked with an ‘X’ in the “Use” column are selected as input to the key finding algorithm (we might know in advance that certain fields will not be keys and we can deselect them). Figure 6 shows the key finding window displaying previously discovered results. At the top of the window is a collection of filtering parameters for the algorithm (to minimize the number of expensive queries necessary to search for a key). The algorithm found two 2-field keys (AS\*R and EQ\*D, AS\*R and PR\*E), and four 3-field keys. Note that this table has no single field key.

The key finding algorithm computes the counts of unique values of pairs, triples, etc. of the fields of the table. Therefore as a side effect the key finding algorithm finds *approximate dependencies*, in which one set of fields determines another. The key finding algorithms will not find all such dependencies (because it avoids doing all counts), and does not check the quality of the dependence. These dependencies are marked accordingly when they are entered into the profile repository.



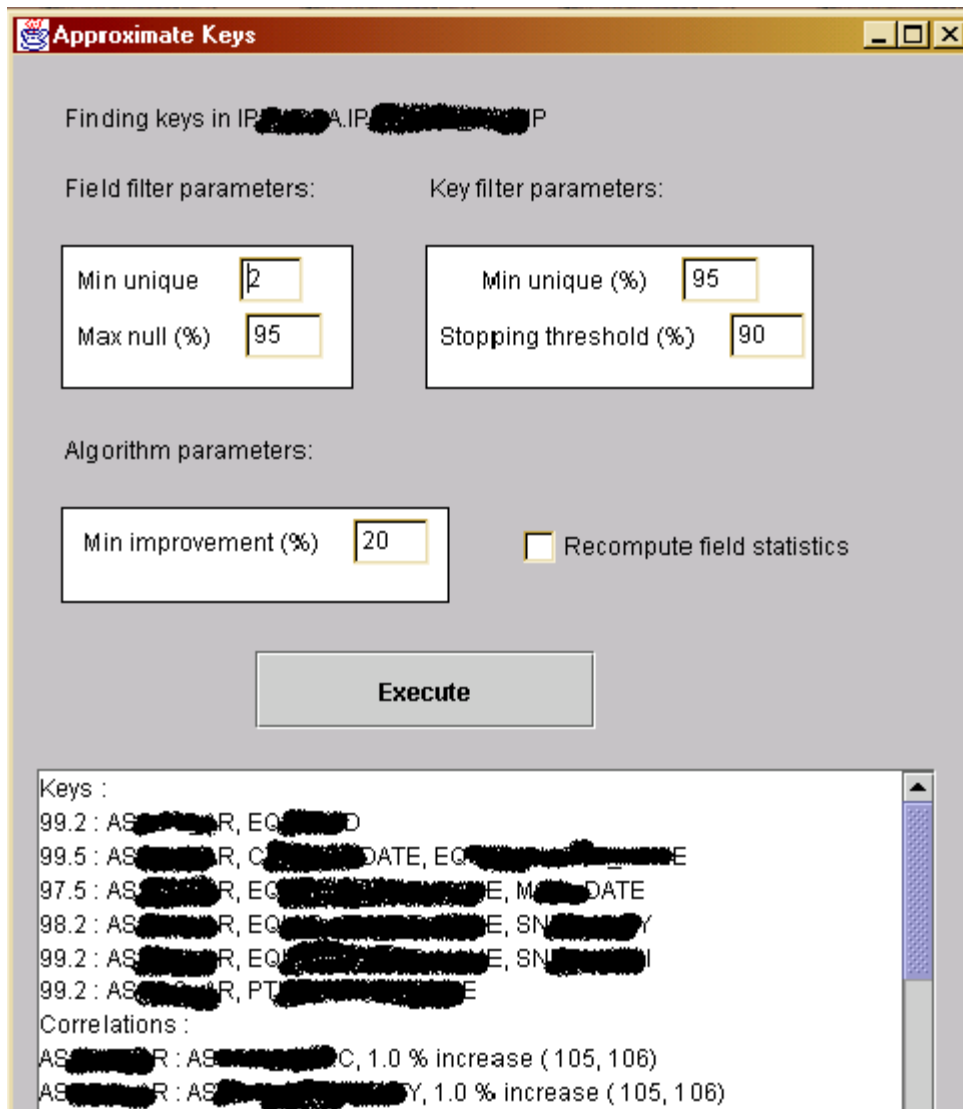


Figure 6: Finding keys and correlations.

Another important element in understanding the structure of the database is the finding of the fields which link tables together (i.e., can be joined). We make use of a *sketch* [9] (a special type of sampling) to represent the set of values in a field using only a few numbers (in the case of Bellman, 50 numbers). By comparing the sketches of two fields, we can approximate the size of their intersection – and therefore whether the tables can be joined using these fields.

In Bellman, we compute a sketch for each field with at least 20 unique values (a total of 4700 fields qualified). Figure 7 shows the set of fields found to have a large intersection with the ES\*.SNM\*.ROU\* field. The *resemblance* field contains the value which can be estimated directly from the sketch. Using the resemblance and the number of unique values in the fields, we can estimate the intersection size (in the (*intersection*) column. The “compute intersection” button will compute the actual intersection size, listed in the rightmost column. Some of the

actual intersection sizes are computed to be zero, because the fields contain values with the same text but of different data types (in this case, fixed-length character strings versus variable-length character strings). The estimated intersection size is a rough approximation of the actual intersection size, but it is able to distinguish between small, medium, and large intersections. We note that the manes of the similar fields are all significantly different; indicating that the use of field names alone to identify joins paths does not work well.

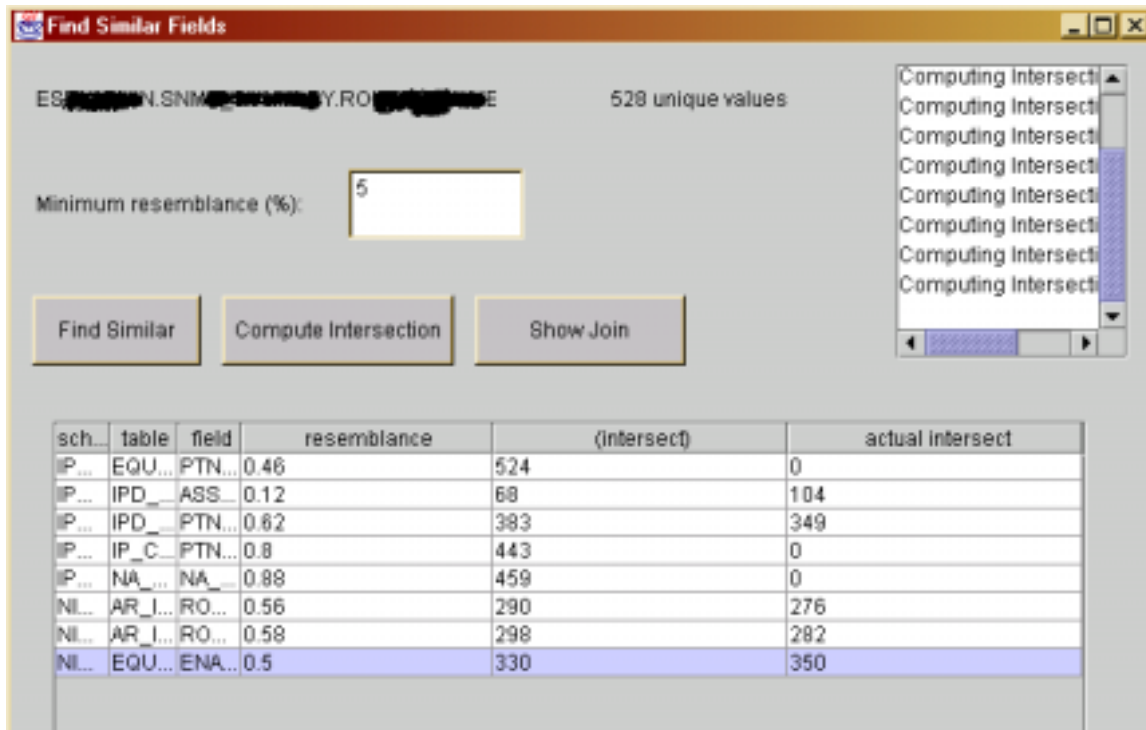


Figure 7. Finding similar fields (an “actual intersection” of zero occurs when the field values contain the same text but are of different data types).

Whenever Bellman performs an analysis task, it stores the profile data in the profile repository. One advantage of storing this data is to make the browser feel interactive – all of the screen shots in this paper appear within a second of the mouse click which requests it. By storing the profile data in the database itself provides another advantage – the profile data is available for ad-hoc queries. In Figure 8, we show a query that asks for fields that might relate to sites, and a portion of the result (the first three fields are the tablespace, the table, and the field). By examining the number of unique values of the fields, we obtain a hint about promising join paths. Other profiled data is also stored in the database and is available for querying – including the approximate keys.

Note that one of the fields in the profile repository table is the date at which the profile data was created. This field allows us to track changes over time.

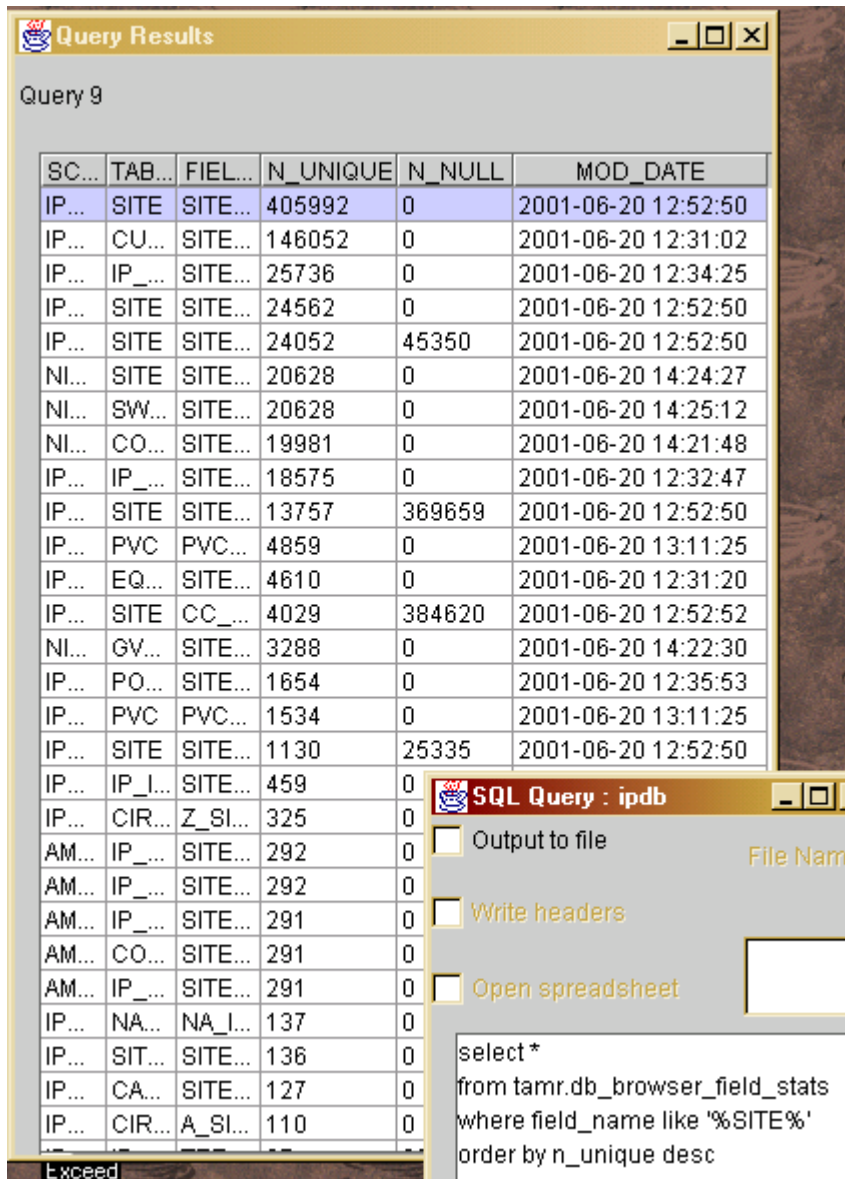


Figure 8: Querying the profile metadata.

## Conclusions

In spite of the simple and limited analysis that the current version of Bellman provides, it is already a very useful tool for the exploratory analysis of large, complex, poorly documented, and dirty databases. We have used it to find hidden default values and to explore join paths between different databases (which appear as different tablespaces in our examples). Information about table sizes, and counts of unique and null values allowed us to eliminate many dead ends when searching for join paths. Finding hidden null values is automated by having a

pre-formulated query available at the click of a button. More sophisticated analyses will allow us to automate our exploration even further.

We are continuing to Bellman, adding profile data and special purpose analyses to automate the process of exploring new databases and finding data quality problems. Tasks that we plan to support with automation include analyzing textual similarity fields, visualizing the structure of a database, and finding potentially corrupted data.

## **Bibliography**

- [1] *Data Quality Issues in Service Provisioning & Billing*, T. Dasu and T. Johnson, submitted to IQ2001.
- [2] Evoke Software <http://www.evokesoftware.com/>
- [3] Metagenix Inc. <http://www.metagenix.com/home.asp>
- [4] Knowledge Driver <http://www.knowledgedriver.com/>
- [5] <http://www.evokesoftware.com/pdf/wtpprDPM.pdf>
- [6] *Efficient Discovery of Functional and Approximate Dependencies Using Partitions*, Y. Huhtala, J. Karkkainen, P. Porkka and H. Toivonen, International Conf. On Data Engineering, 1998.
- [7] *Dependency Inference*, H. Mannila and K.J. Raiha, Proc. 13<sup>th</sup> Intl. Conf. On Very Large Data Bases, pg. 155-158, 1987.
- [8] *Bottom-up Induction of Functional Dependencies from Relations*, I. Savnik and P. Flach, AAAI Workshop (KDD '93), pg. 174-185, 1993.
- [9] *On the resemblance and containment of documents*, A. Z. Broder, Compression and Complexity of Sequences, pg. 21-19, 1997.
- [10] *Quality Information and Knowledge*, K.-T. Huang, Y. W. Lee, R. Y. Wang, Prentice-Hall 1999.

# **From Databases to Information Systems – Information Quality Makes the Difference**

Felix Naumann  
IBM Almaden Research Center  
650 Harry Road  
San Jose, CA 95120  
felix@almaden.ibm.com

**Abstract:** Research and business is currently moving from centralized databases towards information systems integrating distributed and autonomous data sources. Simultaneously, it is a well acknowledged fact that consideration of information quality—IQ-reasoning—is an important issue for large-scale integrated information systems. We show that IQ-reasoning can be the driving force of the current shift from databases to integrated information systems.

In this paper, we explore the implications and consequences of this shift. All areas of answering user queries are affected – from user input, to query planning and query optimization, and finally to building the query result. The application of IQ-reasoning brings both challenges, such as new cost models for optimization, and opportunities, such as improved query planning. We highlight several emerging aspects and suggest solutions toward a pervasion of information quality in information systems.

## **1. Information Quality**

The development of the Internet—especially the World Wide Web—has made it possible to access a multitude of data sources on almost any given topic. Web directories guide users to these sources, search engines let users discover sources previously unknown to them, and a huge number of Web sites act as data sources and provide the actual data. Most often, a user may choose between many alternative sources and source combinations to obtain the desired information item. This choice is advantageous but also time-consuming. It is advantageous to choose the most renowned, the fastest, or the most accurate sources. But it is time-consuming to come to this choice through trial and error. And it is even more time-consuming to access several sources in a row if the desired information is not provided by a single source, but is spread across those sources.

Consider search engines as data sources. Most users have chosen their favorite search engine, possibly based on personal experience in response time, relevancy of the results, ranking method, usability, etc. However, users might miss just the right Web page, simply because that page was not yet indexed or ranked sufficiently high by the search engine of choice. Meanwhile, other search engines might have already indexed this Web page. The user might turn to one of these

and may eventually find the wanted Web page. A meta-search engine solves this problem by simultaneously querying multiple search engines with the user's keywords. The results of the different engines are integrated to a combined response to the user. The drawback is that a quality-unaware meta-search engine uses engines of mixed quality, creating an inferior result.

Integrated access to data that is spread over multiple, distributed, autonomous, and heterogeneous data sources is an important problem for information consumers in many areas. In this paper we argue that the user goal of finding data is changing with the move from databases to integrated information systems: Users demand not the *correct* answer but are satisfied with approximate answers. Not the *complete* answer is necessary, but the answer should be relevant. Users demand not a *full* answer with all attributes, but are content with missing values. We also show that the optimization goal of finding a complete answer as quickly as possible has shifted to its dual problem of finding the best possible answer within a given cost/time constraint.

The emergence of Web-based information systems has amplified the known problems of poor information quality, but at the same time has reached an audience with a new requirement profile.

- **Technologies:** Due to the abundance of information sources on the Web, and due to many new technologies and architectures to fuse multiple sources to appear as one, source selection, information integration, and information filtering are important tasks to shield information consumers from data overflow, data errors, or simply low quality data.
- **Users:** The Web has made information available to a much broader audience. The vast majority are casual users who do not have a high stake in the outcome of the query, so that the answer must not be of highest quality. Additionally, users of Web-based systems are more aware of IQ problems and in consequence reduce their expectations. Integrated information systems can take advantage of lower expectations by reducing the amount of resources spent to answer a query.

Not having to or not being able to respond to user queries with maximal quality demands a comprehensive model of information quality.

## 1.1 Databases vs. Information Systems

Reasoning about information quality (IQ-reasoning) comes in two flavors: IQ-reasoning for database management systems (databases) and IQ-reasoning for information systems. Both the way of measuring and improving the quality of query results and the users expectations toward quality differ widely.

Information quality reasoning for databases differs from information quality reasoning for information systems. For illustration, we exaggerate the characterization of the two: Databases provide storage for structured, well-defined data and full query access to the data. In particular, data in databases is either gathered by the users of the database themselves<sup>1</sup>, or the users are able to update or delete the data. In essence, the control of the data lies with the users.

---

<sup>1</sup> Or people working for the same company gather it.

Information systems, on the other hand, are collections of structured, semi-structured, or unstructured information items such as text, tables, images etc. Integrated information systems gather this information (logically or materialized) from multiple, possibly autonomous information sources. Users of such information system have no control over the information it provides. We will argue in Section 2 that answering queries in these two types of systems differs, and that this difference is best described by an information quality model, and is best addressed by IQ-reasoning.

## **1.2 Scenarios**

We point out several exemplary scenarios of data usage that can be enhanced through IQ-reasoning.

**Search.** Searching is one of the most frequent used methods to gain information on the Web, and for inexperienced users the simplest way to pose queries. To search over multiple sources, meta searching techniques that distribute a search term to multiple sources are employed.

Meta-search engines are the most prominent example. However, search engines differ widely in the number of Web pages they have indexed, the amount of information they return for each page, their response time etc. An IQ-aware meta-search engine could improve results by taking such quality scores into account when deciding which individual sources to send the query to.

Other search-scenario examples are Web-based telephone and email directories, which can be integrated to increase the chance of finding a person (white pages) or company (yellow pages). Reasoning about their quality can identify large sources, sources with much additional data about a person like fax-number and email-address, source with up-to-date data, etc. Integration of the sources using this metadata can greatly enhance the final result and help filter out duplicates.

**Information integration.** Information integration is the process of taking multiple query results and merging them into a single response to the user. IQ-reasoning can enhance the integration of incoming query results in two ways: (i) Conflict resolution benefits from IQ-reasoning and (ii) result tuples can be ranked by their quality.

A data conflict occurs when two sources report different data values about the same real world entity. Resolution functions are employed to resolve these conflicts by deciding which value to include in the final result. Having knowledge about the quality of the sources, resolution functions can favor the value from the qualitatively better source. For instance, when deciding which address to include in a result for a person search, the address of the source with the higher update frequency can be chosen. Further examples are search engines that export a date attribute specifying the last update of the page index. In this case, the more recent data about the Web page should be chosen. Quality-dependent resolution functions enhance the query result by favoring high quality information over low quality information.

The presentation of the final integrated results also profits from IQ-reasoning. The quality determined for a source, a part of a plan, or an entire plan represents the quality of the data generated by the plan. Instead of dropping this information once the data is received, it can be used to rank

the query results. If the user does not specify another order, high quality tuples should be ranked first.

**Data mining.** Data Mining is the process of extracting previously unknown information from a set of data, such as a data warehouse. Data mining techniques are especially sensitive towards poor data quality [LLLK99]. For instance, outliers, i.e., data points that lie far from the average, severely skew the results of data mining algorithms. Outliers are usually produced where the data itself is generated: Sensors give incorrect output, a human accidentally adds a decimal to a number, etc. Therefore, any data mining method is preceded by a data cleaning technique to improve data quality, before applying the actual mining algorithms [Pyl99].

Also, other aspects of information quality play an important role for data mining. The completeness of the data is of importance so as not to mine on a subset of the available data. If the data, such as consumer behavior data, is obtained from a third party, the reputation and objectivity of the source are an important factor. IQ-aware data mining can improve the quality of the results.

### **1.3 Conclusion**

We define IQ-reasoning as the integration of IQ aspects to the process of planning and optimizing user queries against databases and information systems. IQ aspects include a set of IQ criteria, IQ assessment methods, and an IQ measure. When information sources store data and information about the same real world objects, information quality aspects constitute the main difference between the sources. These observations and others, such as those in [CZW98, Wei99, Wie99, MRV00], give rise to the following axiom:

*Information quality is the main discriminator of Web data sources, and information quality reasoning should be used to improve integrated query results.*

Or condensed:

*Information quality is the response time of the Web age.*

The rest of the paper is organized as follows: Section 2 analyses the traditional problem of query answering and optimization, and then describes the changes introduced by query processing over integrated sources. In Section 3 we present several necessary IQ components that enable IQ-reasoning for databases and information systems. Section 4 concludes the paper with an appeal to IQ-aware design and deployment of future information systems.

## **2. A Problem Shift**

Information quality (IQ) is the main discriminator of data and data sources on the Web. As we have seen in the previous section, the autonomy of Web data sources renders it necessary and useful to consider their quality when integrating their data. The information system paradigm shift—from central database management systems (DBMSs) to distributed multidatabase systems and finally to virtual, integrated World Wide Web information systems—has moved attention from *query processing* to what we call *query planning*.



**Query processing** is concerned with efficiently answering a user query to a single or multidatabase. In this context efficiency means speed. If not the speed of answering one query efficiently, it is the speed of the overall running system that is optimized. Many researchers and developers have designed sophisticated algorithms, index structures, etc., to enhance database efficiency. All those techniques have the same goal: Find the query execution plan that provides users with the correct and complete query result in an efficient manner.

**Query planning** on the other hand is concerned with finding the best possible answer given some cost or time constraint. Query planning involves regarding many query execution plans across different, autonomous sources that together form the complete result. Research has addressed the problem of determining *all* such plans [LRO96, Les98], but to the best of our knowledge only [NLF99] has addressed the problem of finding the *k* best plans, where “best” is defined through a quality model.

## 2.1 Query processing in DBMS

Databases store data and let users pose queries against it. The aim of query processing is to answer those queries with the available data. When answering user queries the DBMS assumes that users require correctness of the answer (R.1) and completeness of the answer (R.2 and R.3):

- **R.1:** The user expects only *correct* results, i.e., only tuples where all query conditions hold true. For example, a user of a data warehouse asking for departments with revenue of at least \$1,000,000 expects in the result *only* such departments.
- **R.2:** The user expects the result to be *extensionally complete*, i.e., to contain *all* correct tuples accessible by the integrated system. Continuing the example above, the user not only expects only departments with the specified revenue, but also *all* those departments (as long as their revenue data is stored in the database).
- **R.3:** The user expects the result to be *intensionally complete*, i.e., to contain all attributes specified in the query and contain non-null values in all the attributes. Continuing the example, if the user asked for the department name, its revenue, and its manager, the user expects all this information to be in the result. The user will not accept missing manager data (again, as long as this data is actually stored in the database).

Completeness and correctness in a DBMS are defined with regard to the content of the underlying database. The assumptions toward this database are that it contains only correct data, and that it contains all relevant data (closed world assumption). For instance, corporate users of a customer database assume that all customer data is correct and that data about all customers is actually stored within the database. He/she will not doubt the data provided, and will not turn to other databases suspecting that there is more customer data stored elsewhere.

Of course, DBMSs may also contain incorrect data; of course DBMSs may also not have all available data. However, compared to Web data sources, the owner of a DBMS has the power to change this situation. If there are inaccurate data, one can correct them, if data is missing, one can insert it. If the overall quality of the system is low, one can take measures to increase the quality aspects that are amiss. Web data sources on the other hand are autonomous. If complete-

ness and correctness or the overall information quality is not satisfying, there is usually nothing the integrating system can do about it.

The query processing component of a DBMS tries to answer a given query as cost-efficiently as possible, where cost-efficiency is usually defined as *response time*. Response time is the time a user must wait after submitting a query until reception of the complete result. A DBMS predicts response time using a cost model, which calculates the cost of database operations, such as join or selection operations, on different relations. In particular, the optimization component of a relational DBMS solves the following (simplified) problem:

*Given a set of relations, a user query against them, and a cost model, find the most cost-efficient order to access and combine the relations.*

The problem definition becomes more complicated for multiple parallel processors, multiple queries and multiple DBMSs. The basics however remain the same: The desired result (and hence, also its quality) is fixed—the aim of query processing is to generate this fixed result as efficiently as possible.

## 2.2 Query Planning in Integrated Information Systems

Query planning in information systems reverses this paradigm, as we will see: In general, the completeness and correctness assumptions about the underlying database do not hold for Web data sources in an open world—quite the contrary: A search engine will never have indexed *every* available Web page on the World Wide Web; stock information systems do not provide data on every stock; Web-based telephone directories only store data about some people, but never cover all telephone networks. That is, Web data sources are usually not complete. Correctness is also never guaranteed: Web pages may change after a search engine has indexed them; stock information systems purposely return delayed and thus outdated stock quotes; etc.

Further, typical users and Web servers have resource constraints: There might be technical constraints, such as a limited network bandwidth or limited access to the underlying data sources. Users may have constraints, such as a limited budget or limited time. Finally, users might have non-technical constraints, such as an unwillingness to browse a large result set. For example, a meta-search engine does not need to download all hits from all search engines it uses; instead, integrating the top ten hits usually suffices.

Knowing about incompleteness and limited correctness of Web sources, and having limited resources in terms of time and money, users of Web-based information systems make three concessions (C.1 – C.3) corresponding to the three requirements (R.1 – R.3) of the previous section:

- **C.1:** Users accept tuples where attribute values are incorrect but *close* to their selection condition. For example, a user querying for cars with a price lower than \$10,000 might also find cars for \$10,500 agreeable in the result. Allowing plurals or synonyms of search terms can extend the results of a search engine.
- **C.2:** Users accept *extensionally incomplete* answers in the presence of constrained resources. If, for any reason, the extensionally complete answer cannot be returned, the best

possible answer should be returned. A user of a search engine usually does not demand the entire result set but is satisfied with, say, ten Web pages. However, the result should consist of the Web pages best matching the keywords of the query.

- **C.3:** Users accept *intensionally incomplete* answers or answers with *missing values*—a partial answer is better than no answer. A user of a stock information service asking for companies whose stock quotes have risen more than 10 percent today along with a company profile is at least partially satisfied if the result contains companies without the profile information. Of course, those tuples for which the profile *is* available should be listed first, but others might still be a helpful part of the result. Integrated information systems should not reduce their information offer to the lowest common denominator of the participating sources, in effect throwing away information. For instance, a meta-search engine like MetaCrawler offers only title, description, and URL of a Web page, even though it queries several sources that offer much more information, such as language, size, etc.

In short, users cannot and do not expect the same type of results from a query to a Web-based and integrated information system as they do from a DBMS<sup>2</sup>. Hence, the problem of query processing is reversed:

*Given a set of relations/sources, a user query against them, a quality model, and a cost limit, find the highest quality combination of the relations/sources within the cost limit.*

Like a cost model, a quality model should be able to predict the quality of the result, retrieved from different sources and combinations of sources (see Sec. 3.2). The problem is reversed, because now the cost/efficiency is fixed, while the quality of the result is optimized. Cost can be fixed for several reasons:

- Users might not wait indefinitely for a result, but abort a query after a few minutes. For instance, a meta-search engine will not waste time by waiting for all search engines to return a result. Rather, it will integrate all results that have been returned within the first few seconds. In effect this fixes the time the information system has available to find some (best) answer to the query.
- If systems charge money to access the information, users might specify a spending limit. The higher the limit, the better the result is expected to be.
- To deal with large number of users, the information system itself might wish to spend only a certain amount of bandwidth or time for each query. This may limit the number of sources to access and the amount of data to be retrieved from each source for any given query.

---

<sup>2</sup> In fact, due to varying availability and frequent changes of sources, user cannot even expect two identical queries to produce the same result.

## 2.3 Conclusion

Improved technology has given rise to a new type of information system, which covers much more information at the cost of diminished quality. Simultaneously this technology is available to many more people, who have lower expectations toward the quality. IQ, as the main discriminator of these new systems, should play an increasingly important role in the systems design and deployments. The new paradigm of planning queries across multiple information sources provides quality-driven challenges throughout the integrated system:

- Design a **quality measure** with a set of IQ criteria and a way to measure them.
- Design a **quality model** to determine the quality of combinations of sources.
- Design **optimization algorithms** finding only a few best answers and dealing with quality model properties.
- Design **information integration** techniques that enhance the quality of the result.

The following sections highlight some of the necessary changes to meet the challenges.

## 3. New Components for IQ Pervasion

General definitions for information quality are “*fitness for use*” [TB98], “*meets information consumers needs*” [Red96], or “*user satisfaction*” [DM92]. These definitions are just as non-operational as Pirsig’s: “*Even though quality cannot be defined, you know what it is*” [Pir74]. Rather, we conceive quality as an aggregate value of multiple IQ-criteria. With this definition, information quality is flexible regarding the application domain, the sources, and the users, because the selection of criteria can be adapted accordingly. Also, assessing scores for certain aspects of information quality and aggregating these scores is easier than immediately finding a single global IQ-score.

### 3.1 An IQ Measure

Information quality is defined as a catalog of IQ-criteria. Several research projects have put together such general catalogs [Bas90, CZW98, JV97, Red96, Wei99, WS96] or compiled multiple catalogs [NR00, EW00]. These catalogs are proposals formulated in the most general way to allow for different interpretation depending on applications, data sources, and users. Many criteria are not independent and typically not all criteria should be used at the same time. Rather, an application specific selection of criteria helps to identify qualitatively good data and simultaneously reduces assessment cost. Information quality assessment is the process of assigning numerical values (IQ-scores) to IQ-criteria. An IQ-score reflects one aspect of information quality of a set of data items. Usually this set represents an entire data source, but it might be useful to assign scores to certain parts of data sources as well. We are aware of the difficulties of numerically expressing certain criteria. Because not the absolute IQ-scores are of importance, but rather their relative values, we believe that a numerical approach is reasonable. One of the major challenges is to make IQ-assessment feasible.

IQ-assessment is rightly considered difficult, and there have been only few research approaches addressing it. In [EW00] Eppler and Wittig observe that most existing assessment methods *solely* rely on users to provide IQ-scores [BMY99, WSKL99], even though many criteria can be assessed automatically (e.g., AVAILABILITY), or semi-automatically (e.g., COMPLETENESS) [NR00].

When assessing IQ-scores, it is necessary to observe the tradeoff between precision and practicality.

Below, we highlight two criteria that play an especially important role for both databases and integrated information systems (RESPONSE TIME and ACCURACY), and two exemplary criteria that emphasize the need for a broader definition of IQ for integrated information systems (COMPLETENESS and RELEVANCE).

**RESPONSE TIME.** Traditionally, the quality of a database is determined by its ability to respond quickly to queries, i.e., its RESPONSE TIME. The cost models of database optimizers, which have only speed as their goal, reflect this quality measure<sup>3</sup>. While this goal remains important for integrated information systems, methods of achieving low RESPONSE TIME have dramatically changed: In traditional databases much query processing time is spent in CPU-bound tasks such as the optimization algorithm itself, sorting a large set of values, or processing a join operator. Because of the distribution of sources in integrated information systems, this time is by far outweighed by network-bound tasks, such as retrieving a result set over a network, or waiting for a server response. In consequence, cost models of optimizers should adapt to this new situation. The key ability of Web-based information systems is not to answer queries quickly, but to answer them well.

**ACCURACY.** Recently, ACCURACY<sup>4</sup> has found more attention among database users and has been subject of several research projects, such as [HS98, MWS98, GFSS00]. Data quality is a quality measure for the relative amount of erroneous data stored in the database. Integrating multiple information sources is both a source for low ACCURACY and an opportunity to increase ACCURACY.

Autonomous information sources are a source for inaccurate data, or more precisely, a source for data with unknown and unalterable ACCURACY. In a centralized database the consumer of data typically owns the data. Insufficient ACCURACY is created by the consumer and can be remedied by the consumer. This is not the case for autonomous sources, such as sources on the Web.

On the other hand, the ability to access multiple sources to obtain information about the same real world object gives systems the opportunity to combine the data to a more accurate overall representation of the object (see Section 3.4).

**COMPLETENESS.** For many data sources and many application domains, size is everything: The more tuples and the more attributes a source provides, the more attractive it is to users. For instance, users typically prefer large search engines, i.e., search engines that have indexed a large number of Web pages, over small search engines. The rationale is that the larger a search engine is, the higher the probability is, that the result the user is looking for has been indexed by the search engine (and therefore appears in the result). Also, users prefer search engines that return more attributes than others, e.g., knowing the *byte size* of a Web page before clicking on the link is advantageous.

---

<sup>3</sup> In multi-user environments, some DBMS optimize for throughput, sacrificing response time of individual users for overall fast responses to all users. Essentially, the optimization goal remains time-based.

<sup>4</sup> ACCURACY is also known as “data quality”, as opposed to the more general term “information quality”.

Determining the “size” of a data source has only recently become a problem, when such meta-data became desired for autonomous sources of unknown size, such as typical WWW information sources. There are yet few projects striving to model or determine the size of Web data sources [BB98]. Chen and associates, who address query processing in the WWW, mention the quality criteria “size of result” and “number of documents accessed”, but they neither define them, nor point out the difference between the two [CZW98]. Motro and Rakov define a completeness criterion, counting the tuples in a source [MR98].

Calculation or prediction of join result sizes is an important technique for cost-based query optimization in DBMS [Ros81, GP89, SS94]. Mannino and associates give a survey on the suggested statistical values to store, how to maintain them, and how to use them to predict the result sizes of various database operations [MCS88]. Florescu and associates attempt to describe quantitatively the content of distributed autonomous document sources using probabilistic measures [FKL97].

All approaches have in common that they aim to predict the number of tuples/objects in the result, but none consider the amount of information returned per tuple. One source might provide rich information about the objects, another only a few attributes. In [NL00a] we propose to combine these measures with a density measure, which takes this aspect into account and also counts the frequency of null-values in the tuples—a common phenomenon in Web-based information sources.

**RELEVANCE.** RELEVANCE is the degree to which the provided information satisfies the users need. RELEVANCE is a standard criterion in the field of information retrieval [SM83]. There, a document or piece of data is considered to be relevant to the query, if the keywords of the query appear often and/or in prominent positions in the document. That is, word-counting techniques guide the relevance measure [GGMT99].

The importance of RELEVANCE as a criterion depends on the application domain. For instance, for search engines RELEVANCE is quite important, i.e., returned Web page links should be as relevant as possible, even though this precision is difficult to achieve. For instance, a query for the term “jaguar” to a Web search engine retrieves document links both for the animal and the automobile. If the user had the animal in mind, the links to automobile sites should have been considered as not relevant. The use of ontologies can help solve such problems to some extent. In other application domains, RELEVANCE is implicitly high. For instance, a query for IBM stock quotes in an integrated stock information system only returns relevant results, namely IBM stock quotes. The reason for this discrepancy is the definition of the domain: Search engines have the entire WWW as a domain and thus provide much data that is of no interest to the user. The domain of a stock information system is much more clear-cut and much smaller, so a query is less likely to produce irrelevant results.

For our purposes we reduce the RELEVANCE criterion to a correctness criterion. If a result is correct with respect to the user query, we assume that it is also relevant. If it is not relevant, the user query was either incorrect with respect to what the user had in mind, or it was not specific enough.

### 3.2 An IQ Model

An information quality model for integrated information systems takes on the role of cost models in DBMS. Given the quality of the participating sources (using the quality measure of the previous section), a quality model determines the quality of the query result. In a DBMS, the optimizer component explores different alternatives of executing a query (query execution plans), applies the cost model to each alternative and chooses the cheapest one. In an information system, the planner also considers different alternatives of executing a query (different combinations of sources) and applies the quality model to determine the best of those plans.

Given IQ-scores for all sources in all criteria, two problems must be solved: (i) IQ aggregation to determine the IQ-score for a plan in each criterion. (ii) IQ ranking to rank sources according to those multiple, aggregated IQ-scores.

**IQ Aggregation.** We propose merge functions as a method to determine IQ criterion scores of multiple sources. A merge function has a different interpretation for each criterion, reflecting properties of the underlying IQ-measure. For instance, the merge function for a PRICE criterion is the SUM function, because the price of each participating source in a plan must be paid. RESPONSE TIME has MAX as merge function, assuming parallel access to all sources in a plan. Merge functions can be quite complex, such as for the COMPLETENESS criterion [NF00].

Merge functions must be commutative and associative, so that a change of the execution order has no effect on its IQ-score. This property is desirable, as the user perceives the quality of the query result and not the quality of how the query result is obtained. The result of IQ aggregation for each combination of sources (plan) is a vector of IQ-scores with one dimension per criterion.

**IQ Ranking.** Given the IQ-vectors for a number of plans, we want to find a—possibly complete—qualitative ordering of them, to decide which one to execute. Methods to solve this problem are called ranking methods or Multi-Attribute Decision-Making methods (MADM). These face three general problems:

1. The range and units of the IQ-scores of the criteria varies. *Scaling methods* solve these problems.
2. The importance of the criteria varies in the eyes of a user. *User weightings* specified as a weight vector solve this problem.
3. The IQ-scores place the data sources into a multi-dimensional space with one dimension per IQ-criterion. Because there is no natural order on a multi-dimensional space, the *ranking methods* determine an ordering among the sources or combinations of sources (for an overview see [Nau98]).

After scaling and weighting the IQ-vectors, ranking methods map them to single scalar IQ-scores, which determine the rank among them. In a simple scheme, the best plans are subsequently queried, until the cost limit is reached. The following section describes more sophisticated approaches.

### 3.3 IQ Optimization

Query answering on the Web can be enhanced both in effectiveness and efficiency by using IQ-reasoning. As argued before, in the presence of resource constraints it is often not possible to execute all plans for a query. When not all plans can or should be executed, it is beneficial to restrict execution not to arbitrary plans, but to the best plans according to a quality model.

Recently, there has been some research on retrieving only the top  $N$  answers to a query [CK97, CG99, TGO99], where “top” is not in reference to information quality, but to some similarity measure. For instance, Chaudhuri and Gravano justify the relaxed requirement with a query for houses at a certain price and with a certain number of rooms against a real estate database. Obviously, the user does not expect only houses that *exactly* match the query, rather, the  $N$  results *best* matching the query should be returned. In an earlier article the authors based the top  $N$  approach on multimedia repositories, where objects typically match conditions only to a certain degree [CG96]. Therefore, it does not suffice to only return exact matches, nor is it feasible to return all objects that match to even the slightest degree. In their paper, the user must specify a minimum matching degree for result objects. This research amounts to the consideration of concession C.1 for query planning.

**Pre-optimization.** The potential number of plans for a user query is exponential in the number of relations in the user query and the number of sources. For instance, given 10 search engines, a meta-search engine could answer a user query by accessing any of the  $10! = 3,628,800$  combinations of them. Therefore, it is desirable to decrease this number before starting to generate these combinations. To this end, we use the source-specific IQ-criteria to “weed out” sources that are qualitatively not as good as others. Our goal is to find a certain number or percentage of best sources independently of any user-specific weighting or preference.

Mihaila and associates recently suggested using IQ-metadata for source selection [MRV00]. To this end, the authors suggest an extension of SQL with fuzzy conditions so that the user can specify the desired quality of the result.

**Optimization.** Essentially, an optimizer trying to find the best set of sources under some cost constraint must solve the Knapsack problem [GJ79]. The Knapsack problem is proven to be NP-complete, but there are many approximation algorithms that efficiently find near optimal solutions. The Knapsack problem assumes that combining sources has monotone benefit, i.e., adding a source to a combination never decreases overall quality. For general quality model we cannot assume this property. Consider the ACCURACY criterion. Adding an inaccurate source to a combination can decrease overall accuracy. In such cases, more quality-aware algorithms must be employed to guarantee certain optimality [NL00a]. An additional problem arises in a Web-based environment, where sources can fail without warning. Optimization algorithms must be able to dynamically adapt to such situations, for instance, by re-optimizing after each source failure, or by anticipating failures in the plan. Of course, consideration of an AVAILABILITY criterion for each source could reduce source failures in a plan: Unreliable sources will be valued at a lower quality and will less likely enter a plan in the first place.



**Post-optimization.** The order in which the results arrive from the participating sources is not necessarily the best order to present them to the user. The IQ-scores already obtained can be used to rank the result tuples, presenting the highest quality information first.

### 3.4 Information Integration

Data about a real world entity may be stored with differing attribute values at different sources. In strict, duplicate removing relational semantics, those tuples would appear individually in the result of any operator. Even in the presence of a unique ID-attribute identifying the entity, a relational operator returns multiple tuples about the same entity. Integration of results is reduced to concatenation of results. It is left to the user to identify and resolve data conflicts. We propose to only represent one result tuple per real world entity. To this end, traditional operators must be enhanced to include resolution functions as presented earlier.

Generally speaking, data sources overlap in two ways: extensionally and intensionally. The extensional overlap between two sources is the set of real world entities that are represented in both sources. The intensional overlap between two sources is the set of attributes both sources provide.

To make use of overlap and to integrate data in a meaningful and useful way, we must recognize identical entities represented in different sources (object identification), and we must be able to resolve any data conflicts between values (conflict resolution). Especially during conflict resolution, IQ-reasoning can greatly improve the result.

**Object Identification.** Integrating data from different sources requires that different representations of identical real world entities be identified as such [Ken91]. This process is called object identification. Object identification is difficult, because the available knowledge about the objects under consideration may be incomplete, inconsistent, and sparse. A particular problem occurs if no natural IDs exist. For instance, the URL of a Web page is a natural ID for the page. A meta-search engine can use the URL of reported hits to find and integrate duplicates. On the other hand, a used car typically has no natural ID. An integrated information system for used cars has no easy way of finding identical cars being advertised in different data sources.

Object identification in the absence of IDs, which is essentially the same problem as duplicate detection, record linkage, or object fusion [NL00, New88, PAGM96], is typically approached by statistical methods, for instance, using rough set theory [Zia99]. After having identified a set of tuples representing the same real world entity, they must be combined to a single representation. If their data values differ in some attributes, conflict resolution must be applied.

**Conflict Resolution.** Once different tuples have been identified as representing the same entity, the data about them can be integrated. In general, a result that is integrated from tuples of different sources, contains tuples where

1. some attribute value is not provided by *any* of the sources,
2. some attribute value is provided by *exactly one* source, and
3. some attribute value is provided by *more than one* source.

In the first case, it is obvious how the result is merged: Because the sources do not provide a value, the tuple in the result has no value either (null-value). In the second case, there is also no data conflict; thus, when constructing the result, the one attribute value can be used for the result tuple. Depending on the type of attribute and the type of sources, the fact that the data is missing in some sources can be taken into account as well, when determining the final attribute value.

The third case demands special attention. Several sources compete in filling the result tuple with an attribute value. If all sources provide the same value, that value can be used in the result. If this is not the case, there is a data conflict and a *resolution function* must determine what value shall appear in the result table.

Internal resolution functions are of various types, depending on the type of attribute, the usage of the value, and many other aspects [KCGS95, YM98]. A simple resolution function might concatenate the values and annotate them with the source that provided the value. Especially conflicts in textual attributes may be resolved in this way. Resolution functions need not only depend on the two conflicting attribute values. A resolution function could additionally depend on quality scores like AGE, favoring the more recent data value. In general, resolution functions should include IQ-scores in their decision and favor sources of higher quality.

## **4. Conclusion**

The surfacing of Web-based, integrated information systems has altered the way queries can be answered, and it has altered the expectations of users. In both cases, information quality is the main discriminator of the changes: Now, more queries can be answered with a larger underlying information space, at the cost of decreased quality of the answers. With more and more publicly available data and more and more autonomous sources, the problem will increase in the future. Now, more users can access the information sources and the information need of more users is covered. The expectations towards the quality of the answers to such queries are low.

To make full use of the opportunity to integrate large amounts of data from various sources, IQ-reasoning methods must be applied at all levels of the integration process. We hope that our findings about information quality and our IQ-reasoning techniques will find their way into integrated information systems, thereby regaining the ability to deliver high quality query results to users, once lost in the transition from centralized database management systems to systems integrating autonomous information sources.

## **References**

- [Bas90] Reva Basch. Measuring the quality of the data: Report on the fourth annual SCOUG retreat. *Database Searcher*, 6(8):18-24, October 1990.
- [BB98] Krishna Bharat and Andrei Broder. A technique for measuring the relative size and overlap of public web search engines. In *Proceedings of the International World Wide Web Conference*, Brisbane, Australia, April 1998.
- [BMY99] Mónica Bobrowski, Martina Marré, and Daniel Yankelevich. A homogeneous framework to measure data quality. In *Proceedings of the International Conference on Information Quality (IQ)*, pages 115-124, Cambridge, MA, 1999.

- [CG96] Surajit Chaudhuri and Luis Gravano. Optimizing queries over multimedia repositories. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD)*, pages 91-102, Montreal, Canada, 1996.
- [CG99] Surajit Chaudhuri and Luis Gravano. Evaluating top-*k* selection queries. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 397-410, Edinburgh, Scotland, 1999.
- [CK97] Michael J. Carey and Donald Kossmann. On saying "Enough already!" in SQL. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD)*, pages 219-230, Tucson, AZ, 1997.
- [CZW98] Ying Chen, Qiang Zhu, and Nengbin Wang. Query processing with quality control in the World Wide Web. *World Wide Web*, 1(4):241-255, 1998.
- [DM92] W.H. Delone and E.R. McLean. Information systems success: the quest for the dependent variable. *Information Systems Research*, 3(1):60-95, 1992.
- [EW00] Martin J. Eppler and Doerte Wittig. Conceptualizing Information Quality: A review of information quality frameworks from the last ten years. In *Proceedings of the International Conference on Information Quality (IQ)*, Cambridge, MA, 2000.
- [FKL97] Daniela Florescu, Daphne Koller, and Alon Levy. Using probabilistic information in data integration. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 216-225, Athens, Greece, 1997.
- [GFSS00] Helena Galhardas, Daniela Florescu, Dennis Shasha, and Eric Simon. An extensible framework for data cleaning. In *Proceedings of the International Conference on Data Engineering (ICDE)*, page 312, San Diego, CA, 2000.
- [GGMT99] Luis Gravano, Hector Garcia-Molina, and Anthony Tomasic. GLOSS: Text-source discovery over the Internet. In *ACM Transactions on Database Systems (TODS)*, 1999
- [GJ79] Michael R. Garey and David S. Johnson. *Computers and Intractability*. W.H. Freeman and Company, New York, NY, 1979.
- [GP89] Danièle Gardy and Claude Puech. On the effects of join operations on relation sizes. *ACM Transactions on Database Systems (TODS)*, 14(4):574-603, 1989.
- [HS98] Mauricio A. Hernández and Salvatore J. Stolfo. Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery*, 2(1):9-37, 1998.
- [JV97] M. Jarke and Y. Vassiliou. Data warehouse quality design: A review of the DWQ project. In *Proceedings of the International Conference on Information Quality (IQ)*, Cambridge, MA, 1997.
- [KCGS95] W. Kim, I. Choi, S. Gala, and M. Scheevel. On resolving schematic heterogeneity in multidatabase systems. In W. Kim, editor, *Modern Database Systems*, chapter 26, pages 521-550. ACM Press, New York, NY, 1995.
- [Ken91] William Kent. The breakdown of the information model in multi-database systems. *SIGMOD Record*, 20(4):10-15, 1991.
- [Les98] Ulf Leser. Combining heterogeneous data sources through query correspondence assertions. In *Workshop on Web Information and Data Management*, in conjunction with CIKM'98, pages 29-32, Washington, D.C., 1998.
- [LLLK99] Mong-Li Lee, Tok Wang Ling, Hongjun Lu, and Yee Teng Ko. Cleansing data for mining and warehousing. In *Proceedings of the International Conference on Data-*

- base and Expert Systems Applications* (DEXA), volume 1677 of *LNCS*, pages 751-760, Florence, Italy, 1999. Springer.
- [LRO96] Alon Y. Levy, Anand Rajaraman, and Joann J. Ordille. Query-answering algorithms for information agents. In *AAAI National Conf. on Artificial Intelligence*, pages 40-47, Portland, OR, 1996.
- [MCS88] Michael V. Mannino, Paicheng Chu, and Thomas Sager. Statistical profile estimation in database systems. *ACM Computing Surveys*, 20(3):191-221, 1988.
- [MR98] Amihai Motro and Igor Rakov. Estimating the quality of databases. In *Proceedings of the International Conference on Flexible Query Answering Systems (FQAS)*, pages 298-307, Roskilde, Denmark, May 1998. Springer Verlag.
- [MRV00] George A. Mihaila, Louiqa Raschid, and Maria-Esther Vidal. Using quality of data metadata for source selection and ranking. In *Proceedings of the ACM SIGMOD Workshop on The Web and Databases (WebDB)*, pages 93-98, Dallas, TX, 2000.
- [MWS98] Steve Mohan, Mary Jane Willshire, and Charles Schroeder. DataBryte: A proposed data warehouse cleansing framework. In *Proceedings of the International Conference on Information Quality (IQ)*, pages 283-291, Cambridge, MA, 1998.
- [Nau98] Felix Naumann. Data Fusion and Data Quality. In *Proceedings of the New Techniques and Technologies for Statistics Seminar (NTTS)*, Sorrento, Italy, 1998.
- [New88] H.B. Newcombe. *Handbook of Record Linkage*. Oxford University Press, Oxford, UK, 1988.
- [NF00] Felix Naumann and Johann Christoph Freytag. Completeness of Information Sources. Technical Report HUB-IB-135, Humboldt University of Berlin, February 2000.
- [NL00] Mattis Neiling and Hans-Joachim Lenz. Data integration by means of object identification in information systems. In *Proceedings of European Conference on Information Systems*, Vienna, Austria, 2000.
- [NL00a] Felix Naumann, Ulf Leser. Cooperative Query Answering with Density Scores. In *Proceedings of the Conference on Management of Data (COMAD 00)*, Pune, India, 2000.
- [NLF99] Felix Naumann, Ulf Leser, and Johann-Christoph Freytag. Quality-driven integration of heterogeneous information systems. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 447-458, Edinburgh, UK, 1999.
- [NR00] Felix Naumann, Claudia Rolker. Assessment Methods for information quality criteria. In *Proceedings of the International Conference on Information Quality (IQ)*, Cambridge, MA, 2000.
- [PAGM96] Yannis Papakonstantinou, Serge Abiteboul, and Hector Garcia-Molina. Object fusion in mediator systems. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 413-424, Bombay, India, 1996.
- [Pir74] Robert Pirsig. *Zen and the Art of Motorcycle Maintenance*. Bantam Books, New York, 1974.
- [Pyl99] Dorian Pyle. *Data Preparation for Data Mining*. Morgan Kaufman Publishers, San Francisco, CA, 1999
- [Red96] Thomas C. Redman. *Data Quality for the Information Age*. Artech House, Boston, London, 1996.

- [Ros81] Arnon Rosenthal. Note on the expected size of a join. *SIGMOD Record*, 11(4):19-25, 1981.
- [SM83] Gerard Salton and Michael J. McGill. Introduction to Information Retrieval. McGraw-Hill, Inc., New York, NY, 1983
- [SS94] Arun Swami and K. Bernhard Schiefer. On the estimation of join result sizes. In *Proceedings of the International Conference on Extending Database Technology (EDBT)*, volume 779 of *LNCS*, pages 287-300, Cambridge, UK, 1994. Springer.
- [TB98] Giri Kumar Tayi and Donald P. Ballou. Examining data quality. *Communications of the ACM*, 41(2):54-57, 1998.
- [TGO99] K.L. Tan, C.H. Goh, and B.C. Ooi. On getting some answers quickly, and perhaps more later. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 32-39, Sydney, Australia, 1999.
- [Wei99] Gerhard Weikum. Towards guaranteed quality and dependability of information systems. In *Proceedings of the Conference Datenbanksysteme in Büro, Technik und Wissenschaft (BTW)*, pages 379-409, Freiburg, Germany, 1999.
- [Wie99] Gio Wiederhold. Trends for the information technology industry. Technical report, Stanford University under sponsorship of the Japan Trade Organization, October 1999.
- [WS96] Richard Y. Wang and Diane M. Strong. Beyond accuracy: What data quality means to data consumers. *Journal on Management of Information Systems*, 12(4):5-34, 1996.
- [WSKL99] Richard Y. Wang, Diane M. Strong, Beverly K. Kahn, and Yang W. Lee. An information quality assessment methodology. In *Proceedings of the International Conference on Information Quality (IQ)*, pages 258-265, Cambridge, MA, 1999.
- [YM98] C. Yu and W. Meng. *Principles of database query processing for advanced applications*. Morgan Kaufmann, San Francisco, CA, 1998.
- [Zia99] Wojciech Ziarko. Discovery through rough set theory. *Communications of the ACM*, 42(11):54-57, November 1999.

# Information Engineering Approach for Decision-Making in Textiles<sup>1</sup>

Yatin Karpe, George Hodge, Neil Cahill\* & William Oxenham

North Carolina State University, NC

\*Institute of Textile Technology, VA

**Abstract:** Information Engineering is defined as a technique for extracting the "meaning" contained in information to allow the understanding needed by a user to make a "right" decision. This paper mainly describes and discusses the concepts underlying the Information Engineering approach, viz. Knowledge Management and Information Quality, and emphasizes their role and application as they relate to the Decision-Making process. The importance and concept of modeling will also be discussed, specifically with respect to one type of universally accepted form of modeling called IDEF (Integrated Definition Language). A case study using IDEF (IDEF0 and IDEF1x) for knit machine operation is presented in brief. Research is in progress to develop an Information Engineering methodology for mapping (using IDEF principles) and simplifying the decision-making process for a particular decision in textiles, resulting in more effective and efficient decision-making by the textile personnel.

## 1. INTRODUCTION

The information systems developed over the last 30 years have been heavily technology based, while decision-making remained a human thinking process. It can be envisioned that the information system was a sort of a pipeline through which information would flow past various

Users "Tap" Information Flow to make Decisions

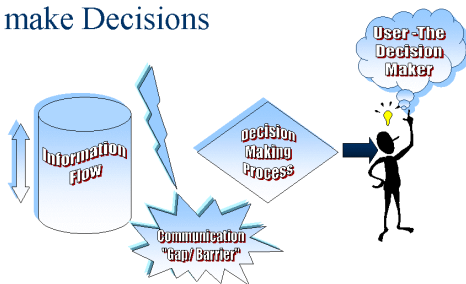


Figure 1: A Typical Communication System

users in the organization. As particular users desired/needed some information to make decisions, they "tapped" the pipeline. This basic approach of people tapping the information flow as needed to make decisions is basically the same today. Unfortunately, as businesses became more complex and the system could generate increasing quantities of information, then the discriminating power of the user to select and digest the "right" information was stretched to the limit. This phenomenon, also known as information overload, resulted in inferior or downgraded decision-making, due to the sheer volume of information that had to be processed in a given decision-making time frame. As this dilemma

of the information system and the human user increased, it evidently developed into a communication gap. Information systems primarily involve generating and distributing information throughout an organization. Such information transmission is the necessary first step in developing any communications capability. But information has no use, and therefore no value, until a decision-maker utilizes it. It is the human decision-maker who constitutes the

<sup>1</sup> Part of this paper was presented at the Textile Institute World Conference, Australia, 2001

second component of a communications system. The point of integration, where the human decision-maker “taps” into the information system is what forms the interface, and the design of this interface will influence the proficiency of converting information into decisions.

The ability of the decision-maker to make “right” decisions does not depend on information itself, but on the meaning and understanding derived from that information. If information access is a key driver, providing the right information filtering capabilities emerges as a major challenge. It is here that Information Engineering plays a vital role. Based on the above discussion and research conducted, the data-to-decision cycle model (The Decision Cycle) was developed to better understand the decision cycle and the Information Engineering approach. The conversion process by which raw data is translated into decisions of high quality is the Data-to-Decision cycle model. A parallel can be drawn between the components of the model and the present Information technology advancements. This process would assist in

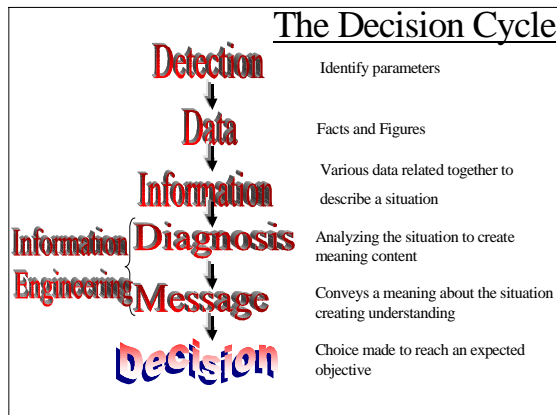


Figure 2: The Data-to-Decision Cycle Model

further diversifying the research so as to explore new and unique areas that would encompass the entire depth of the model. Different industry segments of the textile supply chain are being studied for the different Information Engineering stages.

## 2. PRE-INFORMATION ENGINEERING STAGE

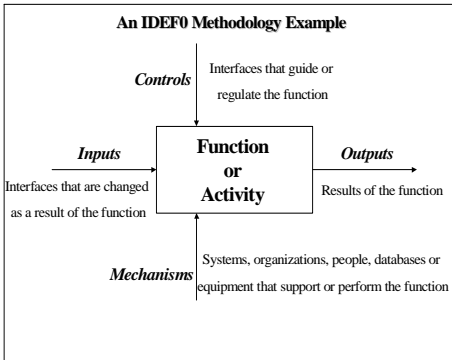
The “Detection” and “Data” parts of the model represent a data warehouse, which is a repository of the company’s historical data. In regards to these stages, studies are being conducted on the profiling, classification and standardization of the data. Attempts were made to describe and define the data elements of specific relevance to decision-makers, such as superintendent, foreman, operator, maintenance personnel, etc., in knitting mills, one of the components of the textile supply chain, which in turn will assist in better understanding the decision-making process. Modeling (IDEF Modeling in particular) is used for this purpose. Results obtained will be utilized in mapping out a particular decision-making process in the weaving industry of the textile supply chain.

### 2.1 IDEF MODELING

IDEF stands for Integrated DEFinition language. It is a methodology for describing, managing and improving complex processes and systems. IT provides a common, public-domain language for modeling and describing processes, data, requirements, as well as functions (Cete, 2001). It was first developed as part of the US Air Force ICAM (Integrated Computer Aided Manufacturing) Program in the early 1980s (ICAM, 1981). Since then, it has become the most well known and widely used method worldwide for modeling because of its simplicity. Originally, IDEF method comprised of three non-integrated modeling techniques, namely - IDEF0 (for functional modeling), IDEF1x (for data and information modeling) and IDEF2 for dynamic modeling (Vernadat, 1996). IDEF0 added features to the SADT methodology, which

made it a standard for use as the language to describe decisions, actions and activities that make up today's complex organizational environments (Wizdom Software, 1998).

## 2.2 IDEF0 PROCESS MODELING



An illustration of a basic IDEF0 model is shown in the figure. IDEF0 is a method designed to model decisions, actions and activities of an organization or system. IDEF0 models help to organize the analysis of a system and to promote good communication between the analyst and the customer. As a communication tool, IDEF0 enhances domain expert involvement and consensus decision-making through simplified graphical devices (Cete, 2001). As an analysis tool, it assists the modeler in identifying what functions are performed, what is needed to perform those functions, what the current system does

right and wrong. Thus, IDEF0 models are often created as one of the first tasks of a system development effort. The text in the box is the name of the activity for which it stands, typically a verb or verb phrase. Each side of an activity box has a specific meaning. The left side is reserved for inputs, the topside is reserved for controls, the right side is reserved for outputs, and the bottom side is reserved for mechanisms (resources). This reflects system principles; Inputs are transformed to outputs; Controls constrain or dictate under which conditions transformations occur; and, Mechanisms describe the resources needed to accomplish a function. A top-down diagramming method such as IDEF0 goes from the general to the specific, from a single diagram that represents an entire system to more detailed diagrams that explain how the subsections of the system work. The IDEF0 Methodology is primarily used for understanding the *AS-IS* (Present State) environment - the functions that are carried out, the relationships between them, and the logical breakdown of those functions into their sub-functions. The *AS-IS* scenario is then utilized to design and develop the *TO-BE* (Future Proposed State) environment, thus allowing for process or function or decision improvement.

## 2.3 KNITTING PROCESS MODEL

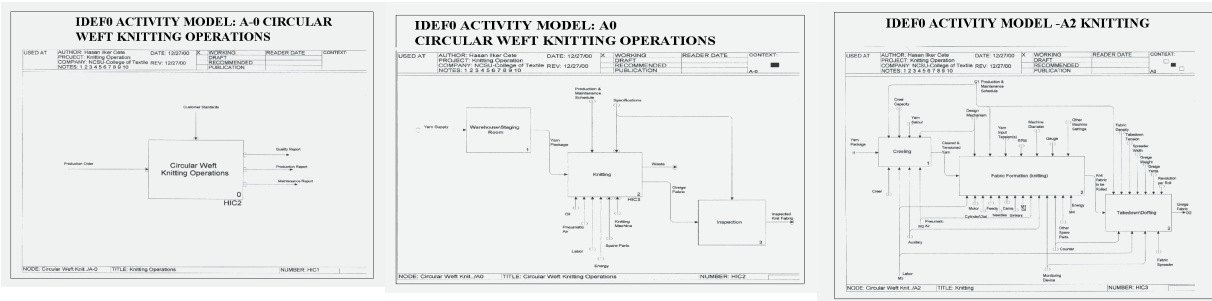


Figure 4: IDEF0 Knitting Models

A case study of IDEF0 modeling is shown in the three figures above (Cete, 2001). The base model is the A-0, followed by its decomposed diagrams. A-0 diagram (top) can also be called as the context of the function. When one clicks on the “circular weft knitting operations” activity box, the box pops up into another page that shows the presented A0 diagram (middle). A0 diagram can go one level down from “knitting” activity box. The lower-level diagram is named



as A2 (bottom) because the operation selected in A0 diagram is second one, “knitting”. The number seen on the lower right corner in “knitting” activity process box on A0 diagram formulates the lower level diagram node. Knitting represents the large diameter circular weft knitting machines’ operations.

### 2.4 IDEF1x DATA/INFORMATION MODELING

IDEF0 describes the activities needed to perform functions. IDEF1x describes the information or data needed to perform the same functions, both automated and nominal (Cete, 2001). It is important to model information. In order to avoid business problems, information needs to be accurate, timely, in the right place and in the right format. When the model is made, it is important to define all the information that is needed to meet the mission and goal of the organization. The IDEF1x standard defines what is to be known to do what is to be done. It is

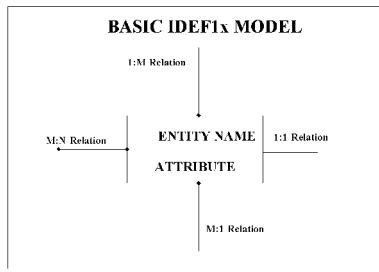


Figure 5: Basic IDEF1x Model

easy to communicate with others through standard syntax and representations by using IDEF1x. IDEF1x is based on the primitive form of the entity-relationship model as shown in the adjoining figure. Each information object is modeled as an entity (represented by a named rectangle and defined by its list of attributes, which can be listed in the box). Entities can be connected by named lines representing the relationships. Relationships can be of type 1:1, 1:n, or m:n, as shown in the figure 5.

### 2.5 KNITTING DATA MODEL

As part of the ongoing research in IDEF Modeling, and as a continuum to the IDEF0 modeling examples developed, a data/information model was developed for large diameter circular weft knitting as shown in the adjoining figure (Cete, 2001). A data model defines the entities and their attributes along with the relationship among the entities. A specific relationship is always

involved with two entities. In this data model there are 12 entities defined. They are production order, style card, yarn package, creel, knitting machine, takedown/doffing unit, monitoring device, greige fabric, maintenance report, quality report, production report and knit operator. All these entities involve knitting machine operation. The activity models and the data model were prepared by using WizdomWorks98 Office software; ProcessWorks and DataWorks.

#### IDEF1x DATA MODEL: CIRCULAR WEFT KNITTING OPERATIONS

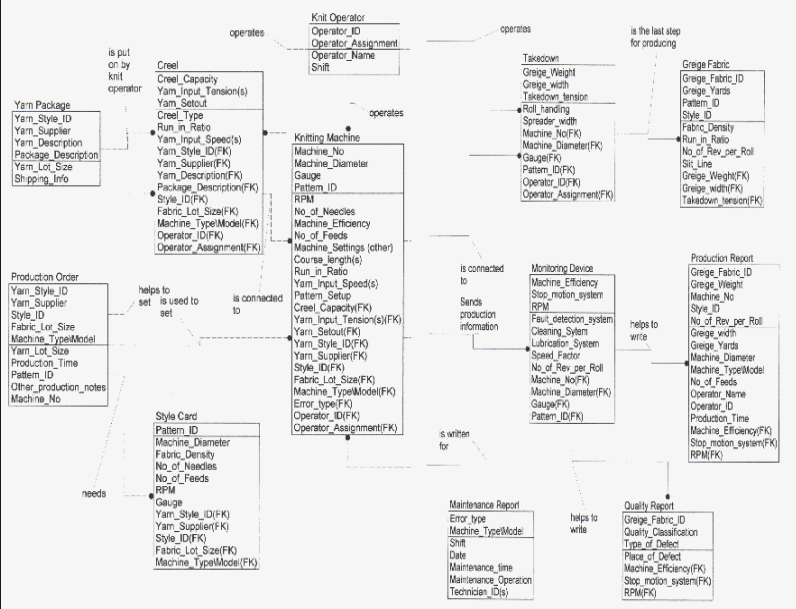


Figure 6: IDEF1x Knitting Data Model

(Results of this research (IDEF process and data models) will be discussed in detail during the presentation). These process and data models will further assist in defining the concepts of the decision-making process in knitting or similarly in any other sector of the textile supply chain. Thus, IDEF can be used as a tool to model the decisions and assist in developing an Information Engineering methodology. But before that is done, research is in progress to define and understand the various concepts (such as Knowledge Management and Information Quality) underlying the Information Engineering approach as being used in this specific research context.

### **3. INFORMATION ENGINEERING**

While the data being generated and information processed is at one end, the outcome of the decision being made is on the other end. But if we look at the center of the cycle, we realize that right decisions are not made merely by obtaining information, but by the correct diagnosis of the meaning of that information. If we interpret the meaning correctly, then we get the right message, which means we will probably make the right decision. It is here, in the center of the Decision cycle that machine intelligence can be created and it is here that Information Engineering can be applied to the manufacturing system. Information Engineering can thus be used to bridge the gap between the Data to Decision phases. It is the right decision that leads to favorable outcomes for that company and this is where information actually creates value.

Information Engineering is defined as a technique for extracting the "meaning" contained in information to allow the understanding needed by a user to make a "right" decision. According to one of the authors (Neil Cahill) "When one has to make a decision, it is the meaning contained in the information that is needed to make a "right" decision, and not the information alone. Of all the information available in existing plant reports today, only about 10-15% of the information contained in these reports is actually utilized. This low information utilization occurs due to the desired information (vital information) being buried in the report and requiring more diagnostic time than the user can provide" (Cahill, 1997). It must be realized that 80% plus of the time to reach a decision is used simply to find the right information. According to Myers, "While all communication contains information, not all information has communication value" (Cahill, 1985). Therefore, the goal should be to optimize the quality of the messages transmitted through the interface from the information system to human user. Information Engineering assists in this process. The ability of the user to make "right" decisions does not depend on information itself, but on the "meaning and understanding" derived from that information. The sender attempts to convey meaning through the message of the information. It is the message contained in the information that transfers meaning. This suggests that one way to improve the value of information is the designing of a message interface. This interface enhances the meaning of information in order for the user to better understand the business situation in which he/she must make a decision. Information is the raw material of the human thinking. But it is the "meaning and understanding" that is the raw material of decision thinking. Information by itself has no meaning or understanding. The human decision-maker acquires meaning and understanding not from the raw information, but rather from the "message content" of that information. This conversion process by which raw data is translated into decisions is the Data-to-Decision cycle model (Karpe, 2000). And it is in the center of this decision cycle that Information Engineering plays a vital role.

Information Engineering is a technique for identifying appropriate information for specific sets of decisions, and then tailoring and relaying this information to support effective management decisions. Therefore, by designing information in such a way that it's fit for use - making what we can call "actionable information" in today's fast-paced, information overloaded environment - one can construct meaning out of the clutter of disjointed data fragments. This means that when information is converted into a meaningful format, it leads to knowledge of that particular situation, resulting in effective and efficient decision-making. Information Engineering could prove to be a tool in knowledge mobilization, one of the twenty-four "Critical Business Practices" identified for the creation of an agile enterprise (Dove & Hartman, 1998). Hence one approach of understanding the model would be to study and analyze the knowledge management process, which draws a close resemblance to the decision cycle model.

#### **4. KNOWLEDGE MANAGEMENT**

Knowledge Management (KM) can be compared to the Industrial Revolution, where the work shifted from hand-centric labor to machine-centric processes leading to an explosive rise in production and new technologies. In the same way, KM drives the shift from the manual generation of information (paperwork, which is still common today) to complete electronic processing (with the ability to effectively use and apply information). This Knowledge management revolution leads to faster rates of producing knowledge assets, and new technologies for adapting knowledge faster (Leibmann, 2000). A recent study by the Cambridge Information Network found that 85% of chief information officers surveyed believe that managing knowledge creates a competitive advantage by fostering better decision-making (Taft, 2000). The primary goal of knowledge management is to deliver the intellectual capital of the firm to the knowledge workers who make day-to-day decisions that in aggregate determine the success or failure of a business (Microsoft(a), 2000). Developing such capabilities is what this research is all about. But before we try to develop a tool to implement KM practices, it will be useful for us to lay some groundwork on the fundamentals of knowledge management.

##### **4.1 DEFINITIONS: KNOWLEDGE AND KNOWLEDGE MANAGEMENT**

Defining knowledge is an essential first step when investigating knowledge management. Knowledge can be defined in several different ways (Beckman, 1997; Van der Spek, 1997), one of them being information that has been organized, analyzed and reasoned to make it understandable and applicable to actively enable performance, problem-solving and decision-making (Turban, 1992). Knowledge is composed of two main types, focal and tacit: focal being knowledge about the object or phenomenon in focus (or of explicit interest) and tacit is knowledge that is used as a tool to manage or improve what is in focus. For example, if a certain piece of information represents focal knowledge (say end-breaks in ring spinning), how a person perceives that information and operates on it is driven by his tacit knowledge (decision to continue or stop that machine, examine the physical attributes of the spindle or yarn, and so on). Knowledge Management also has been defined in several different ways by different authors (Petrasch, 1996; Wiig, 1997; O'Dell, 1996). One definition being the practice of identifying, capturing, organizing and processing information to create knowledge, which is then distributed and otherwise made available for others to use and to create more knowledge (Radding, 1998), which in turn creates more value, and it is precisely this value-enhancement in decision-making that is needed to be achieved with the Information Engineering approach.

## 4.2 KNOWLEDGE PROCESS, FUNCTIONS AND PHASES

The very process that is used to create, communicate and apply knowledge results in new knowledge. New knowledge almost always begins with an individual. A brilliant researcher has an insight that leads into a new patent. A shop-floor worker draws on years of experience to come up with a new process innovation. In each case, an individual's personal knowledge is transformed into organizational knowledge, valuable to the company as a whole. The result is a knowledge cycle in which data is transformed into information. The information is then culled and enhanced and transformed into knowledge. The knowledge is then applied and the results are documented, creating new data and information and recommencing the process. And this in essence is the aim of this project. Once the knowledge is engineered for specific end-use and assists in decision-making for a particular decision, it could be made available for not only enhancing the efficiency and effectiveness of that decision, but also used as a basis for other decisions across various other functions and departments.

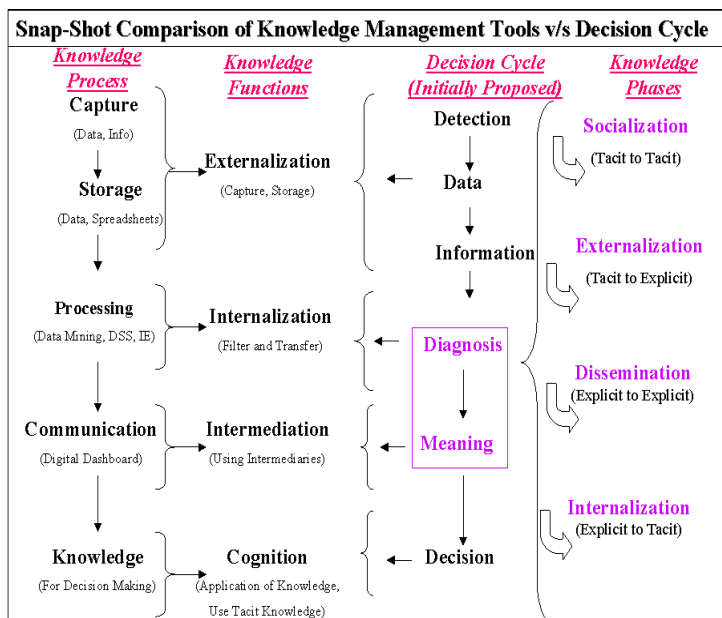


Figure 7: Knowledge Management Tools v/s Decision Cycle

The following figure is a summarized version of understanding the relation of the various components of knowledge management versus the decision model and the possible approach that could be used for the purpose. The knowledge process (Radding, 1998) and the knowledge functions (Frappaolo, 1998) are self-explanatory and draw a close resemblance to the knowledge cycle explained above and the Decision cycle model (Karpe, 2000). Related to this knowledge process and functions are the four basic knowledge phases (Radding, 1998; Malhotra, 2000):

Socialization (Tacit -to- Tacit) - The conversion from tacit knowledge to tacit knowledge through sharing of experiences, imitation and practices. This type of activity occurs during coaching, in apprenticeships, at conferences and seminars or simply during employee interaction during recesses.

Externalization/Articulation (Tacit -to- Explicit) - Also referred to as capture. The conversion from tacit knowledge to explicit knowledge, usually by articulating the tacit knowledge and turning it into explicit form, such as a report or document.

Dissemination/Combination Phase (Explicit -to- Explicit) - The conversion from explicit knowledge to explicit knowledge by the owner's sharing it with one another. Dissemination is the primary way knowledge is leveraged throughout the organization.

Internalization (Explicit -to- Tacit ) - The conversion from explicit knowledge back to tacit knowledge, enabling workers to incorporate the knowledge into the way they respond and behave when faced with a situation or problem, to which the knowledge applies.

This spiral of knowledge elicits one fact; knowledge creation results form efficient and effective use of the existing information. And in order for the user or decision-maker to utilize this information effectively, the quality of that information is of prime importance. Enhancing the quality of information results in better understanding of the situation, resulting in effective decision-making.

## **5. INFORMATION QUALITY**

Information quality problems hamper virtually every area of a business, from the mailroom to the executive office. Every hour the business spends hunting for missing data, correcting inaccurate data, working around data problems, scrambling to assemble information across disintegrated databases, resolving data-related customer complaints, and so on, is an hour of cost only, passed on in higher prices to the customer. That hour is not available for value-adding work. Senior executives at one large mail order company personally spend the equivalent of one full-time employee (senior executive) in reconciling conflicting departmental reports before submitting them to the Chief Executive Officer. This means that the equivalent of one senior executive's time is spent or wasted because of redundant and inconsistent (nonquality) data. According to Bill Inmon, 80 to 90 percent of the human efforts in building a data warehouse are expended in handling the interface between operational and data warehouse environments (Inmon, 1992). The bottom line is that information quality problems hurt the bottom line. The social and economic impact of poor-quality data costs billions of dollars (Wang, 1995; Strong et al, 1997). Quality experts agree that the costs of nonquality are significant. Quality consultant Phil Crosby, author of *Quality is Free*, identifies the cost of non-quality to manufacturing as 15-20 percent of revenue (Crosby, 1979). J.M. Juran pegs the cost of poor information quality at 20 to 40 percent of sales (Juran, 1988), Kearny puts this cost at 25 to 30 percent of sales dollars, while in service companies, poor quality can amount to an increase of 40 percent in operating costs (Boyle, 1992). Furthermore, as much as 40 to 50 percent or more of the typical IT budget may actually be spent in "information scrap and rework", a concept well known in manufacturing. Following the analogy between manufacturing and information systems, we can clearly see that there is a significant economic benefit to be gained if data or information quality can be managed effectively (Wang, 1992). Information quality is a business issue and information quality improvement is a business necessity. For organizations in a competitive environment, information quality is a matter of survival, and then of competitive advantage. For organizations in a public and not-for-profit sectors, information quality is a matter of survival, and then of stewardship, of stakeholder resources.

### **5.1 INFORMATION QUALITY: DEFINITION -**

Information Quality is defined as "consistently meeting the knowledge-workers and end-customers expectation", through information and information services (English, 1996), enabling them to perform their jobs efficiently and effectively. Information quality describes "the attributes of the information that result in user (customer) satisfaction (Nayar, 1996). There are two significant attributes or definitions of Information Quality. One is inherent and the other is pragmatic information quality (English, 1999).

*Inherent Information Quality* is the correctness or accuracy of the data. If all facts that an organization needs to know about an entity are accurate, that data has inherent quality - it is an electronic representation of reality.

*Pragmatic Information Quality* is the degree of usefulness and value data has to support the enterprise processes that enable accomplishing enterprise objectives. In essence, pragmatic information quality is the degree of customer satisfaction derived by the knowledge workers who use it to do their jobs.

Information can be represented by the formula (English, 1999):

$$\text{Information} = f(\text{Data} + \text{Definition} + \text{Presentation})$$

The three components that make up the finished product of information are separate and distinct components that must each have quality to have information quality. If we do not know the meaning (definition) of a fact (data), any value will be meaningless and we have non-quality. If we know the meaning (definition) of a fact, but the value (data) is incorrect, we have non-quality. If we have a correct value (data) for a known (defined) fact, but its presentation (whether in the form of a written report, on a computer screen, or in a computer-generated report) lacks quality, the knowledge worker may misinterpret the data, and again we have non-quality. Thus information quality is not an esoteric notion; it directly affects the effectiveness and efficiency of business processes.

## 5.2 DIMENSIONS OF INFORMATION QUALITY -

Studying the 5-dimensional model creates another form of representing the information quality phenomenon (Albrecht, 1999). As seen in the figure below, the Data logistics and data protection parts are concerned with the capture and storage of the data. The Information behavior encompasses what human beings do with the data and information, viz. recording information manually or by computer, paraphrasing, getting information from others, etc.

Information design is at the heart, using software and other tools to create new information and

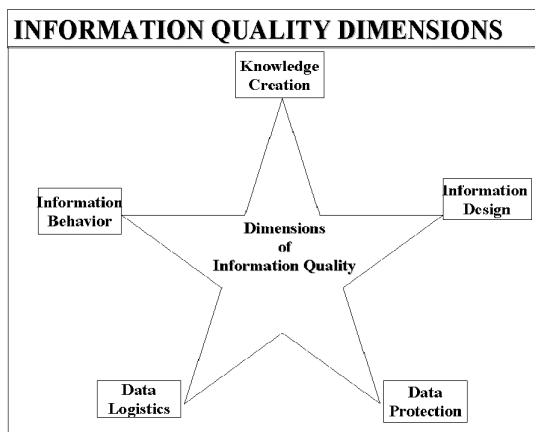


Figure 8: Information Quality Dimensions

knowledge by transforming source information into meaningful form. This meaning is then used for knowledge creation, wherein the human skill of drawing insights and conclusions from the existing information comes into play. It could also lead to new inventions, conceptualizing new ideas, conceiving new strategies and building new models and rethinking existing beliefs. Thus, the five dimensions further strengthen the Knowledge Process/creation and the Knowledge phases concepts that were discussed earlier and will provide us with a strong foundation to design a new approach.

In addition to these concepts, literature has also been reviewed in the areas of cognition (memory, inferences, etc.) and problem solving and reasoning (heuristics, biases, etc.). All this literature reviewed will be utilized to design and develop an effective research approach towards meeting the main objective of this research.

## 6. RESEARCH APPROACH -

Based on the literature reviewed in different areas, information will be gathered using a case study approach, focused on the current decision-making process of textile companies. Case study research is a widely used and accepted approach in the development of modern management

theories and models, since it enables development of new ways of describing reality using qualitative information. Furthermore, it provides reasonably good potential for result generalization from very few cases, or even a single case, based on the opportunity for holistic view of a phenomenon or series of events. The industry sample will include manufacturing sites from one segment of the textile supply chain complex, viz. weave room (weaving). Thus, the unit of research analysis will be a textile (weaving) manufacturing plant, and not the entire textile industry. The main study will focus only on approximately 3 to 5 textile plants for its study. In order to obtain information for studying and analyzing the present decision-making process, concepts and ideas learned from the literature reviewed and the knitting industry study will be utilized and interpreted to formulate a structured interview. Concepts, such as problem solving heuristic, the knowledge phases (Socialization, Externalization, Dissemination, Internalization) and cognition (different biases) will be used to define the “As-Is” scenario and possibly propose a structured “To-Be” approach for effective and efficient decision-making with regards to the weave room efficiency decision. Initially, a pilot study will be undertaken in order to evaluate the functionality and the result generation capability of the interview. This process will assist in restructuring or refining the interview process for more effective and efficient response generation. Efforts will be made to collect and analyze plant reports, forms and such other documents that are deemed fit for information collection and analyzing situations. The information obtained will be graphically represented using the Wizdom software for process modeling (IDEF0) and data modeling (IDEF1X) methodology. This process will also assist in identifying and isolating the inherent deficiencies in the present decision-making process and standardize a potentially simple map of a decision-making process for the weave room efficiency decision, which could later be utilized across the entire textile chain for different other decisions in different parts of the textile chain.

## **7. SUMMARY -**

The goal of the current research is to fundamentally enhance the decision-effectiveness of the textile personnel on the plant floor, using the data-to-decision cycle model as the basis. Since this research is in progress and has several angles to it, this paper specifically deals only with the research foundation and the concepts underlying the approach that will be adopted. Results of the approach adopted will be ready for presentation at the next years' conference. The Information Engineering approach could prove to be a valuable asset in improving data and information quality with the use of knowledge management and modeling tools. Thereby reducing the overload (information overload) that tends to occur in the present generation highly automated machinery and making it simpler for the personnel on the plant floor to make the right decision. In combination with Knowledge Management and Information Quality, Information Engineering can eventually lead to the development and creation of a kind of Digital Decision Dashboard (D<sup>3</sup>), which would be the decision-making tool of the next generation for the textile industry. A digital dashboard is defined as a customized solution for the knowledge workers that consolidates personal, team, corporate and external information and provides a single click access to analytical, and collaborative tools (Microsoft, 2000). It brings an integrated view of a company's diverse sources of knowledge to an individual's desktop, enabling better decision-making by providing immediate access to key business information. The D<sup>3</sup> can be a similar tool for decision-making in textiles, capturing and disseminating vital management information for effective and efficient decision-making, thus addressing a critical need presently facing the textile industry.

## ACKNOWLEDGEMENTS

The authors would like to thank the National Textile Center (NTC) for providing the required support for this research.

## REFERENCES

1. Albrecht, K. "Information, the next quality revolution?", Quality Digest, June 1999.
2. Beckman, T. "A Methodology for Knowledge management", IASTED- AI and Soft Computing Conference, Banff, Canada, 1997.
3. Beckman, T. "The Current State of Knowledge Management", The Knowledge Management Handbook, Jay Liebowitz, ed., CRC Press, 1999
4. Boyle, S. "Quality, Speed, Customer Involvement and the New Look of Organizations" seminar, Excel, pg.17, 1992.
5. Cahill, N. "Analyzing textile plants of the 21<sup>st</sup>. Century", ISA - textile division, June 1997.
6. Cahill, N. "*Information Engineering – Measuring use value of information*", ITT Report, 1985.
7. Cete, H. "Information Technology and Data Modeling in Large Diameter Circular Weft Knitting with Data Standardization and Profiling", M.S. Thesis, North Carolina State University, Raleigh, NC, 2001.
8. Crosby, P. Quality is Free, Penguin Group, New York, 1979.
9. Dove, R & Hartman, S. "An Agile Enterprise Reference Model". Available: <http://www.parshift.com/aermodA0.html>
10. English, L. "Defining Information Quality", Improving Data Warehouse and Business Information Quality, Wiley & Sons, New York, 1999.
11. English, L. Information Quality Improvement: Principles, Methods and Management, Seminar 5th. Ed., Information Impact International, Inc., Brentwood, Tennessee, 1996.
12. Frappaolo, C. "Defining knowledge management: four basic functions", ComputerWorld, February 1998
13. ICAM. "US Air Force Integrated Computer Aided Manufacturing (ICAM) Architecture", Part II, Volume IV-Functional Modeling Manual (IDEFO), Air Force Materials Laboratory, Wright-Patterson AFB, Ohio 45433, AFWAL-tr-81-4023, 1981.
14. Inmon, B. "Data Warehouse - Into the 90's", All About IRM'92 Conference, Beaver Creek, CO, July 1992.
15. Juran, J. Juran on Planning for Quality, Free Press, New York, 1988.
16. Karpe, Y., Hodge, G., Cahill, N. & Oxenham, W. "Information Engineering: Enhancing Decision Effectiveness in Textiles?", Proceedings of the 80<sup>th</sup> World Conference of the Textile Institute, Manchester, UK, April 2000.
17. Leibmann, M. "Building Knowledge Management Solutions: A Way to KM Solutions (Technical Deployment Guide), <http://www.microsoft.com/solutions/km/KMSols.htm>, Microsoft Corporation, 2000
18. Malhotra, Y. The Brint.com Knowledge Management Portal. <http://www.brint.com/km/>, 2000
19. Microsoft Corporation. "Digital dashboard Overview", Microsoft Solutions, <http://agent.microsoft.com/solutions/km/Ddoverview.html>, 2000



20. Microsoft Corporation (a). "The Microsoft Knowledge Management Strategy- Practicing Knowledge Management: Turning Experience and Information into Results (White Paper) <http://www.microsoft.com/solutions/km/KMpract.htm>, Microsoft Corporation, 2000
  21. Nayar, M. "A Framework for achieving information Integrity", IS Audit & Control Journal, Vol. II, 1996.
  22. O'Dell, C. "A Current Review of Knowledge Management Best Practice", Conference on Knowledge Management and the Transfer of Best Practices, Business Intelligence, London, December, 1996.
  23. Petrash, G. "Managing Knowledge Assets for Value", Knowledge -Based Leadership Conference, Linkage, Inc. Boston, October 1996
  24. Radding, Alan. "Executive Summary", Knowledge Management: Succeeding in the Information-Based Global Economy, CTR Corporation, SC, USA, 1998.
  25. Strong, D., Lee, Y. & Wang, R. "Data Quality in Context", Communications of the ACM, Vol. 40, (5), May 1997.
  26. Taft, D. "Stopping knowledge overflow", Computer Reseller News, Manhasset, February 2000.
  27. Turban, E. Expert Systems and Applied Artificial Intelligence, Macmillan, 1992.
  28. Van der spek, R. & Spijkervet, A. "Knowledge Management: Dealing Intelligently with Knowledge", Knowledge Management and its Integrative Elements, Liebowitz & Wilcox, eds., CRC Press, 1997.
  29. Vernadat, F. "Enterprise Modeling and Integration: Principles and Applications", Chapman & Hall, London, 1996
  30. Wang, R. & Kon, H. "Toward Total Data Quality Management (TDQM), Sloan School of Management, MIT, <http://web.mit.edu/tdqm/papers/92>, June 1992
  31. Wang, R., Sotrey, V & Firth, C. "'A Framework for Analysis of Data Quality Research", IEEE Transactions on knowledge and Data Engineering, Vol. 7, (4), 1995
  32. Wiig, K. "Knowledge Management: Where did it come from and where will it go?" Expert Systems with Applications, Pergamon Press/Elsevier, Vol. 14, Fall 1997.
  33. Wizdom Software. "Introduction to Wizdom Software and Business Process Engineering", 1998
- 

### **Correspondence Addresses:**

*Dr. George Hodge, Dr. William Oxenham and Yatin Karpe*  
College of Textiles, North Carolina State University  
Raleigh, NC 27695-8301, USA

*Mr. Neil Cahill*  
Institute of Textile Technology, 2551 Ivy Road  
Charlottesville, VA 22903-4614, USA

# An assessment of the theory underpinning the role of information quality in the single-loop decision making model

Raul M. Abril, M. C. Sc. \*  
Teradata, division of NCR

**Abstract:** This paper offers comments on the single-loop decision-making model. The underpinning theoretical bases of such a model are assessed. The role of information quality is hypothesized in relation to customer relationship management decision making in the context of a data warehouse. The importance of learning aspects of decision making is explained. A focus group was set up in order to (1) validate the importance of our research problem from a management perspective (i.e. the interest of practitioners), (2) test the single-loop decision making model parsimony, and (3) get some clues about the potential constructs that would be of relevance in assessing decision performance. The results of this preliminary study support the single-loop decision-making model as a suitable framework for CRM decision making. Specific recommendations are made for further work on this piece of research.

## 1. Introduction

What is a good decision? [1]. Is it possible to help an inexperienced manager to detect problems? [2]. Is it possible to get some support when stating the facts or describing the situation? [3]. How can a manager be more confident in his/her decisions? [4]. We were interested in applying these questions in a given organization property and context. The organization property is the Customer Relationship Management, CRM in short, process and the organization context is the data warehouse of the firm. Therefore, our research problem consists in understanding the relationships (if any) between personal decision making, problem solving and information in a CRM process supported by a data warehouse. Although decisions are made at all levels in an organization [23] we will focus on decisions in CRM [24] at the individual level.

To the best knowledge of this researcher, the argument that decision effectiveness is a benefit provided by data warehouses has not been empirically validated. As such, our research question is: ***“What is the role of a data warehouse’s information quality in predicting decision effectiveness in CRM”***. This research question belongs to the MIS evaluation category of questions, more specifically to the decision support systems evaluation category of questions which is concerned with “the dependent variable in MIS research” [25]. Dependent variable here means MIS success. Our research is of theory-testing type; therefore empirical research guidelines should be rigorously applied [38]. We have selected the positivist paradigm passive observation for our research methodology. Remenyi et al. [38] recommend the following steps for this type of research: Literature review, assessment of established theoretical frameworks, assessment of ground theory in case of weak theoretical basis, theoretical conjecture and

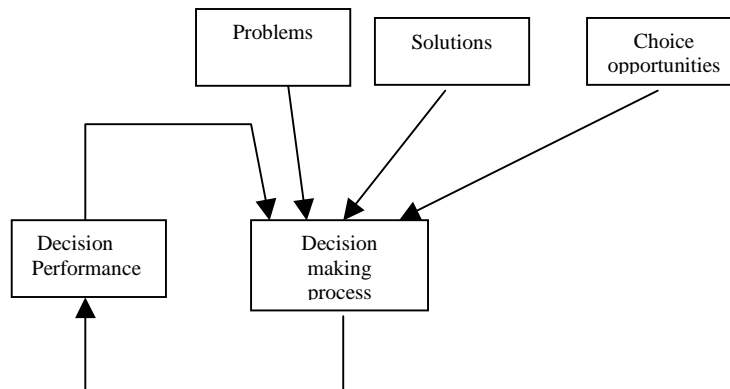
---

\* I am grateful for the comments provided by Dr. Robert M. O’Keefe, Dr. Joe F. Hair, Jr., and the reviewers of this paper.

hypotheses formulation, evidence collection design, primary and control evidence, testing and analysis, confirmation of theory and development of further/refined theory. As the title suggests, we report our findings after having completed the hypotheses formulation step, including the results of an exploratory research. Specific recommendations are made for working further on this piece of research.

## 2. The theoretical conjecture

The theoretical background for our research consists of the following two widely accepted theories: (1) The theory of decision making of Cohen, March and Olsen [66], and (2) the theory of problem solving behavior of Simon [5]. The theory of decision making developed by Cohen, March and Olsen [66] describes the links between input exogenous variables, the decision making process (named ‘garbage can’ process) and the output variable (i.e. a decision). Later March [59] emphasized the idea of the impact of preceding decisions, see Figure 1. This theory is applicable for both organizational and managerial decision making [9]. Because we are considering this theory only at the individual level, the exogenous variable “participants” [66] is implicitly considered in each of the decisional activities in the decision making process. In this theory there are the following five links of interest to us: *Decision performance* → *Decision making process*, *Decision making process* → *Decision performance*, *Problems* → *Decision making process*, *Solutions* → *Decision making process*, and *Choice opportunities* → *Decision making process*.



**Figure 1.** Cohen, March and Olsen’s ‘garbage can’ decision making process.

As Butler commented [9] in relation to this theory “a decision is constrained by the performance of preceding decisions and will, in turn, affect succeeding decisions by its own performance”. Learning is an important aspect in decision making. People learn ways of interpreting and dealing with situations (tasks, problems, and other conditions) by interacting with them [67]. Butler [68] found two dimensions of decision effectiveness. The first, like Marakas in [1], objective attainment, which is the extent to which prior objectives are reached. The second is the extent to which a decision and its associated processes lead to learning. This concept of learning is supported by the feedback of preceding decision performance in the garbage can decision making process.

Decision-making occurs under uncertainty and, potentially, involves subdecisions and concurrent decisions [9]. This theory links decision making to a time frame for the different streams of

decision making activities [9]. The contribution of this theory is twofold: (1) it gives an explanation of the [external view of] managerial decision making process, and (2) it gives a [controversial] explanation of the internal view of a managerial decision making process where this process is seen as a ‘garbage can’. This internal view of the decision making process (i.e. the garbage can decision making process), in simple terms, tells us that *per se* there is not necessarily a causal relationship between problems and solutions. The streams of decision-making activities include detecting new problems, compiling old problems, solutions to old problems and solutions to [potential] future problems. Decision makers “throw” the compiled problems and solutions into the garbage can [their personal ‘repository’] awaiting a choice opportunity (e.g. money spent, responsibilities allocated) when he/she will apply complex mental constructs (e.g. analogies) in order to, for example, map one or several of the available solutions to the problem in a given situation. This model does not prevent the fact that a decision-maker creates an *ad-hoc* solution to a new type of problem or even to an old type of problem. The sequence of streams of activities is entirely up to the decision-maker. Information is central to the process of coping with uncertainty and hence many studies of decision making have investigated the use of information [9]. Also, the fact that, as the garbage can process suggests, decision makers may go about looking for situations where their solutions and problems “fit” – where they make sense- simply means that there are situations that are manageable in terms of the available solutions. It implies nothing about the utility of other solutions [39].

Cohen, March and Olsen [with their input-garbage can process-output model, constructs and causal relationships] created a theory [70] of managerial decision making that has either support or no objections by other researchers [9] [39] [71] [72] [73] [74].

Decision-making can be thought of as a special type of problem solving [1]. Nobel prize-winning scholar Herbert A. Simon stated in his seminal work [5] and other subsequent works [26][51][52][53][54] a theory on decision making behavior contending that (1) as a result of cognitive constraints -i.e. cognitive limits to rationality- and uncertainty, characterized by scarcity of information, a person makes decisions with bounded rationality, (2) unprogrammed decisions will tend to involve problematic searches, with the need for alternatives prompted by a crisis or the availability of a solution and the use of ‘satisficing’ criteria to make a choice, (3) decision making consists of a general three-stage approach for problem-solving including intelligence, design, and choice where choice is a judgmental activity, and (4) in each of the problem solving activities humans are considered information-processing systems [51]. Massey [50] provides additional explanation to these three stages in terms of the following decisional activities: problem identification, problem definition encompassing problem structuring and problem formulation, solution building (named “alternative generation” by Massey), choosing (named evaluation and selection by Massey), and implementing.

Newel and Simon’s theoretical further development on the problem solving behavior theory contends that a problem solver, during the problem-solving process, (1) gets input information on the situation (e.g. a problem, solving task, choice opportunity) and makes an internal representation, (2) makes a problem space where there is information about the problem and about the solution, (3) problem solving takes place within such problem space, and (4) the task instructions and previous experience in solving similar tasks contribute significantly to the determination of the problem space. This view of a problem solver as an information-processing system fits with the garbage can process without entering into contradiction. The bounded rationality pattern of choice is individualistic and therefore judgmental, and context is more

important as a source of information for an interpretation of the decision task, as well as for its performance [39]. Dery [3] comments on this element of Simon's theory arguing that managers do not "normally face a choice situation, but events that call for evaluation and interpretation and that problems do not present themselves as structured or ill-structured, nor do they come as decision problems, complex or simple....To state the facts or describe the situation is to interpret not to copy it. Structured decision problems are structured because we choose to treat them as such". Simon's theory of problem solving behavior is applied to most models of management decision-making [1]. In addition, the literature addresses the application of the problem-solving process in marketing [47][48].

### 3. Research question and research model

The *problems, solutions and choice opportunities* variables of the garbage can process are [mental] internal representations of the decision maker about situations made after forming information, which means that a construct *Information on situations* that describes such a variables can be used instead. The following three constructs and links in the two referred theories are of interest for our research problem:

- *Information on situations* is the interpretation of problems, solutions and choice opportunities that stem from a specific situation that a decision-maker assigns to data by means of the known conventions used in their representation [60].
- *Decision performance* is (1) the extent to which prior objectives that gave rise to the need for a decision were reached within the boundaries and constraints imposed by the problem's context, and (2) the extent to which a decision and its associated processes lead to learning [9].
- *Decision making process* is the streams of activities that a decision-maker follows in order to cope with uncertainty over time under conditions of bounded rationality [9]. Such streams of decisional activities are problem identification, problem definition encompassing problem structuring and problem formulation, solution building, choosing, and implementing [50].
- The links are: *Information on situations* → *Decision making process*, *Decision performance* → *Decision making process*, and *Decision making process* → *Decision performance*.

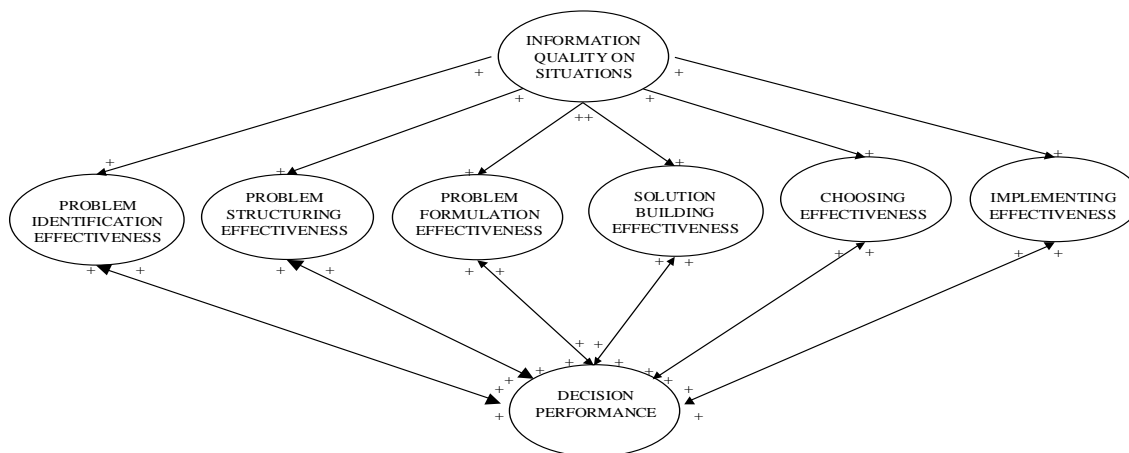
Our research question is: "***What is the role of a data warehouse's information quality in predicting decision effectiveness in CRM***". Our unit of analysis is the organizational member working in marketing units performing CRM processes and making decisions. We choose quality as the aspect of *information on situations* and effectiveness as the aspect of the *decision making process* that we will address respectively. We focus on the following constructs:

- *Information quality on situations* (IQ) is the extent to which the data warehouse enables the formation of information on situations that is fit for use by decision-makers in CRM. This definition is a specification of information quality as defined by Huang and colleagues [77]. *Information quality on situations* is an independent variable in our research model.
- *Decision performance* (D) as defined before. *Decision performance* is both a dependent and an independent variable in our research model giving sense to the 'single-loop' denomination.

- *Decision making process effectiveness* (DM) is the extent to which the decisional activities have been performed with effectiveness. That is *problem identification effectiveness*, *problem structuring effectiveness*, *problem formulation effectiveness*, *solution building effectiveness*, *choosing effectiveness*, and *implementing effectiveness* are dependent variables in our research model. This construct can be described as follows  $DM = \beta_1PI + \beta_2PS + \beta_3PF + \beta_4SB + \beta_5C + \beta_6Imp$  where *Problem identification effectiveness* (PI) is the extent to which a decision maker perceives symptoms that indicate or anticipate the presence of problems [1], *Problem structuring effectiveness* (PS) is the extent to which a decision maker collects the elements or variables needed to define the problem accurately [50], *Problem formulation effectiveness* (PF) is the extent to which a decision maker identifies and explores the relationships between these variables in order to define the problem [50], *Solution building effectiveness* (SB) is the extent to which a decision maker identifies and plans solution alternatives within the boundaries and constraints imposed by the problem's context, *Choosing effectiveness* (C) is the extent to which a decision maker chooses an acceptable solution plan from a selected set of alternatives, and *Implementing effectiveness* (Imp) is the extent to which a solution plan is realized.

The single-loop decision making model, see figure 2, has been derived from the theory of Cohen, March and Olsen about garbage can decision making and Simon's theory of problem solving considering the aspects of our research question. The denomination 'single-loop' applied to decision making is inspired by the 'single-loop model of learning' denomination [56]. Argyris introduced the 'single-loop' and 'double-loop' models of learning as crucial processes before the decision is made. Argyris explicitly argued the positive impact of such models in relation to the 'garbage can' model. Kolb [78] presented an overlay relating managerial learning styles to problem solving in order to illustrate the similarities and how learning styles affect problem solving success. Kolb concluded that problem solving and learning represent "...the same basic process of adaptation viewed from different perspectives".

There are studies reporting the following factors that can individually or collectively determine the relative difficulty of a pending decision: structure [5] [6], uncertainty [1][7], risk [8], alternatives and multiple objectives [1][9], cognitive limitations [10], contextual limitations [11] [12] [13] [14] environmental limitations [15] [16] [17] [18] and psychological limitations [19] [20] [21] [22]. We have to conclude that all of them are relevant with respect to their impact on managerial decision making. Recognizing the limited specification of the single-loop model to the constructs that we have explained we argue that this model is complete. From our literature review we have not identified any other independent variable, with strong empirical support, that might cause the dependent variables in the single-loop decision-making model.



**Figure 2. The single-loop decision making model.**

With that, our research question can be addressed by testing the following three hypotheses:

Hypothesis 1:  $IQ \rightarrow DM$ . The greater the *information quality on situations* derived from the data warehouse, the greater the *decision making process effectiveness* in CRM decision-making.

Hypothesis 2:  $D \rightarrow DM$ . The greater the *decision performance* in CRM decision making, the greater the *decision making process effectiveness* in CRM decision-making.

Hypothesis 3:  $DM \rightarrow D$ . The greater the *decision-making process effectiveness* in CRM decision-making, the greater the *decision performance* in CRM decision making.

#### 4. Theory assessment

The relevance of the theory of decision making of Cohen, March and Olsen for our research problem is that it gives us (1) a set of links between the variables *decision performance*, *problems*, *solutions*, *choice opportunities* and *decision making process*, and (2) a context of managerial decision making as described by the input-garbage can process-output approach. The relevance of Simon's theory of problem solving to our research is that (1) the *problems*, *solutions* and *choice opportunities* variables of the garbage can process are [mental] internal representations of the decision maker of situations made after forming information, which means that a construct *Information on situations* that describes such a variable can be used instead, (2) the theory provides further detail on the streams of decision making activities in the garbage can process, and (3) scarcity of information and/or cognitive limitations condition the variable *decision making process*.

We see in the literature support for the theory of Cohen, March and Olsen about garbage can decision making together with Simon's theory of problem solving. However, in order to address potential criticisms of this theoretical framework for the single-loop decision-making model as lacking in robustness, we wanted to find evidence in the literature supporting its links.

#### ➤ Hypothesis 1: $IQ \rightarrow DM$

The assertion that information quality is an antecedent of decision performance is common sense and has been the object of many studies (e.g. [30][40][42][43][44][45]). Particularly interesting

for our research are the effects on decision performance of i) integrated data [46] which is a characteristic of data warehouses [31], ii) data accuracy [58], completeness [58][84], and consistency [58], which are critical aspects of data quality in a data warehouse [33] [34], iii) information presentation [69], iv) data load [87], and v) information load [86]. In summary, information reduces uncertainty [75]. However, there are not many field studies giving evidence of a link with the decision making process. From the efficiency point of view, there is evidence supporting that i) knowledge in the problem domain will reduce the time spent on pre-decisional activities [28], ii) integrated data results in faster decision making [46], and iii) information load impacts decision time [86]. From the effectiveness point of view, there is evidence that i) lack of required information is negatively related to procedural rationality [57], ii) information framing has biasing effects on the decision making process [20], iii) information sources providing information of higher perceived quality will be used more frequently than will be those of lower quality [87], and iv) information sources that are more accessible will be used more frequently than will be those that are less accessible, of lower quality. Berthon et al. [85] have found that marketing managers who operate in organizations with more extensively developed repositories of relevant information including rules, policies, etc. will perceive a decision making context composed of higher proportion of structured than unstructured problems. Therefore, although we can conclude that there is enough evidence to support this link, we cannot conclude this at decisional activity level.

➤ **Hypothesis 2: D → DM**

The assertion that [preceding] decision performance is an antecedent of the decision making process either from the effectiveness or efficiency point of view, is postulated in the literature by studies addressing, among other topics, control as a behavioral strategy, learning, experience, and commitment escalation. However, few empirical studies addressing this single-loop, to the best knowledge of this researcher, have been published. The empirical studies supporting this single-loop link that this researcher is aware of had the following focus: i) Mintzberg et al. [27] collected data concerning the extent to which decisional activities were performed in twenty-five decision cases and emphasized the essential circularity of decision-making, and ii) feedback on decision accuracy leads to more normative-like processing of information and improved performance [29].

➤ **Hypothesis 3: DM → D**

Oz et al. [36] state that there are two dominant schools of thought on good decision making: one emphasizes the process, the other the outcome. The assertion that process is an antecedent of process' outcome is overwhelmingly accepted and the central subject in the quality literature. However, in relation to our research problem, there is limited empirical evidence supporting this link. Some remarkable findings from an effectiveness perspective are: i) a match between the information emphasized by the problem-solving tool and by the decisional activity results in superior problem solving [55], ii) comprehensiveness of the decision making process is negatively related to performance in an unstable industry [61] and positively related in a stable industry [76], iii) procedural rationality is positively related to strategic decision effectiveness [57], managers who collect information and use analytical techniques make decisions that are more effective than those who do not [57], quality of implementation is positively related to strategic decision effectiveness [57], iv) computation (e.g. internal rate of return), as a decision making strategy, is a necessary condition for effective investment decisions [68]. From the



efficiency perspective, it is known that time pressure impacts negatively decision effectiveness [83] and decision confidence [81].

The marketing literature has highlighted the significance of CRM. Kotler [62] states that the seller who knows how to build and manage strong relationships with key customers will have plenty of future sales from these customers. Reichheld and Sasser [63] state that companies might boost profits by almost 100% by retaining just 5% of their customers. Companies earn a higher return from getting repeat sales from current customers than from spending money to attract new customers [64]. Yet the nature of marketing strategy implementation and the reason for its success or failure are poorly understood. Furthermore, little is known about the factors influencing managers vested with implementation responsibilities [65]. There are not many empirical works addressing our research problem [65], which from the theory validation point of view represents a good opportunity for research [35]. Therefore, decision performance in CRM decision making becomes a key issue. With respect to IT investments in the marketing function (e.g. data warehouses), Beaumont [41] stated that there is no relationship between the scale of the investment and its benefits. The payoffs are dependent on the quality of the management of the systems and databases rather than the quality of the investment [41]. Industry expectations are that the overall data warehousing market will experience robust growth, at the compound annual growth rate (CAGR) of 28%, through 2004. The CRM-centric data-warehousing segment is expected to have an even higher CAGR of 37%, growing from \$4.2 billion in 1999 to \$20.1 billion by 2004. Consequently, our research question and the (dis)confirmation of our research hypotheses represent a required step in explaining decisional aspects for the success or failure in CRM processes. This has been so far neglected in the literature.

## 5. Exploratory research

Paradoxically with respect to the academic and economic relevance of data warehouses, industry surveys show that few organizations are measuring tangible or quantifiable returns from their data warehouses. Of those that are, reported returns have been modest. We wanted (1) to validate our perception of the importance of our research problem from a management perspective, (2) test the single-loop decision making model parsimony, and (3) get some clues about the potential constructs that would be relevant when assessing decision performance. Therefore, we conducted an exploratory field intervention by arranging a focus group following the recommended guidelines of Stewart and Shamdasani [82]. This study was conducted during the annual international conference that the user community of NCR's data warehouses organizes addressing data warehouse and CRM subjects. Details of this exploratory research are included in appendix A.

In general, the results from our exploratory research support the managerial significance of our research problem of understanding the relationships (if any) between personal decision making, problem solving and information in a CRM process supported by a data warehouse. As a result of this exploratory research, we found managerial support for a potential extension of our research problem including personal variables such as CRM knowledge competence, and information management competence. Our focus group research, see results #7 and #8 in appendix A, supports the inclusion of *CRM knowledge competence* and *information management competence* as independent variables in the specification of the single-loop decision making model. In order to address this potential weakness in the model (i.e. a potential specification

error [32] missing a critical predictor variable), we have tried to find empirical support for their inclusion. Considering, (1) the lack of strong empirical support for these constructs as predictors of CRM decision making process performance, and (2) our intention of looking for model parsimony [32] avoiding inserting variables indiscriminately [32], we decided not to include these two variables. In relation to the potential constructs that would be of relevance in assessing decision performance, results #2 to #5 (see appendix A) show that organizational performance type of measures is the preferred construct, with decision performance measures being the second preferred type of measures.

## **6. Next steps: Research methodology, operationalization and analysis**

### **Research methodology**

The next steps in our research require (dis)confirmation of our research hypotheses. Our research is of theory-testing type; therefore empirical research guidelines should be rigorously applied [38]. We have selected the positivist paradigm for our research methodology. Field study is our choice of [positivistic] research method. Our choice is based on (1) our literature review, (2) the purpose of our study, and (3) the nature of our research question as explained below.

- Field studies, mainly using a survey questionnaire, have become an increasingly common way of investigating decision making as a way of overcoming lack of generalizability of single cases, and the lack of a real-life feel of laboratory experiments (e.g. with students as surrogates of managers). Some studies create types of decision out of patterns in the associated processes and draw conclusions as to the likely conditions to which each type is best-suited [27] [49] [68]. Other studies test the causal relationship between variables to either create ground theory or support existing theory [30] [36] [40] [42] [43] [44] [45] [46] [58]. Therefore, we find that our choice of large-scale survey has strong support in the literature for our category of research questions.
- The purpose of our research is explanatory and there is a good match with field study as stated in [79]. Also, research will be nomothetic in nature in that it is studying general laws and finding empirical evidence from the research findings supporting such laws [38]. This aspect (i.e. nomothetic) requires evidence collection in such a way that we can generalize research results.
- Our research question is concerned with antecedents of decision effectiveness, which require a quantitative technique to produce quantitative descriptions of the predictors [80].

Therefore, field study is an appropriate choice as our positivist research method.

### **Operationalization**

DeLone's framework [25] is a helpful way of categorizing measures according to ways of assessing information system success. The categories of measures that we find applicable to the single-loop model are IS output quality, user satisfaction, individual impact and organizational impact.

	Technical		Semantic			Effectiveness or Influence	
	← Level	↔ Level	↔ Level	→ Level →			
<b>Variables in our research model</b>	System Quality	IS Output Quality	IS Output Use	User Satisfaction	Individual Impact	Organizational Impact	
- Information quality on situations		✓		✓			
- Decision performance					✓	✓	
- Decision making process effectiveness					✓		

Figure 3. Potential categories of measures applicable to the constructs in the single-loop model.

### Testing and analysis

The [theoretical] relationships X-Y in our research model and hypothesized in our research hypotheses could be (dis)confirmed as a logical conclusion of the following two types of validations: Construct validity and causal validity (empirically established). Causation requires a sufficient degree of correlation between the two variables, that one variable is the outcome of the other, and that there are no other reasonable causes for the outcome [32]. In addition, it is required to demonstrate that there are no spurious relationships. The traditional procedure for hypothesis testing should be used. Type I and Type II errors should be avoided. Structural equation modeling is particularly useful when (1) one dependent variable becomes an independent variable in subsequent relationships, and (2) we need to examine a series of causal relationships simultaneously [32].

### 7. Discussion

From our assessment of the theory underpinning the role of information quality in the single-loop decision making model we conclude that (1) there is enough evidence in the literature supporting information quality on situations as an antecedent of decision making process effectiveness but the role that information quality plays in each of the decisional activities is still not well understood, (2) there is very limited evidence supporting decision performance as an antecedent of decision making process effectiveness and with little detail at decisional activity level, (3) there is very limited evidence supporting decision making process effectiveness as an antecedent of decision performance, and (4) our research question is neglected in the literature. We argue that the single-loop decision making model that we present in this paper represents an innovative and integrative framework suitable for CRM research (i.e. our research question) on the role of a data warehouse’s information quality in predicting decision effectiveness. This can be addressed by testing three hypotheses that we have formulated based on the well respected theory of Cohen, March and Olsen about garbage can decision making and Simon’s theory of problem solving. Recommendations are made about the research methodology (i.e. field study, using a survey questionnaire), operationalization (i.e. DeLone’s framework), testing and analysis (e.g. SEM) that best fit the requirements of the (dis)confirmation of such hypotheses. The single-loop decision-making model supports Beaumont’s point regarding investment in IT for marketing. This author stated that (1) databases and systems do not possess value from their existence per

se, but from their application in different decision making domains, and (2) value should be measured against the productivity of management as it feeds into their decision making.

A remarkable characteristic of the single-loop decision making model is the interesting possibility that the variable *decision performance* as a dependent and independent variable offers for considering decision calibration [37] (i.e. decision confidence vs. decision accuracy).

## **Appendix A: Focus group**

### **Description**

Session Title: Data warehouses in production. Focus group  
 Context: Partners 2000 conference. Orlando (FL)  
 Date, time, place: Tuesday, September 26, 2000, 4:30-5:50 P.M., Europe 6 (Dolphin Hotel)  
 Facilitator/Moderator: Raul M. Abril  
 Participants: Retail: 1 (Denmark)  
 Financial : 1 (Israel), 1 (Netherlands), 1(Argentina)  
 Telecommunications: 1 (Spain), 1-fix line- (Austria), 1-cellular- (Austria)

### **Method**

Potential barriers to open communication were avoided by limiting the group to one participant per industry and country. The fix line provider and the cellular provider of Austria did not regard each other as a competitor. Two invited retail firms were not able to attend. The qualification criteria for participating was: (1) More than one year of data warehouse in production, (2) Responsibility for the usage of the data warehouse either as a user or as a service function. Five questions were sent in advance to the participants. Clarifications were offered over the phone. Answers were provided in a round table discussion with open discussion after each question/round. The questions were: What type of measures do you have for the value contribution of your Data Warehouse?, What measures would you recommend for the value contribution of your Data Warehouse?, What kind of barriers do you find in promoting usage?, What type of queries do you have?, and regarding data warehouse planning, give details about the horizon and kind of financials. The facilitator had a questionnaire ready for recording the answers. The group answers were e-mailed one week later to the participants asking for confirmation. After active follow up we had 3 confirmations out of seven without changing the initial answers. We did not have any (dis)confirmation from the other four.

### **Focus group results**

Legends:

**IQ** information quality on situations      **D** decision performance      **DM** decision making process effectiveness      **DWH** data warehouse

1. Most of the participants have measures for the value contribution of their DWH.
2. **IQ** and organizational financial performance are the measures used more frequently.
3. **D**-measures were reported as the second used more frequently measure by three participants.
4. Organizational financial performance measures were recommended by the majority (five) of the participants for evaluating a DWH.

5. **D**-measures were recommended by three of the participants for evaluating a DWH.
6. **IQ** was recommended by three of the participants for evaluating a DWH.
7. The majority (five) of the participants considered information management competence as a barrier to promoting usage of the information derived from their DWH.
8. Knowledge about the business process (e.g. CRM) was considered by three participants to be the second barrier to promoting usage of the information derived from their DWH.
9. Only one participant considered [lack of] experience as a barrier to promoting usage of the information derived from their DWH.
10. Most of the participants estimated that (1) between 75% and 90% of the queries are of a “What happened” nature (2) between 5% and 20% of the queries are of a “Why did it happen?” nature, and (3) between 0% and 5% of the queries are of a “What will happen?” nature. Two participants reported that they did not have any queries of a “What will happen?” nature.
11. Most of the participants (six) reported planning activities for their DWH with a planning horizon of one to three years based on a budget. All of them reported that they do not make estimates about ROI. They expressed an interest in better understanding end user training expenses.
12. Almost none of the participants expressed an interest in problem complexity, and problem solving skills.

### **Focus group discussion**

Result #11 is consistent with industry surveys in relation to the lack of ROI measures. This result would explain the inconsistency between result #4 (i.e. desired status) and result #2 (i.e. reality). We conclude that the participants experience difficulties implementing organizational performance measures for evaluating the contribution of their data warehouses. Results #2 and #6 reveal a consistent interest by the participants in **IQ** as a factor impacting the contribution of their data warehouses. Therefore, we find practitioners support the inclusion of **IQ** in our research problem. Results #3 and #5 reveal a consistent interest by the participants in decision performance **D** as a factor impacting the contribution of their data warehouses. Therefore, we find practitioners support the inclusion of **D** in our research problem. We only can explain result #12 as a consequence of the utilization (i.e. type of queries) that the participants reported in result #10 indicating that most of the queries are of a reporting nature with low predictive queries. We did not find practitioners support the inclusion of problem solving skills as part of our research problem. However, given that (1) the literature reports that decision making (**DM**) can be thought of as a special type of problem solving, and (2) the application of Simon’s problem solving steps is done for most models of management decision making, then we find theoretical and empirical justification for including decision making (**DM**) as part of our research problem. Results #8 and #7 indicate that lack of competence in information management and lack of knowledge of the business process (e.g. CRM) are barriers to promoting usage of the information derived from their data warehouses. As reported in the literature, for example see [45] in relation indirect users, result #9 indicates that, in the opinion of the practitioners, experience (e.g. number of years) of the business process is not a barrier to

promoting usage of the information derived from their data warehouses. Now, this has to be interpreted with care because the practitioners might have meant that they do not find such barrier because most of their decision-makers actually have experience in the business process.

### References:

- 1 Marakas, G.M. *Decision support systems in the 21st century* (Prentice-Hall, 1998)
- 2 Marples, D. "Studies of managers." *Journal of Management Studies* vol. 4, no. 3, (1967) pp. 282-299
- 3 Dery, D. "Decision-making, problem-solving, and organizational learning." *OMEGA* vol. 11, (1983) pp. 321-328
- 4 Aldag, R.J., Power, D. J. "An empirical assessment of computer-assisted decision analysis." *Decision Sciences* vol. 17, no. 4, (1986) pp. 572-588
- 5 Simon, H.A. *The new science of management decision* (Harper and Row, 1960)
- 6 Gorry, G.A., Scott Morton, M.S. "A framework for MIS." *Sloan Management Review* vol. 13, no. 1, (1971) pp. 76-88
- 7 Thompson, J. D. *Organizations in action* (McGraw-Hill, 1967)
- 8 Morris, P. W. G., Haugh, G. H. *The Anatomy of major projects: a study of the reality of project management* (Wiley, 1987)
- 9 Butler, R. "Decision making." in *International Encyclopedia of Business and Management International Encyclopedia of Business and Management* pp. 951-1001 edited by Warner, M. , 1996)
- 10 Miller, G. A. "The magical number seven, plus or minus two: Some limits to our capability for processing information." *Psychological Review* vol. 63, no. 2, (1956) pp. 81-97
- 11 Wang, R. Y., Reddy, M. P., Kon, H.B. "Toward quality data: An attribute-based approach." *Decision Support Systems* vol. 13, no. 3/4, (1995) pp. 349-372
- 12 Redman, T. C. "Improve data quality for competitive advantage." *Sloan Management Review* vol. 36, no. 2, (1995) pp. 99-107
- 13 Raghunathan, S. "Impact of information quality and decision-maker quality on decision quality: a theoretical model and simulation analysis." *Decision Support Systems* vol. 26, no. 4, (1999) pp. 275-286
- 14 Brockhoff, K. "Forecasting quality and information." *Journal of Forecasting* vol. 3, no. 4, (1984) pp. 417-428
- 15 Mintzberg, H. *Structure in fives: Designing effective organizations* (Prentice-Hall, 1983)
- 16 Witteloostuijn, A.V. "Contexts and environments." in *International Encyclopedia of Business and Management* pp. 752-61 edited by Warner , 1996)
- 17 Emery, F. E., Trist, E. L. "The causal texture of organizational environments." *Human Relations* vol. 18, (1965) pp. 21-32
- 18 Miles, R. E., Snow, C. C. *Organizational strategy, structure and process* (McGraw-Hill, 1978)
- 19 Tversky, A., Kahneman, D. "The framing of decisions and the psychology of choice." *Science* no. 211, (1981) pp. 453-458
- 20 Mowen, M. M., Mowen, J. C. "An Empirical examination of the biasing effects of framing on business decisions." *Decision Sciences* vol. 17, no. 4, (1986) pp. 596-602
- 21 Kashima, Y., Maher, P. "Framing of decisions under ambiguity." *Journal of Behavioral Decision Making* vol. 8, no. 1, (1995) pp. 33-49

- 22 Huber, O., Debeutz, A., Pratscher, J., Quehenberger, I. "Perceived control in a multistage decision task." *Journal of Behavioral Decision Making* vol. 3, no. 2, (1990) pp. 1123-137
- 23 Anthony, R. N. *Planning and control systems: A framework for analysis* (Harvard University Graduate School of Business Administration, 1965)
- 24 Fletcher, K. *Marketing management and information technology* (Prentice-Hall, 1995)
- 25 DeLone, W. H., McLean, E. R. "Information systems success: The quest for the dependent variable." *Information Systems Research* vol. 3, no. 1, (1992) pp. 60-95
- 26 Simon, H. A. *Administrative behavior: A study of decision making process in administrative organizations* (The Free Press, 1976)
- 27 Mintzberg, H., Raisinghani, D., Théorêt, A. "The structure of unstructured decision processes." *Administrative Science Quarterly* vol. 21, no. 2, (1976) pp. 246-275
- 28 Day, D. V., Lord, R. G. "Expertise and problem categorization: The role of expert processing in organizational sense-making." *Journal of Management Studies* vol. 29, no. 1, (1992) pp. 35-47
- 29 Creyer, E. H., Ettman, J. R., Payne, J. W. "The impact of accuracy and effort feedback and goals on adaptive decision behavior." *Journal of Behavioral Decision Making* vol. 3, no. 11, (1990) pp. 1-16
- 30 Gelderman, M. "The relation between user satisfaction, usage of information systems and performance." *Information & Management* vol. 34, no. 1, (1998) pp. 11-18
- 31 Inmon, W. H. *Database machines and decision support systems. Third way processing* (QED Information Sciences, 1991)
- 32 Hair, J. F. J., Anderson, R. E., Tatham, R. L., Black, W. C. *Multivariate data analysis* (Prentice-Hall, 1998)
- 33 Lyon, J. "Customer data quality: Building the foundation for one-to-one customer relationship." *Journal of Data Warehousing* vol. 3, no. 2, (1998) pp. 38-47
- 34 Shanks, G., Darke, P. "A framework for understanding data quality." *Journal of Data Warehousing* vol. 3, no. 3, (1998) pp. 46-51
- 35 Day, G. S., Montgomery, D. B. "Charting new directions for marketing." *Journal of Marketing* vol. 63, (1999) pp. 3-13
- 36 Oz, E., Fedorowicz, J., Stapleton, T. "Improving quality, speed and confidence in decision-making: Measuring expert systems benefits." *Information & Management* vol. 24, no. 2, (1993) pp. 71-82
- 37 Chung, J. "Auditor's confidence and the audit expectation gap." *Australian Accountant* no. June, (1995) pp. 26-30
- 38 Remenyi, D., Williams, B., Money, A., Swartz, E. *Doing research in business and management. An introduction to process and method* (SAGE, 1998)
- 39 Bahl, H. C., Hunt, R. G. "Problem-Solving strategies for DSS design." *Information & Management* vol. 8, no. 2, (1985) pp. 81-88
- 40 Iivari, J., Ervasti, I. "User information satisfaction: IS implementability and effectiveness." *Information & Management* vol. 27, no. 4, (1994) pp. 205-220
- 41 Beaumont, J. R. "Information technology in marketing." in *International Encyclopedia of Business and Management* pp. 2118-24 edited by Warner, M. , 1996)
- 42 Seddon, P., Kiew, M.-Y. "A partial test and development of the DeLong and McLean model of IS success" pp. 99-110 *International conference on information systems*, 1994)
- 43 Igarria, M., Tan, M. "The consequences of information technology acceptance on subsequent individual performance." *Information & Management* vol. 32, no. 3, (1997) pp. 113-121

- 44 Etezadi-Amoli, J., Farhoomand, A. .F. "A structural model of end user computing satisfaction and user performance." *Information & Management* vol. 30, no. 2, (1996) pp. 65-73
- 45 Gatian, A.W. "Is user satisfaction a valid measure of system effectiveness?" *Information & Management* vol. 26, no. 3, (1994) pp. 119-131
- 46 Goodhue, D. L., Klein, B. D., March, S. T. "User evaluations of IS as surrogates for objective performance." *Information & Management* vol. 38, no. 2, (2000) pp. 87-101
- 47 Titus, P. A. "Marketing and the creative problem-solving process." *Journal of Marketing Education* vol. 22, no. 3, (2000) pp. 225-235
- 48 Wierenga, B., Bruggen, G. H. V. "The integration of marketing problem-solving modes and marketing management support systems." *Journal of Marketing* vol. 61, no. July, (1997) pp. 21-37
- 49 Hickson, D., Butler, R. J., Cray, D., Mallory, G., Wilson, D. C. *Top decisions: Strategic decision making in organizations* (Basil Blackwekk, 1986)
- 50 Massey, A. P., Clapper, D. L. "Element finding: The impact of a group support system on a crucial phase of sense making." *Journal of Management Information System* vol. 11, no. 4, (1995) pp. 149-176
- 51 Newell, A., Simon, H. A. *Human problem solving* (Prentice-Hall, 1972)
- 52 Simon, H. A., Newell, A. "Human problem solving: The state of the theory in 1970" *American Psychologist* vol. 26, (1971) pp. 145-159
- 53 Simon, H. A. *The shape of automation for men and management* (Harper and Row, 1965)
- 54 Simon, H. A. "The architecture of complexity" pp. 467-482 *American Philosophical Society*, 1962)
- 55 Agarwal, R., Sinha, A.P., Tanniru, M."Cognitive fit in requirements modeling: A study of object and process methodologies." *Journal of Management Information System* vol. 13, no. 2, (1996) pp. 137-162
- 56 Argyris, C."Single-loop and double-loop models in research on decision-making." *Administrative Science Quarterly* vol. 21, no. 3, (1976) pp. 363-375
- 57 Dean Jr., J.W., Sharfman, M.P."Does decision process matter? A study of strategic decision-making effectiveness." *Academy of Management Journal* vol. 39, no. 2, (1996) pp. 368-396
- 58 Wixom, B. H., Watson, H. J. "An empirical investigation of the factors affecting data warehousing success." *Management Information Systems Quarterly* vol. 25, no. 1, (2001) pp. 17-41
- 59 March, J. G. *Decisions and organizations* (Blackwell, 1988)
- 60 Everest, G.C. *Dabase Management. Objectives, system functions and administration* (McGraw-Hill, 1986)
- 61 Fredrickson, J.W., Mitchell, T.R."Strategic decision processes: Comprehensiveness and performance in an industry with an unstable environment." *Academy of Management Journal* vol. 27, no. 2, (1984) pp. 399-423
- 62 Kotler, P. *Marketing management. Analysis, planning, implementation, and control* (Prentice Hall International, 1988)
- 63 Reichheld, F. F., Sasser, W. E. "Zero defections: Quality comes to services." *Harvard Business Review* vol. 68, no. September-October, (1990) pp. 105-111
- 64 Shelth, J. N., Parvatiyar, A. "Relationship marketing in consumer markets: antecedents and consequences." *Journal of Academy of Marketing Science* vol. 23, no. 4, (1995) pp. 255-271



- 65 Noble, C. H., Mokwa, M. P. "Implementing marketing strategies: Developing and testing a managerial theory." *Journal of Marketing* vol. 63, no. 4, (1999) pp. 57-73
- 66 Cohen, M. D., March, J. G., Olsen, J. P. "A garbage can model of organizational choice." *Administrative Science Quarterly* vol. 17, (1972) pp. 1-25
- 67 Hunt, R. G. "Coping with institutional racism: A model for social problem solving." in *Producing useful knowledge for organizations* edited by Kilmann, R.H. et al. (Praeger, 1985)
- 68 Butler, R. J., Davies, L., Pike, R., Sharp, J. *Strategic investment decisions: Theory, practice and process* (Routledge, 1993)
- 69 Zmud, R.W., Blocher, E., Moffie, R.P. "The impact of color graphic report formats on decision performance and learning." pp. 179-193 *Fourth international conference on information systems*, 1983)
- 70 Weick, K. E. "What theory is not, theorizing is." *Administrative Science Quarterly* vol. 40, no. 3, (1995) pp. 385-390
- 71 Masuch, M., LaPoint, P. "Beyond garbage cans: An AI model of organizational choice." *Administrative Science Quarterly* vol. 34, no. 1, (1989) pp. 38-67
- 72 Padgett, J. F. "Managing garbage can hierarchies." *Administrative Science Quarterly* vol. 25, (1980) pp. 583-604
- 73 Carley, K. "Efficiency in a garbage can: implications for crisis management." in *Ambiguity and command* pp. 165-94 edited by March, J.G. and Weissinger-Baylon, R. (Pitman, 1986)
- 74 Anderson, P.A., Fischer, G. W. "A Monte Carlo model of a garbage can decision process." in *Ambiguity and command* pp. 140-64 edited by March, J.G. and Weissinger-Baylon, R. , 1986)
- 75 Gadenne, D., Iselin, E.R. "Properties of accounting and finance information and their effects on the performance of bankers and models in predicting company failure." *Journal of Business Finance & Accounting* vol. 27, no. 1, (2000) pp. 155-193
- 76 Fredrickson, J.W. "The comprehensiveness of strategic decision processes: Extension, observations, future directions." *Academy of Management Journal* vol. 27, no. 3, (1984) pp. 445-466
- 77 Huang, K.-T. , Lee, Y.W., Wang, R.Y. *Quality information and knowledge management*, 1998)
- 78 Kolb, D.A. "On management and the learning process." in *Organizational Psychology* pp. 27-42 edited by Kolb, D.A. et al. , 1974)
- 79 Marshall, C., Rossman, G. *Designing qualitative research* (SAGE, 1995)
- 80 Pinsonneault, A., Kraemer, K.L. "Survey research methodology in management information systems: An assessment." *Journal of Management Information System* vol. 10, no. 2, (1993) pp. 75-105
- 81 Smith, J.F., Mitchell, T.R., Beach, L.R. "A cost-benefit mechanism for selecting problem-solving strategie: Some extensions ans empirical tests." *Organizational behaviour and human performance* vol. 29, (1982) pp. 370-396
- 82 Stewart, D.W. "Focus group research." in *Handbook of applied social research methods* edited by Bickman, L. and Rog, D.J. , 1998)
- 83 Svenson, O., Edland, A., Slovic, P. "Choices and judgements of incompletely described decision alternatives under time pressure." *Acta Psychologica* vol. 75, (1990) pp. 153-169

- 84 Ahituv, N., Igarria, M., Sella, A. "The effects of time pressure and completeness of information on decision making." *Journal of Management Information System* vol. 15, no. 2, (1998) pp. 153-172
- 85 Berthon, P., Pitt, L.F., Ewing, M.T. "Corollaries of the collective: The influence of organizational culture and memory development on perceived decision-making context." *Journal of the Academy of Marketing Science* vol. 29, no. 2, (2001) pp. 135-150
- 86 Iselin, E.R. "The effects of information load and information diversity on decision quality in a structured decision task." *Accounting, Organizations and Society* vol. 13, no. 2, (1988) pp. 147-165
- 87 Iselin, E.R. "The effects of the information and data properties of financial ratios and statements on managerial decision quality." *Journal of Business Finance & Accounting* vol. 20, no. 2, (1993) pp. 249-266

**Information Envelope and its Information Integrity Implications:  
For a complex, changing environment, modeling a generic business process  
as an integral to a closed loop information and control system  
characterized by uncertainty**

Vijay V. Mandke  
Research Leader,  
Center for Information Integrity Research,  
Unitech Systems (India) Pvt. Ltd.,  
B-64 (First Floor), Gulmohar Park,  
New Delhi – 110049, India.  
E-mail: vijaymandke@satyam.net.in

Madhavan K. Nayar  
President,  
Unitech systems, Inc.,  
1240 E. Diehl Road, Suite 300,  
Naperville, Illinois 60563, USA.  
E-mail: mnayar@unitechsys.com

Kamna Malik  
Professor,  
Dept. of Information Technology,  
Institute of Management Technology,  
Post Box No. 137, Rajnagar,  
Ghaziabad - 201001, India.  
Kamnamalik@imt.ac.in

**Abstract:** Physical and informational works are strongly interrelated in a business process. This paper facilitates a control's interpretation of model of business process as an integral part of a closed loop information and control system. Various uncertainties (due to 5Cs) affect this model and raise the Information Integrity issue for the same. This complex information and control system delivers a flexible information decision to control business process in a changing environment. The paper argues that a more useful view of information decision is as a process of information gathering and processing rather than the conventional view of decision as 'choice between various alternatives'. This flexible decision process should have the stages viz. Operable goal setting, Definition of complexity criteria, Construction of opportunity and constraint spaces, Development of information structure dynamics model, Customized planning and design of alternatives and, finally, the Choice of alternative. The conclave of information bases for these information gathering and processing stages characterized by their respective contexts is normally not considered for business process IS modeling. It is this conclave that the paper defines as 'Information Envelope', and shows it to be central to the information and control system to which the business process is integral. And more importantly, the paper shows that it is each of these information bases that is affected by further uncertainty of the type not encountered earlier; thereby resulting in further loss of Information Integrity for a business process operating in a complex and changing environment.

## **1. Introduction**

The research investigations on Information Flow Model (IFM) for Integrity Analysis presented at IQ 1999 [5] studied the Information Integrity (I\*I) problem in the context of 'errors in networked computerized information systems that are made but not corrected.' The investigations categorized these error components in IS in terms of errors with deterministic descriptions caused by singular events like software failure, and errors with stochastic descriptions caused by general, judgmental, and systems factors. In the process, the investigation proposed a workable approach to developing IFM with capability for information accuracy, consistency and reliability, i.e. Integrity Analysis and Improvement Plan (IAIP), by viewing IFM in its totality. The total view of IFM includes Data Origin Stage activities, Data Conversion Stage activities and Output Stage activities, each of them having further sub-activities. Accordingly, for each of these sub-activities of a computerized information system, the investigation proposed IFM for achieving I\*I through IAIP implementation.

Though they provide a basis for generating error databases for implementing integrity analysis leading to integrity technologies discussed in literature [3], understandably, these information flow models constitute only core IS views for respective IS sub-stages for integrity analysis and improvement. This, in turn, calls for developing a systematic methodology for developing a structural model for gathering information to be processed by them and for studying the totality of I\*I implications as emerging consequently.

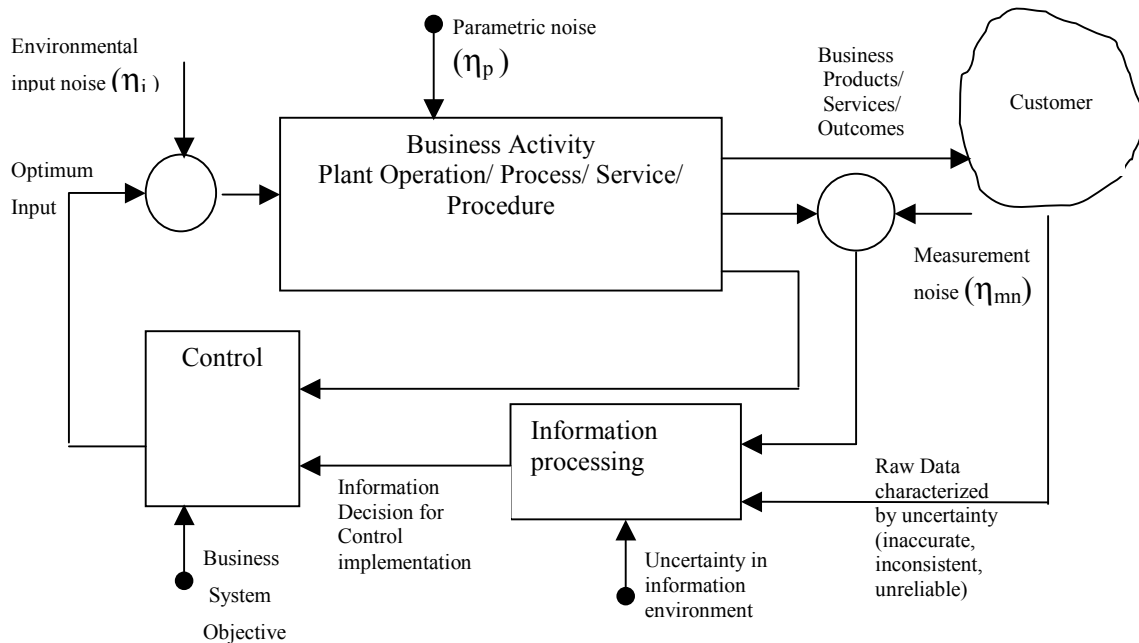
Traditionally, the business enterprise had computerized information systems (IS) developed in isolation, but there was no effort to optimize data or information for improved decision making. The requirement was in terms of automation of functions of 'hard' components, i.e. of 'mechanical' or 'physical' work, so as to add value to the product produced. However, with data-driven technologies keyed to the flow of digital data throughout an enterprise and on the Net and with pressures of achieving business objectives of effectiveness and efficiency, business enterprise has a further requirement for utilizing data/information decisions 'smarter' [6,12].

This calls for automation of 'informational work' carried out by the soft components of the enterprise wherein 'data' is seen as raw material, 'data processing or transformation or conversion' as the system function and 'data product' or 'information' as processed data used to trigger information use (decision making stage included) so as to deliver 'information decision' in the form of information to add value to the product [4,6].

This is an application of flexible automation accounting for product innovation, customer needs (product requirements) and constraints of costs and capabilities - a structural variant from inflexible automation. Specifically, the flexible automation is becoming possible due to (a) availability of on-line computers, (b) computers providing capability for moment by moment optimization of processes and decision-making, and (c) availability of system integration capability so as to yield a computer integrated system for attaining business objectives.

What makes it possible now to 'put it all together' in a total production, delivery and service system is technological reality of digital data as medium of information flow across the enterprise. Further, most importantly, such systems can be applied to both hard components of production like processes, machinery and equipment, and soft components like information flow and data bases --- the informational work systems [10,6].

It is within the above framework of interrelationship between informational and physical work systems and with reference to IFM as already mentioned, that it then becomes possible to reinforce often articulated proposition that whatever else a business does, it processes information [5]. For the purpose of the research investigation at hand, this business process IS view indeed is a very helpful observation. It suggests that system engineering techniques used in understanding the dynamics and responses of physical systems could, therefore, be used for understanding and predicting the operation and performance of more subjective and probabilistic description of business processes controlled by the requirements of flexible information decisions for control implementation [Figure (1)].



**Figure (1): A Business Process IS View - A systems representation of a generic business process as integral to a closed loop information and control system.**

In what follows, this paper addresses this research issue in the context of a generic business process. It may be mentioned that as the model emphasizes information, it is applicable to manufacturing, production, or service activity. As a result, choice of any specific activity to represent the business process is only illustrative; the conclusions drawn being applicable to all types of business activities. Further, for understanding the integrity implications, the business process IS view can be arranged as per the levels of controls applied. Figure (2) gives this information and control system based model of the business process IS view for an environment characterized by uncertainty along with its I\*I implications.

## 2. Uncertainty in Business Process IS view and its Integrity Implications

Due to the system environmental factors of 5 Cs [Change (C1), Complexity (C2), Communication (C3), Conversion (C4), and Corruption (C5)], this information and control system constituting business process IS view is characterized by uncertainty at various levels as described here [4,8]. Traditional systems, emphasizing individual production machines, exhibit the existence of uncertainty at plant operations level and first control level. At plant operation level, the uncertainty is in the input ( $\eta_i$ ), operations ( $\eta_p$ ), and output ( $\eta_{C1,2,4}$ ). At first level of control, the uncertainties are due to measurement or observation noise ( $\eta_{ob}$ ). Measurement error factors and uncertainty at the plant/ process operations ( $\eta_{C1,2,4}$ ) may render information observed at plant output to be inaccurate and incomplete, i. e., affected by measurement or observation noise.

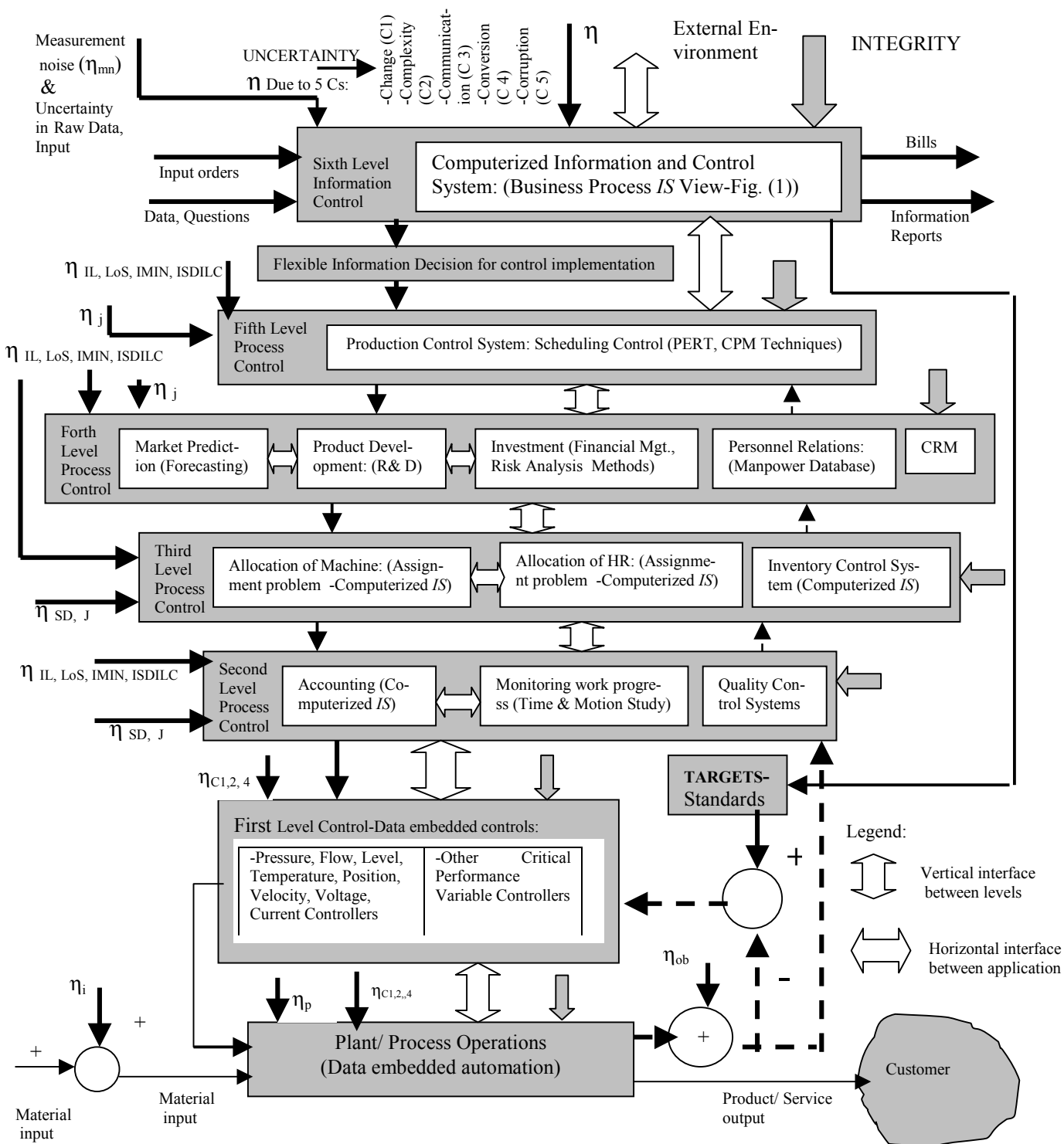


Figure (2): Business Process IS View Model describing a generic business process as integral to an information & control system for a business environment characterized by uncertainty and its Information Integrity Implications.

## 2.1. Uncertainty types newly emerged due to ‘application’ emphasis with system non-integration

With the advent of computer technology, further impetus for automation initiatives came in the form of higher-level process controls [Figure 2]. Specifically, these were ‘applications’ of computerized information systems justified (a) initially on the cost reduction aspects of processing structured and periodic information, the business work clerical in nature being the obvious choice, and (b) later as management tools for planning, direction and control [6,7]. Figure (2) shows different process control levels - higher than the first level control, with the feedback information from lower level control to the higher level, and the reference, i.e. feed-forward information, from higher level to lower level. For the reasons mentioned, businesses developed more and more of these ‘applications’, each with its own terminology, procedures and data sources giving rise to new uncertainties. Further at these higher levels, the human-machine interface is also prevalent. Within this framework, following uncertainties are identified.

- i) *Uncertainty types present at all process control levels ( $\eta_{IL, LoS, IMIN, ISDILC}$ ):*
  - a) Uncertainty due to information overload ( $\eta_{IL}$ ),
  - b) Uncertainty due to lack of standardization ( $\eta_{LoS}$ ),
  - c) Uncertainty due to lack of relationship between the data in several applications (problems arising from emphasis on integration minimization) ( $\eta_{IMIN}$ ),
  - d) Uncertainty due to errors in hardware, software, data entry, or accidental or intentional failures (including human failures, etc.), i.e., uncertainty due to errors in information system development and implementation life cycle ( $\eta_{ISDILC}$ ).
- ii) *Uncertainty types at process control levels 2 & 3:* The process control levels 2 & 3 deal with managerial decisions at middle level [Figure (2)]. In addition to types of uncertainties identified above, these levels are also characterized by uncertainty due to incomplete knowledge of system dynamics ( $\eta_{SD}$ ) and due to judgmental errors at human-IS interface ( $\eta_j$ ) [Section (1)]. These levels are characterized in much more rudimentary and uncertain way by the deterministic and stochastic models of linear and non-linear programming decisions as against the plant/process and first-level controls that can be fully described by deterministic model.
- iii) *Uncertainty types at process control levels 4 & 5:* The process control levels 4 & 5 deal with higher management level decisions [Figure (2)]. Understandably these levels are characterized by human-machine systems in which humans start playing dominant part in decision making. Particularly, the process controls at level 4 are often described by decision theory models, while process control level 5 which may comprise production and scheduling controls (planning control included) differs from conventional control in that it includes humans as part of the process to be controlled. All this adds to uncertainty at process control levels 4 and 5 ( $\eta_j$ ).
- iv) *Uncertainty type at information control level 6:* While automating (optimizing) production process with the help of five control levels as above put in operation in isolation, what has not been possible is to optimize design continually, i.e. in *on-line* fashion. This continuity is the basis for production line delivering mass-customized

products for continually changing business environment (product innovation included) with emphasis on integration maximization across the supply chain.

The technological reality of the sixth level information control makes this possible. Specifically, the sixth level control is a business process *IS* view, and it comprises human - machine systems. Thus, very little is understood about the physical structures governing the sub-systems and components of the sixth level control system. As a result this level is normally described by an inductive model which is developed based on observations made on the real-world business operations, and, as the problem is, these observations are invariably noisy. In other words, one is faced with the problem of implementing the sixth level information control, when the data available to develop the control model is characterized by uncertainty ( $\eta_{mn}$  and uncertainty in raw customer data) [Figure (1)].

## 2.2. Uncertainty types due to increased complexity

And even as there is an increased emphasis on ‘applications’ for competitive business advantage, microprocessors and data driven technology keyed to the flow of information across the enterprise have led to total shift toward system integration. Resulting reduction in information processing costs and the competitive advantage of the systems developed have further accelerated this shift [7]. Thus, on the one hand, one sees a dramatic increase in the use of computers in the form of ‘embedded systems’ over a widest range of systems [11]. On the other hand, the business enterprise has its goal shifted from that of ‘cost minimization’ to that of ‘financial optimization’. At every level all this has, understandably, led to use of components and systems complex in nature, thereby further adding to the uncertainty due to system integration as follows ( $\eta_{C1,2,3,4,5}$ ):

- i) *Uncertainty in plant operations*: Process failures may occur due to complex error mechanisms coming from design, manufacturing, commissioning and maintenance phases and acting with delay ( $\eta_{C1,2,4}$ ).
- ii) *Uncertainty in plant and first and higher level control operations due to failure of ‘embedded systems’*: Traditionally, hardware has been considered to be reliable. However, with embedded systems all this has changed. This failure can emerge due to inadequate tests undertaken; due to incompatibility between electrical components and maintenance errors. It is these failures of ‘embedded systems’ that then result in uncertainty in plant and first (and higher) level control operations ( $\eta_{C2}$ ).
- iii) *Uncertainty due to presence of system interfaces* ( $\eta_{C1,2,3,4,5}$ ): The system integration impacts all the six levels of controls as also the plant /process operation by introducing system interfaces [Figure (2)]. These interfaces call for the specification of each IS module to include details of its interaction with other modules. This interaction may be formalized in an interface design specification (IDS), which sets out the data or messages sent between modules, and any protocols used. As the levels of information and control system in Figure (2) interact laterally and vertically (not shown in full), it follows that modules that are internal will also have interfaces with modules at the boundary and, therefore, with external system and vice versa. In the wake of emphasis on system integration maximization, more often than not the resulting interactions will be complex, thereby introducing further uncertainty at all levels (plant operation inclusive).



### 2.3. Information Integrity Implications

The presence of uncertainties as above at all levels of information and control system leads to errors in business process IS view (that are made but not corrected in spite of application controls [5, 4]). This results in loss of integrity at the data processing stages, thereby, rendering data and information processed inaccurate, incomplete, not up to date, and unreliable. Figure (2) indicates at critical points the noise inputs discussed above, and acknowledges the presence of systems interfaces. For the purpose of presentational simplicity, the vertical interfaces between levels are shown in complete, while the lateral interfaces between applications at a level are shown nominally.

### 3. Improved Business Process IS View, consequent Uncertainties and Information Integrity Implications thereof for a Complex and Changing Environment

In the study of a business process operating in an environment characterized by uncertainty, the identification of an information and control model as in Figure (1) has made it possible to consider applying system engineering techniques to research the I\*I problem at hand. As shown in Figure (2), what one is dealing with is a multi-level control problem. Though a large system problem, on the face of it, the problem looks tractable. At the lowest level, one is concerned with plant operation and processes, which primarily comprise individual production machines, and microprocessor based data embedded systems that are describable by deterministic mathematical models. Similar is the case with first level control models. These models are deductive and are arrived at by having complete understanding of physical structure of the system and either by analytical consideration or by experiment. Such models are susceptible to being controlled in accordance with the principles of classical theory of automatic control.

As mentioned in Sub-sections [2.1-(ii) and (iii)], the controls at levels 2-5 are amenable to quantitative treatment through various models covered by systems engineering tools and techniques scanning a wide range of interdisciplinary areas. Thus, based on knowledge of system engineering tools and techniques, the Figure (2) problem of information and control system modeling of a business process, looks tractable even as one grapples up to 5<sup>th</sup> level control, which include product innovation, planning and design stage requirements.

However, methodological inadequacy creeps in as one deals with the information control at level 6. As pointed out in Section (1) and Sub-section [2.1-(iv)], now, it is possible to optimize design continually, i.e. in *on-line* fashion (continuous product innovation), as a basis for production line delivering mass-customized products and services for continually changing business environment. In other words, in order not to be blind-sided in rapidly changing markets, the search and relevant information decision must not be restricted to diagnostic routines and procedures ballistic in nature. Instead, senior management needs a measurement and decision system more like the one used by the national weather service. Ground stations all over the country monitor temperature, barometric pressure, relative humidity, cloud cover, wind direction and velocity, and precipitation. Balloons and satellites provide additional data. These are monitored continuously and fed to central location where they can be used to search for patterns of change. Based on these intelligence data, forecasts of impending conditions can be made or revised (flexible information decision) in the light of changing circumstances [9].

As mentioned in Section (1) and Sub-section [2.1-(iv)], in the form of information control at the 6<sup>th</sup> level, thus one has inductive modeling exercise at hand based on real world business observations. This exercise (a) involves multiple goals, many factors, and a large number of interdependent information variables, varying with time, and not completely and correctly observable, and (b) its system dynamics is not well understood. This is a complex problem solving exercise, and significantly the complexity is not of the *order*, but of the *organization* [1,2]. In other words, the information control at 6<sup>th</sup> level involves processing of unstructured (maximal) information as against structured (minimal) information as has been the case up to 5<sup>th</sup> level control [6, 9]. To that extent the 6<sup>th</sup> level information control dramatically distinguishes itself from information processed at lower levels by mainly acquiring an open system character. In fact when system integration is complete, all levels acquire open system character, the degree of openness being directly proportional to the order of the level; and in the process the system at its all levels assumes a high degree of complexity.

One of the unique properties of an open system is it has a purpose or goal or direction. As a result, activities of continual operable goal setting and implementing so as to deliver correct action i.e. implementing with integrity - become critical to the satisfactory functioning of the open system in a constantly changing environment. And as it should be these goal setting and implementing activities in themselves work out to be information processing activities characterized by their own brand of uncertainties; thereby making integrity of information processed through various stages an additional necessary requirement (to the integrity implications as already identified under Figure (2)).

System's research suggests that goals can be of various types: general, specific, positive, negative, clear, unclear [2]. Unclear goals are further characterized by implicit goals, which often may come with time delay. A system can have multiple goals, and, depending on type, goals can be multi-criteria or few (single) criteria. In multiple goal situations, goals can be independent or interdependent. Further, goals are characterized by many factors that may lead to large number of information variables which within themselves may be independent or interdependent (linked positively or negatively). It so works out that complex systems are invariably characterized by multiple, interdependent, conflicting and often unclear goals described by multiple criterion and by many factors and large number of interdependent and time varying information variables. Even seemingly simple open systems are complex; e. g., a simple user interface can add substantial uncertainty and hence complexity.

There is yet another aspect. Specifically, as shown through Figures [3(a)] and [3(b)], the information processing for the operable goal setting is characterized by its own uncertainty; thereby ensuring, in the goal set, ambiguity for strategic uncertainty. This indeed is a welcome requirement as it is this ambiguity that provides a basis for constructing an acceptable opportunity space for business for continuous innovation in a changing environment [2,9]. However, this ambiguity may also constitute an entry point for such planning and design processes and procedures (human behavior included) which may not fit the core business values, and, hence, may not be acceptable. In other words, the ambiguity in goal set would bring in strategic uncertainty and, therefore, a risk element. As a result, the methodology for operable goal implementation would also need to develop information systems for constructing acceptable opportunity (innovation) and constraining (process and procedure) spaces in order to increase the benefits of the positive risk (acceptable opportunity) and reduce the implications of the negative risk (unacceptable procedures and processes).

It is based on these goal setting and opportunity and constraining space defining activities that the subsequent stages in goal implementation can be carried out. Specifically, one can develop the structural model for information variables by observing the changes that the information variables (identified from the operable goal setting exercise) undergo over time and/or through study of their co-variances with time delays. This requires collection and integration of information over time, and thus becomes an information processing activity. This would then need to be followed by developing time sequence development trends, i.e., information dynamics model; so as to model the information structure dynamics. As described through Figures (6-7), respectively, their own types of uncertainties characterize both of these information-processing stages.

Given the customer requirements for products and/or services at any time 't', the business process IS view, in the form of 6<sup>th</sup> level control, would then need to develop (using the model of information structure dynamics obtained as above) the flexible information decision for control implementation within the boundaries of the opportunity and constraining spaces. This then gives the framework for removing the inadequacy in methodology for undertaking the inductive modeling exercise at the 6<sup>th</sup> level information control.

The task of delivering the flexible information decision as a result of the information processing at the 6<sup>th</sup> level in Figure (2) cannot be seen merely as that of forecasting (prediction), evaluation of alternatives and selection (as traditionally suggested under system engineering techniques as also in literature [7]). It must be seen as that of dealing with maximal information involving a process of information gathering and processing which leads from the initial recognition of a problem, i.e. operable goal setting followed by subsequent stages as shown in Figures (3-8). This indeed is an important observation as it offers a workable method for an inductive exercise to identify unstructured information so crucial to understand the information processing that is carried out by an open system. It also develops an improved business process IS view model over what has been suggested in Figure (2).

Further, as each of these information processing stages are impacted by uncertainties at respective stages [see Figures (3–8)], all through there is loss of integrity as the maximal information gets processed. This results in inaccurate, inconsistent and unreliable processing of operable goal set and further stages, thereby so rendering the flexible information decision also.

It is within this framework then the improved business IS view incorporating the maximal information processing stages from operable goal setting to flexible information decision and its control implementation for customized product/service delivery need to be researched for uncertainties therein and for their I\*I implications. Needless to say, each maximal information processing stage as these, by itself, is also a complex system, thereby increasing the complexity of the business IS view by that order. As a result exhaustive I\*I studies for each of these stages offer areas of separate research investigations and are beyond the scope of the present research query. However, to tie the knots together in respect of the components, sub-systems (elements), structure and information variables of the improved business IS view, the Figures (3-8) describes systems representations of these maximal information processing stages along with uncertainties and I\*I implications.

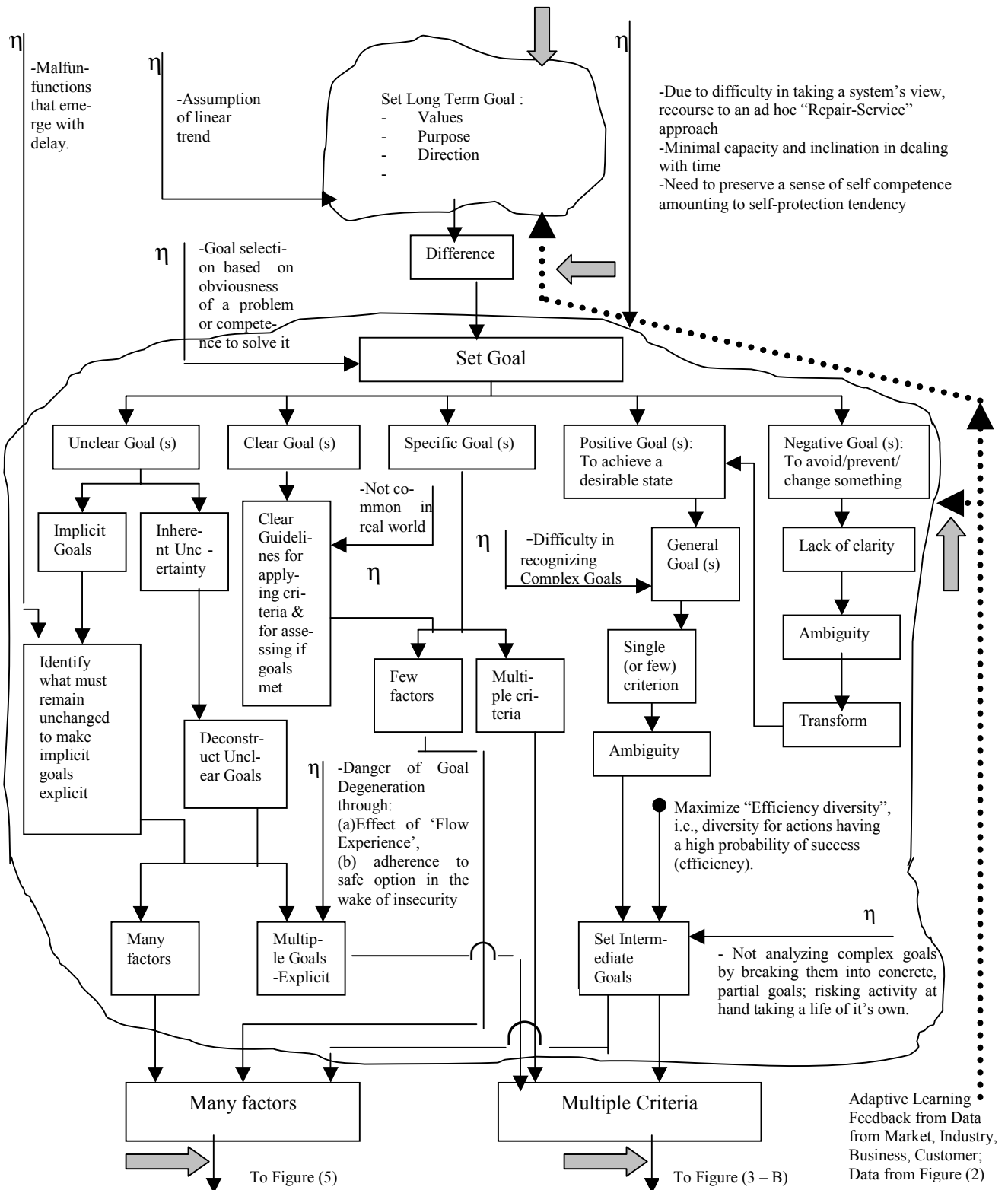


Figure (3 -a) : Systems representation of Information base and its processing for Setting Operable Goal – From ‘Set Goal’ to obtaining ‘Many Factors’ & ‘Multiple Criteria’ characterizing Problem Complexity

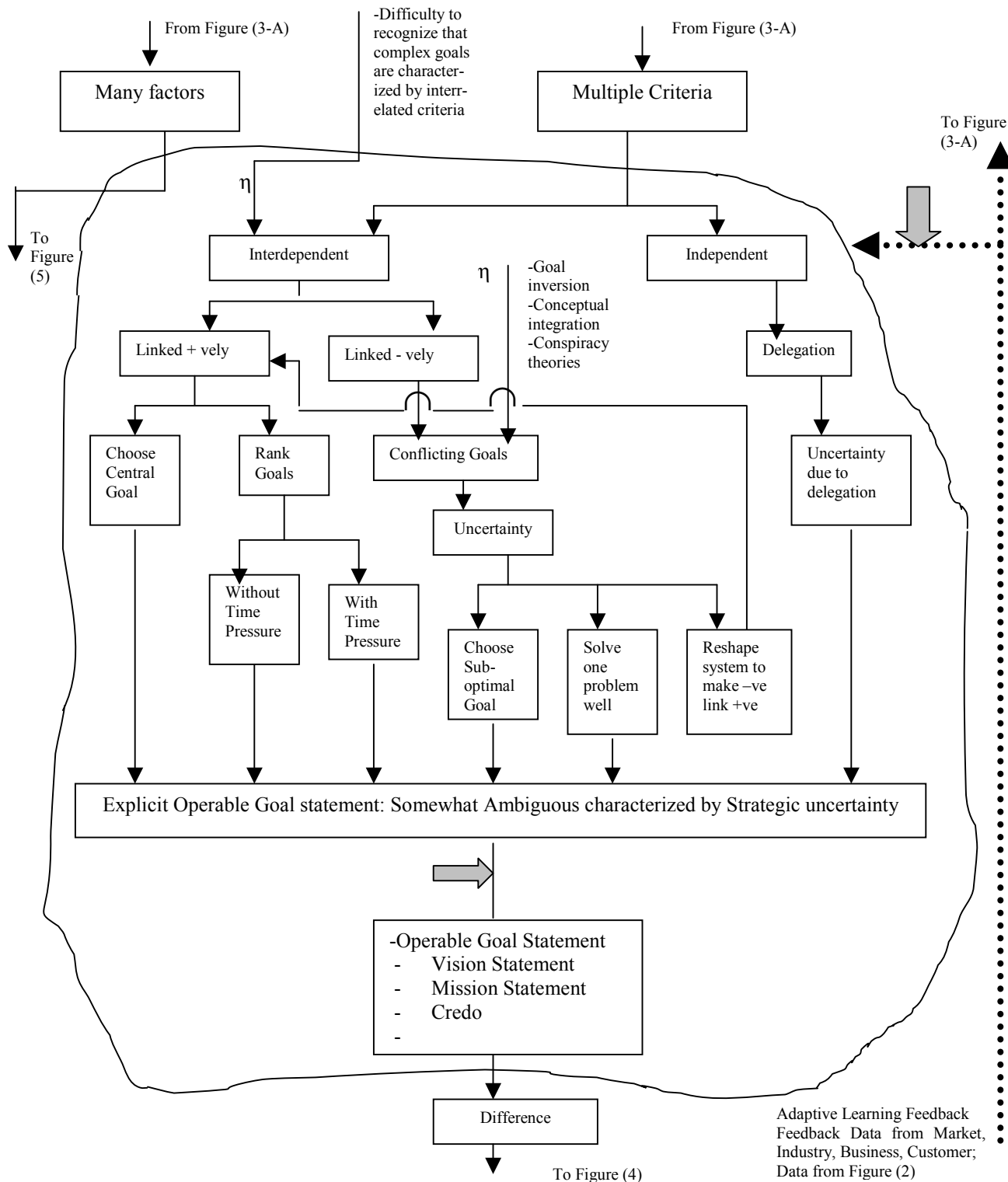
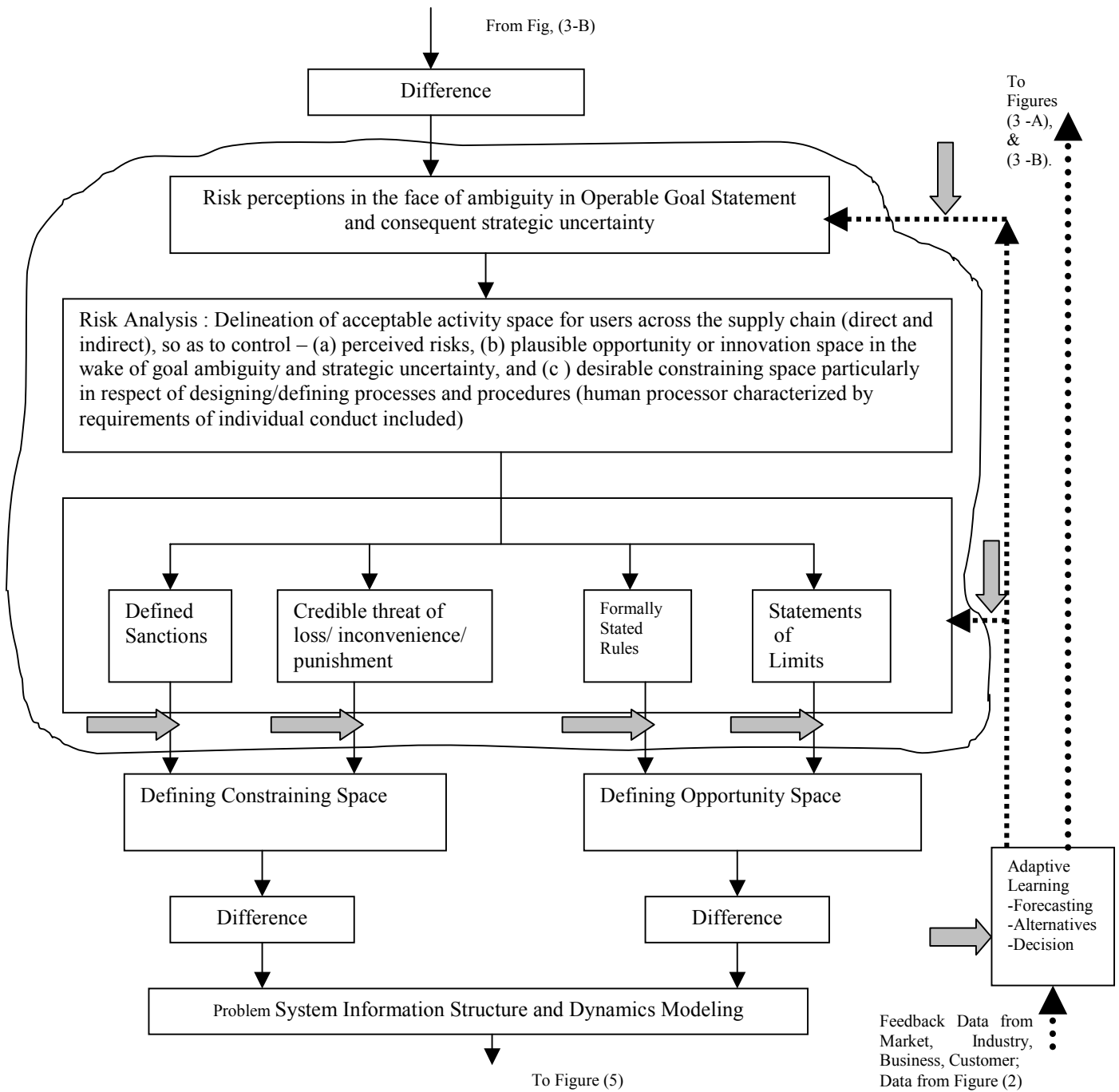


Figure (3 - b): Systems representation of Information base and its processing for Setting Operable Goal – From ‘Many Factors’ & ‘Multiple Criteria’ characterizing Problem Complexity to Operable Goal Statement



**Figure ( 4): Systems representation of Risk Analysis Information base and its processing – From Operable Goal Statements characterized by Ambiguity and Strategic uncertainty to Defining of Plng. & Design Constraining and Opportunity Spaces**

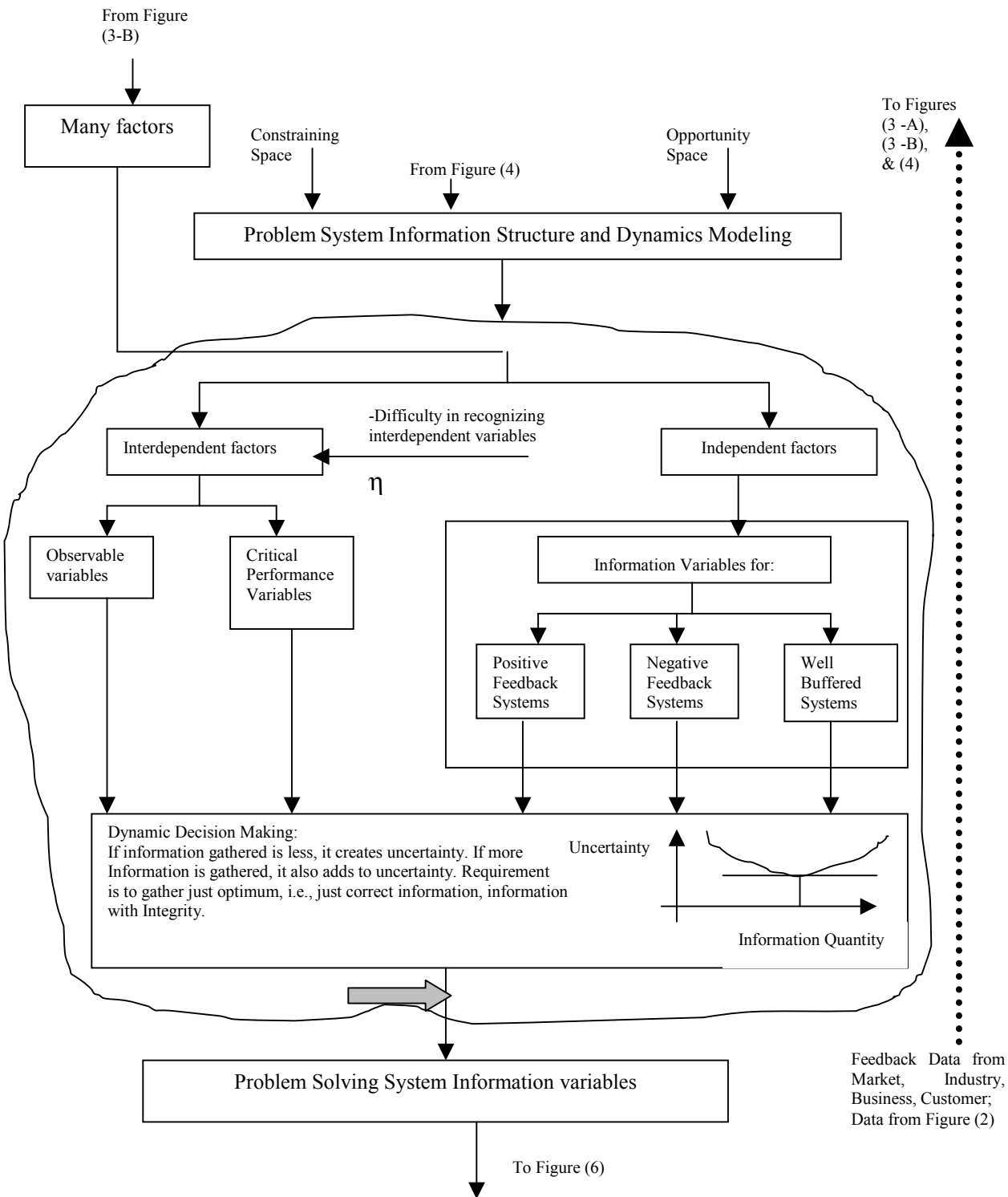
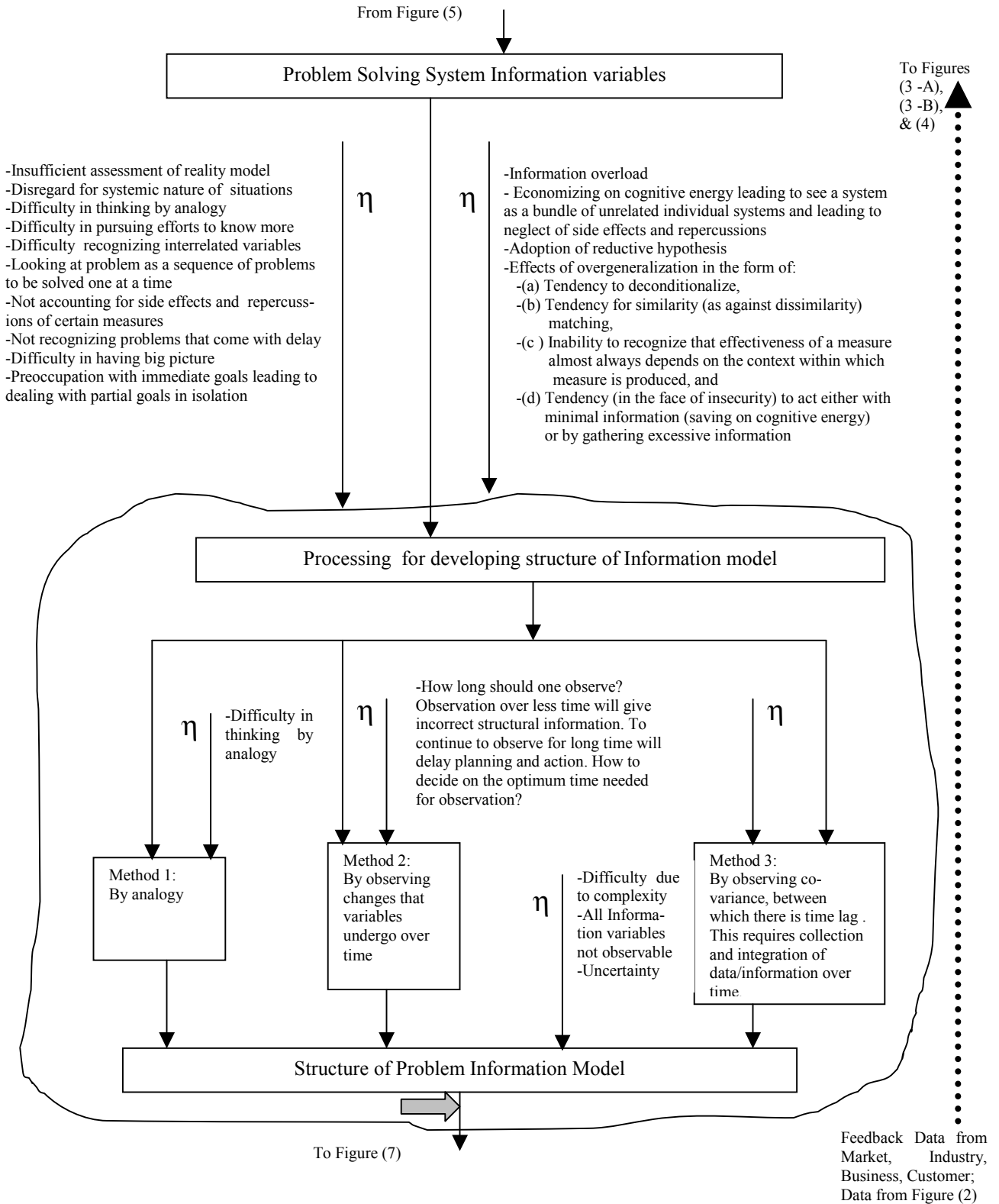
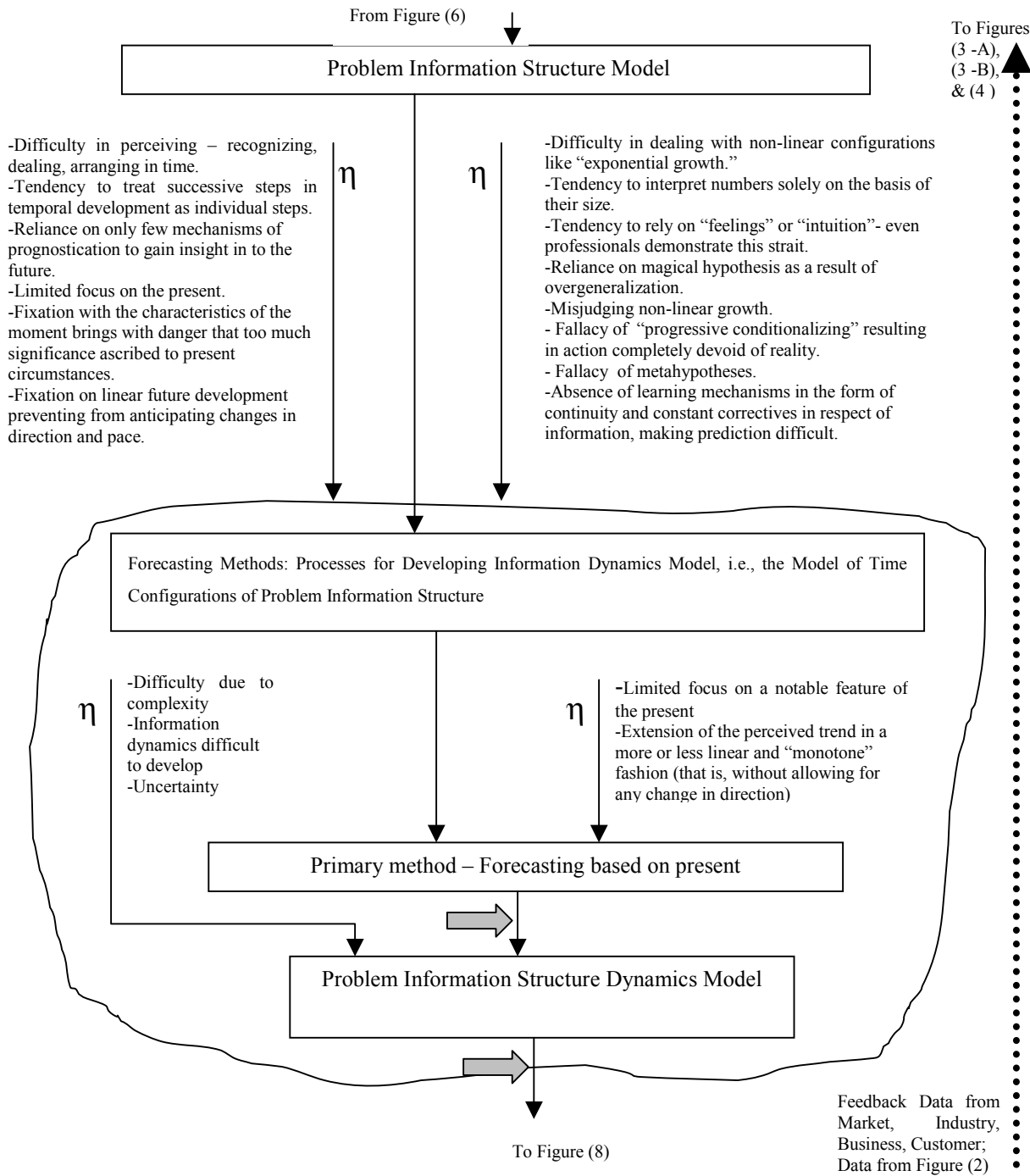


Figure ( 5): Systems representation of Information base for Problem Information Structure Modeling – From Many Factor Information Variables to Problem Solving System Information Variables



**Figure ( 6 ): Systems representation of Information base for Problem Information Structure Modeling – From Problem Solving System information Variables to Problem Solving Information Structure Model**





**Figure ( 7 ): Systems representation of Information base for Problem Information Structure Dynamics Modeling – From Problem Solving System Information Structure Model to Problem Solving Information Structure Dynamics Model**

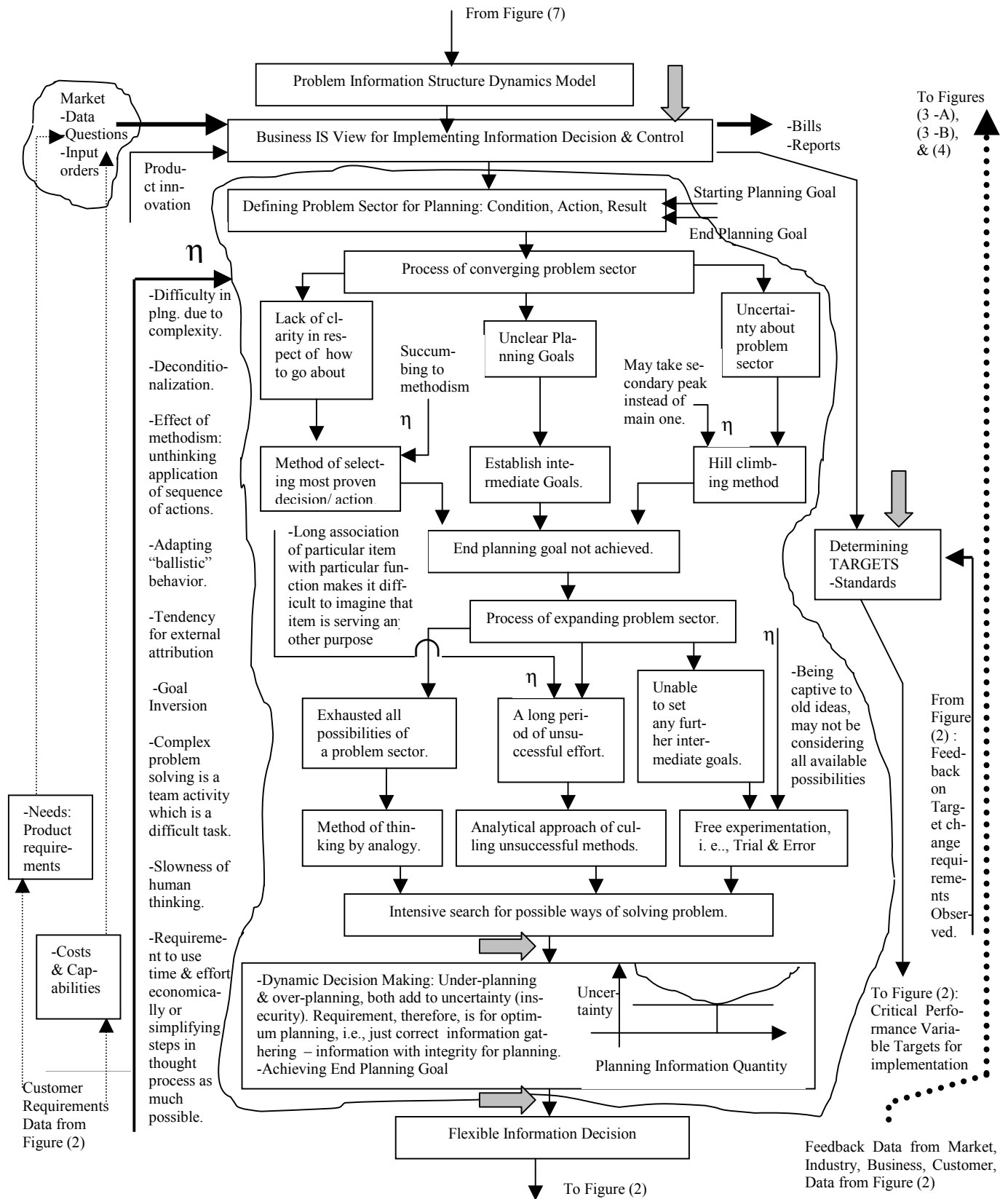


Figure (8): Systems representation of Information base for Flexible Information Decision - From Problem Information Structure Dynamics Model to Flexible Information Decision

#### 4. Defining the Information Envelope

Modeling business process as integral to an information and control system facilitates application of modern control techniques for improved business performance for strategic advantage. Of course this requires means for acquiring process data and information on current basis. The latter requirement can be met with advances in computer integrated systems and with realization of relevant data and information driven technologies. Indeed, it is here that one sees the shift from 'information technology' to 'information' in dealing with the desired objective of strategic business advantage.

Traditionally, with emphasis being on standard products and cost reduction for strategic advantage, business reality model has been viewed as a closed system having structured and repetitive information requirements wherein information content is minimal. Information models were thus developed for meeting the functions of forecasting, evaluation of alternatives, and selection in respect of decision making requirements at various levels of management [1,7].

However, as argued through the paper, this reality model of business process is inadequate. Business process IS view is an open system and, as a result, for strategic advantage emphasis required is not so much on cost reduction in isolation but on maximization of informational value. This requirement in turn goes to suggest a more workable structure for information model comprising information bases as identified through Figures (3-8) in addition to that from Figure (2).

From Figures (2-8), in the form of improved information model, thus, what really one has at hand is a conclave of information bases and the same is termed as 'Information Envelope'. In view of open system character of the business process IS view, it is *for* this Information Envelope that information is required to be continuously gathered and processed. This enables to equip the information and control system model of business to meet the challenges of customization and financial optimization for competitive advantage in a complex and changing environment; in turn making the Information Envelope based informational view of the generic business process the central theme.

Figure (9) gives systems view of an Information Envelope as above characterizing an open, complex system.

#### 5. Emergent All Encompassing View of Information Integrity

And, as shown through Figures (2) and (3-8), it is for this Information Envelope that information gathering and processing for each of its information bases is affected by uncertainties of the type not encountered traditionally, resulting in loss of I\*I. This makes I\*I, i.e. accuracy, consistency and reliability of Information Envelope, the key factor in determining the strategic business advantage.

Research investigations suggest I\*I design basis by incorporating automatic feedback control systems [4,6]. Activity of goal setting is an important requirement in the functioning of open systems. Systems techniques have considered learning mechanism as a workable method to

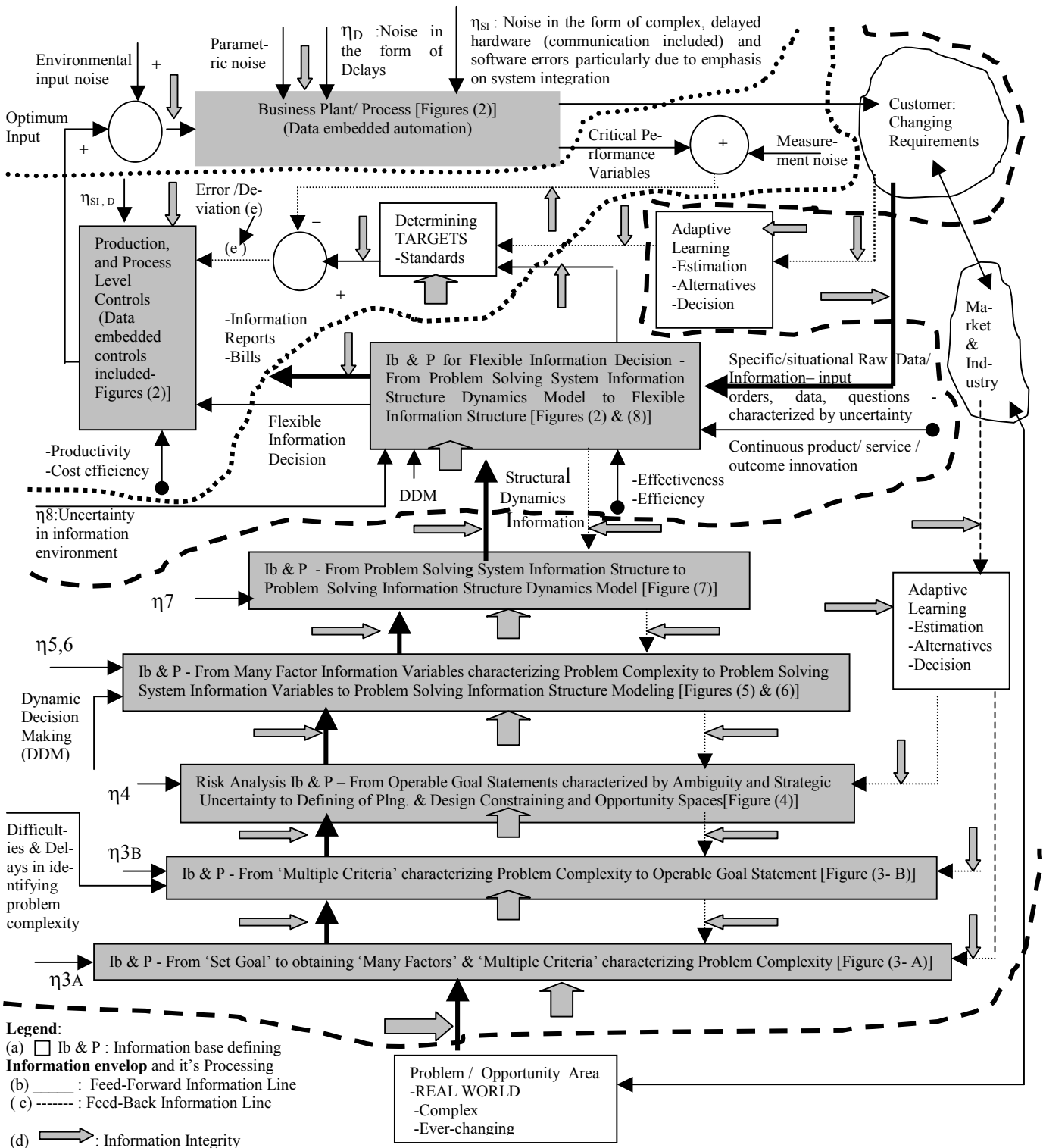


Figure (9): For a complex, changing environment, Systems View of a generic business process as integral part of a closed loop information and control system characterized by Information envelope and its processing in the presence of uncertainty and the emergent all encompassing view of INFORMATION INTEGRITY.

deal with uncertainty in a changing environment [6,13]. Also, as the concept of 'feedback' is implicit in 'learning' mechanism design, literature suggests that the 'automatic feedback control system' conceptualization of I\*I Technology can be further extended to develop an adaptive learning based I\*I planning framework for complex and changing IS environment characterized by uncertainty [6].

Details of these I\*I Technology implementation aspects are outside the scope of the present investigation. However, within their framework and based on the view of the improved business IS model and of uncertainties therein developed in this paper so far, I\*I implications can be conceptually indicated for different information bases and their respective processing stages under the Information Envelope [Figures (2-8)]. In the process what emerges is an all encompassing view of I\*I as it applies across the information and control system model of the business process operating for competitive advantage in a complex and changing environment, and the same is given in Figure (9).

## **6. Conclusion**

Generic business process covers entire supply chain from concept to delivery. A competitive business strategy calls for a good understanding of business process, which in turn requires choice of a good business model. Depending on research need such models could emphasize different facets as material, flow, equipment, money, information, etc. With advances in computer integrated systems and in data and information driven technologies, it has become possible to obtain process data and information on current basis and to manipulate it 'smarter' for strategic advantage. Specifically, what this leads to is an information and control system based model of which generic business process is an integral part. Therefore, competitive advantage can be achieved in a complex and changing business environment by systematically controlling the information processing under this business process IS view.

This requires a clearer perception of the nature of information processing. Most information processing involves some type of data conversion to information in use and, therefore, is closely related to a decision process with an objective. Even when the information is transmitted without changing form, as in a communication system, the issue is to decide the purpose or objective of the transmission.

Traditionally, within the system-engineering framework, decision process is viewed to comprise of stages of forecasting (prediction), evaluation of alternatives and selection. However, information and control system based model of a business process is an open system. For it more workable model of a decision process spans multiple stages. They are: initial problem recognition (goal setting); identifying information variables for a complex problem system; constructing problem solving opportunity and constraining spaces; developing information structure, and information structure dynamics models; and undertaking customized planning & design for development of alternatives for the evaluation of final choice for delivery of flexible information decision for control implementation.

What is significant is that all of the above stages from goal setting to final choice of flexible information decision for control implementation by themselves are complex information processing stages and, therefore, involve information gathering and processing activities with reference to their respective information bases. And of still greater implication is the reality that

at each stage these information gathering and processing activities are affected by uncertainties; resulting in errors in information processed from stage to stage.

The Information Envelope comprising the information bases is thus characterized by loss of I\*I at its all levels; thereby making Information Integrity key factor determining the strategic business advantage in a complex and fast changing environment.

### **References**

1. Beniger J. R., "The Control Revolution", Harvard University Press, Cambridge, USA (1986).
2. Dorner D., "The Logic of Failure – Recognizing and Avoiding Error in Complex situation", Perseus Books, USA (1996).
3. Mandke Vijay V., and Nayar M.K., "Information Integrity Technology Product Structure", Proceedings of 1998 Conference on Information Quality, Edited by Indu Shobba Chengalur Smith and Leo L. Pipino, MIT, Cambridge, Massachusetts, USA (1998).
4. Mandke Vijay V., and Nayar M.K., "Design Basis for Achieving Information Integrity – A Feedback Control System Approach", IFIP TC 11 WG 11.5 second Working Conference on Integrity and Control in Information Systems, Edited by S. Jajodia, W. List, A.W. McGregor and Leon A.M. Strous, Kluwer Academic Publishers, London (1998), pp. 169-190.
5. Mandke Vijay V., and Nayar M.K., "Modeling Information Flow for Integrity Analysis", Proceedings of 1999 Conference on Information Quality, Edited by Yang W. Lee and Giri Kumar Tayi, MIT, Cambridge, Massachusetts, USA (1999).
6. Mandke Vijay V., and Nayar M.K., "Information Integrity Imperative for Competitive Advantage in Business Environment characterized by Uncertainty", Proceedings of Information Technology for Business Management Conference, August 21-25, 2000, Edited by Renchu Gan, 16th World Computer Congress, Beijing, China (2000), pp. 115-125.
7. Matthews Don Q., "The Design of the Management Information System", Auerback Publishers, NY (1971).
8. Peschon J., "Disciplines and Techniques of Systems Control", Blaisdell Publishing Co., NY (1965).
9. Simmons R., "Levers of Control", Harvard Business School Press, Boston, USA (1995).
10. Spectrum Series, "Computers and manufacturing productivity", An IEEE Press Book (1987).
11. Storey N., "Safety-Critical Computer Systems", Addison-Wesley Longman, England (1996).
12. "Supply Chain Management: Building the Agile Enterprise", Gartner Group Symposium, 11-15 October, 1999, Walt Disney World, Orlando, Florida, USA (1999).
13. Tou Julius, "Modern control Theory", McGraw-Hill Co., NY (1964).

## **Progress in Information Quality: Why So Slow?**

**Panelists:** Larry English, Information Impact Int'l Inc  
Stu Madnick, MIT Sloan School of Management  
Ken Orr, Ken Orr Institute, Inc.  
Tom Redman, Navesink Consulting Group  
Tony Tortorice, Predictive Modeling LLC

**Moderator:** Jim Funk, S.C. Johnson and IQ-2001 Conference Co-Chair

**Synopsis/Rationale:** The quality revolution swept much of American manufacturing in the 1980s and early 90s. And during this same time period a number of intrepid souls turned their attention to data and information quality. Everyone knows that data and information are at the heart of the Information Age. And a number of trends, including the astounding growth rates in the quantity of data and information created and the penetration of the Internet, have made data and information even more important.

While any number of organizations have made stunning improvements (and reaped compelling benefits), progress is still not what one would have expected (and certainly hoped). In some organizations a few people are struggling mightily to make the issues, and their importance, known. But most organizations are blissfully ignorant of how poor their data are. And they continue to anger customers, incur added costs, and make bad decisions.

This panel will explore why progress is so slow, with an eye toward identifying what can be done to help progress accelerate.

## **A Conceptual Framework and Belief-Function Approach to Assessing Overall Information Quality**

Matthew Bovee  
mbovee@ukans.edu  
Doctoral Student  
School of Business  
The University of Kansas  
Lawrence, KS 66045

Rajendra P. Srivastava  
rajendra@ukans.edu  
Ernst & Young Distinguished  
Professor and Director  
Ernst & Young Center  
for Auditing Research  
and Advanced Technology  
School of Business  
The University of Kansas  
Lawrence, KS 66045

Brenda Mak  
bmak@ukans.edu  
Assistant Professor  
School of Business  
The University of Kansas  
Lawrence, KS 66045

**Abstract:** We work in an information economy, interact in an information society, and live in an information world. As information availability becomes commonplace, the ability to rapidly define and assess information quality (IQ) for decision-making provides a potential strategic advantage. Yet despite its importance and value, IQ is often ignored or its models and definitions non-intuitive, domain specific, ambiguous or lacking important concepts. A readily applicable, simple and intuitive model bridging features of other key IQ models and addressing pre-existing problems is needed to facilitate assessment. We present such a model based on a user-centric view of IQ adapted from Wang et al. (1995), and discuss its extensions.

The model consists of four essential attributes (or assertions): ‘Accessibility,’ ‘Interpretability,’ ‘Relevance,’ and ‘Credibility.’ Four elements lead to an evaluation of credibility: ‘Accuracy,’ ‘Completeness,’ ‘Consistency,’ and ‘Non-fictitiousness.’

IQ assessment is analogous to audit by evidence aggregation. We anticipate users will be more able to assign comfort or assurance levels to quality parameters based on evidence. Such assignments are readily modeled with belief functions, but not a probability framework. Expression of audit evidence has also been demonstrated to best follow a belief function framework. Therefore we present our model as an evidential network under the belief-function framework to permit user assessment of quality parameters. Several algorithms for combining assessments into an overall IQ measure will be explored. Examples in the domain of medical information are given.

### **I. Introduction**

We work in an information<sup>1</sup> economy (Neef, 1998), interact in an information society, and live in an information world (Stonier, 1991). Identification and management of corporate

---

<sup>1</sup> We refer to a model of information (Bovee, M.W. and Srivastava, R.P, 2001) that encompasses input and data – simple information – as well as more complex and typically recognized forms of information.



information has become a specialized business sub-discipline, but availability of information alone is no longer a strategic advantage – quality of information is (Huang et al., 1999). We implicitly depend on the quality of the information we use in decisions, yet poor quality information is a source of lost productivity or failed enterprise (Huang et al., 1999; Wand & Wang, 1996; Wang & Strong, 1996; Strong, Lee & Wang, 1997). For sources such as the Internet, the quality of information available is of serious concern and its uncritical use poses serious risks. Biermann (1999) and Silberg (1997) cite glaring omissions and inaccuracies in online medical information.

Despite its importance and value, the quality of information from many contexts is often variably or loosely defined, or simply ignored (Fox et al., 1994; Huang et al., 1999; Wang, Storey, & Firth, 1995). Yet a means to assess information quality (IQ) for decision-making is vital. Without clearly defined attributes and their relationships, we are not just unable to assess IQ, we may be unaware or incapable of dealing with the problem. We need to understand the attributes of IQ and to have a broadly applicable, meaningful way to combine them into a single measure of quality. Unfortunately, pre-existing models contain various problems that hinder this: limitation to a specific view of information or quality, missing attributes, and confusion or dependence between attributes and their elements. For example, one well-known product-oriented model of IQ (Wang, Reddy and Kon, 1995) presents a key IQ attribute of Believability, with an element of source credibility. Yet something that is credible is defined as having sufficient evidence to be believed (American Heritage Dictionary, 1992), and thus there is circularity between the levels. Also, since evidence of source credibility may be assessed without examination of the information itself, any weight placed on credibility occurs at the wrong level in the model. We elaborate this concept further in Section III.

In another example, a systems-oriented IQ model (Wand and Wang, 1996) evaluates many intrinsic aspects of information completeness and consistency, but fails to include an attribute such as ‘non-fictitiousness’, an important attribute of information from auditing (e.g., see Mautz and Sharaf, 1964). Non-fictitious information is neither false nor redundant. For example, a hospital’s patient record database would violate a ‘Non-fictitiousness’ attribute if it contained: 1) records for one or more non-existent patients, 2) redundant (i.e. wrongly repeated or duplicate) patient records, 3) fictitious fields or 4) fictitious values for valid fields.

Moreover, an empirically determined model (Wang and Strong, 1996) mixes intrinsic and extrinsic attributes and also mixes quality attributes with items of evidence that provide a level of comfort or assurance that quality attributes are met. For example, ‘Completeness’ deals with an intrinsic attribute of the information whereas in Wang and Strong (1996) model it is classified under ‘Contextual’ quality criteria. Contextual criteria deal with the user’s perspective of information such as ‘Relevancy’ or their level of comfort that criteria are met. To illustrate this point further, consider the earlier example of a hospital’s patient record database. ‘Completeness’ implies that the database contains all the patients’ records with values in all its fields and no patient records or field values are missing. A user who determines a record to be *sufficiently* complete for their purposes is making a judgment or evaluation based on evidence relative to fixed criteria.

Mentioned earlier, the second problem with the empirical model of Wang and Strong (1996) is the mixing of quality attributes with items of evidence. For example, ‘Reputation’ and ‘Believability’ are classified as intrinsic attributes of quality. But reputation is a piece of evidence supportive of one or more intrinsic quality attributes. An information source or provider

with a good reputation should receive a higher level of comfort or assurance that our expected criteria for intrinsic quality attributes are met. A disreputable or unknown source should instead receive a lower level of comfort that such criteria are met. Also, believability is not an intrinsic attribute, as Wang et al. have classified it; rather it is an expression of comfort or confidence based on evidence that the intrinsic attributes are met (Srivastava, 2001). ‘Objectivity’ is another dimension classified as an intrinsic attribute in the empirical model. However, as an expression of lack of bias it refers to the information source or information-generating process, not the information. For example, suppose a hospital’s patients’ database contains a field termed ‘personality.’ This field may contain, the values: ‘pleasant’, ‘average, and ‘grouchy’. These values do not have objective measures. They are subjective judgments. But these values have still the same intrinsic quality attributes of, for example, ‘Accuracy’. If a value is measured objectively, such as a patient’s temperature, then the level of comfort that the value is accurate depends on the (typically high) reliability of the measuring instrument. However, when the value is measured subjectively, as in the case of ‘personality’, the level of comfort that the field value is accurate is not easily assessable. Thus, ‘Objectivity’ does not represent an intrinsic attribute of quality, but how the values are measured.

What is needed is an IQ model flexible enough to work across various domains and purposes of user interest, robust enough to capture criteria of interest and of importance to the user in the production process, with clearly defined theoretical constructs as dimensions for testing against consumer perceptions. A means of combining evaluations assigned to IQ criteria is also needed. This paper presents such a simple and intuitive framework that incorporates features of other key IQ models and addresses pre-existing problems of interdependence, omission and confusion within dimensions. The model is then described as an evidential network under the belief-framework for explicitly tracking user assessment of the level of assurance obtained for various quality attributes and combining them into an overall IQ assessment.

The remaining sections of the paper are as follows: information, quality and IQ definitions; discussion of the strengths and weaknesses of key existing IQ models; description of the modified IQ model; description of the modified IQ model as a evidential network; conclusions; and directions for future research.

## II. Information and Quality

In this section we discuss the definitions used for information, quality, and information quality, and present a categorization of information quality views and models.

### *Information*

The origin of the word *data* is a Latin noun, *datum*, meaning something that is given (Flexnor and Hauck, 1987). An alternate definition is “facts or *pieces of information*” (Flexnor and Hauck, 1987, pg 508, italics added). “Inform” means to give form or character (Davenport and Prusak, 1998; OED, 2001). Thus we use the definition that information is (Bohn, Davenport and Prusak, 1998; Flexnor and Hauck, 1987), or contains (Stonier, 1991), input or pieces of information (data) organized to some purpose<sup>2</sup>.

---

<sup>2</sup> A detailed discussion of this within a molecular model of information, including input, data, information, and transformations between each stage can be found in Bovee, M.W. and Srivastava, R.P. (2001).

There are at least six different schools of thought regarding information (Table 1). Each embodies the concept of information as a signal with senders and receivers (Redman, 1996), and each is consistent with our treatment of information created from structured input or data.

**Table 1. Information Schools of Thought.** (Redman, 1996)

School	Perceptions
Information Management	Processed data
Infological	Knowledge or information used for decision making or action-taking
Statistical	Relevant part or summary of data from an experiment
Everyday Use	Message part that informs
Information Theory	Uncertainty reduction
Thermodynamic	Inverse of entropy

Some information definitions (e.g. Davenport and Prusak, 1998) invoke fitness for the user’s purpose to discriminate data from information. This invites confusion between the structured information, which is stable across user contexts (Stonier, 1991), and its usefulness. Input needs to be organized to *some* purpose to be information, but not necessarily a specific purpose nor that defined by a given user. Fitness of use for the domain and purpose of interest to the user defines information *quality*, not information. Otherwise, we should recognize “useless information” as an oxymoron.

### **Quality**

There is long-standing support for the user-centric, product-oriented approach to defining quality (Juran, 1989; Deming, 1982; Garvin, 1987; Huang, 1999; Wang and Strong, 1996), and there is intuitive simplicity in the approach. Fitness of use as an IQ definition also has an additional advantage. Since information is highly fungible – the same information may be used by consumers with widely variant purposes and grossly dissimilar domains of interest – we need a highly flexible, consistent definition. Unlike other products, typically assessable quality dimensions and their criteria for the definition of fitness for use (Garvin, 1987) applied to information are absent or radically different.

### **Information Quality**

Just as there are multiple perspectives or approaches to the concepts of information and quality, there are multiple views on what defines IQ or its dimensions (see Table 2 for details). These vary based on the definitional approach to quality (intrinsically or extrinsically defined) as well as the model of information (theoretical, system or process output, or product). Theoretical models (e.g. Wang, Reddy and Kon, 1995) define IQ conceptually based on introspection and logical analysis. Process-focused models (e.g. Kinney, 2000) view information as a by-product of measurement. If the measurement process is accurate and properly applied according to user requirements, then the resulting output is expected to be quality information. System-focused models center on specifying the many views and formats involved in the collection, storage, retrieval and display of information (Redman, 1996) such that the information that results from the process or the system should correctly represent the real-world view of interest to the user (e.g. Wand and Wang, 1996). A user-centric model (e.g. Wang and Strong, 1996) defines quality information as meeting user needs according to external, subjective user perceptions.

**Table 2. Information and Quality Model Perspectives.**

Information Model	Theoretical	System/Process Oriented	Product Oriented; User-Centric
Quality View	Intrinsic		Extrinsic
Information View	Intuitive		Empirical
Information Quality	Conceptually derived w/theoretical explication	Depends on system or process design to replicate the user's requirements or world view	Based on user perception

Each of these views has its strengths and weaknesses. Theoretical models provide good explication of constructs and relationships that are grounded in the literature, but they tend to treat quality as an objective construct, ignoring user perceptions. Systems- and process-oriented models tend to capture more details specific to intrinsic attributes of information, but view information as a process output or byproduct. User-centric models capture the broader range of attributes described as important by information consumers, but do not provide clearly defined constructs for these attributes. But, just as there are common dimensions for determining the quality of a type of wood for a given use, despite the plethora of types and uses available (grain, color, hardness, cost, rarity, etc.), general attributes applicable across domains and purposes of interest to information users may provide stable dimensions for assessing its quality.

### III. IQ Models, Problems and the Modified Conceptual Framework

To determine and evaluate IQ criteria we take the perspective of an information user and outline the basic things we require for an information product to be useful. In the process we discuss these criteria relative to key IQ models and the significance of any differences. To clarify the model dimensions and criteria and any comparisons we use the example of a medical patient's clinical evaluation report.

The model may be summarized by a simple, ordered mnemonic of the main criteria: **AIRC** – **A**ccessibility, **I**nterpretability, **R**elevance, and **C**redibility (Table 3).

**Table 3. Basic Aspects of Information Quality Conceptual Framework.**

Criteria	Basic Description
<b>A</b> Accessibility	Ability to retrieve information
<b>I</b> Interpretability	Understandability and meaningfulness of information to the user
<b>R</b> Relevance	Applicability of information to the user's domain and purpose of interest
<b>C</b> Credibility	Degree of belief assigned by the user to information based on whether intrinsic attributes of <i>Accuracy, Completeness, Consistency and Non-fictitiousness</i> are met

Briefly outlined, to determine the quality of information – its fitness for our use – we must: 1) be able to get information which we might find useful (*Accessibility*); 2) be able to understand it and find meaning in it (*Interpretability*); 3) find it applicable to our domain and purpose of interest (*Relevance*); and 4) believe it to be credible (*Credibility*). Note that as an information user we would dismiss or discount information that meets our criteria for all but one of any of the above aspects, each of which may be more than just a binary value. We next describe these major aspects and their respective elements below, and discuss them relative to other key IQ models. Since our reasoning and model closely parallel that of Wang, Reddy and Kon (1995), we

especially note important differences with that model. Explanatory examples are given from the domain of medical information (see also Table 4).

### ***Accessibility***

First we must be able to get information for it to be of use. IQ models that focus on information as a by-product of the system rarely cite information accessibility as a quality criterion (Wang, Storey and Firth, 1995), yet it is obviously critical to the user (Wang, Strong & Lee, 1997; Wang, Reddy and Kon, 1995; Wang & Strong, 1996). Information retrieval may require a certain amount of time or have an associated measure of cost to the user<sup>3</sup>. If information is inaccessible, all other qualities of it are irrelevant.

A hospital medical report on the outcome of patient surgery may not be needed any sooner than the end of the month for statistical purposes, or it may be needed immediately for reference and review during an examination. Off-site clinical access to such information may be free, available as for-pay products or services, or part of a private intranet. To access even different in-house information sources within a hospital intranet may also require widely different times, and have associated costs. Depending on their setting, a physician might conceivably have to decide between results only on hand, available by mail, by fax, or by electronic transfer, and the delays and costs associated with each choice.

### ***Interpretability***

Second, we must be capable of understanding any information retrieved (it must be intelligible) and if it is understandable we need to be able to derive meaning from it. Intelligible information is *capable* of being understood by the user and meaningful information conveys to the user some sense, significance, or meaning (American Heritage Dictionary, 1992; Flexnor and Hauck, 1987; OED Online, 2001). System-focused IQ models tend to assume interpretability of output information is inherent in the correct specifications of the system, the database design or the data production process (Wang, Storey and Firth, 1995; Wand & Wang, 1996; Kinney, 2000). Wang, Reddy & Kon (1995) describe interpretability as the understandability of the syntax and semantics of information. Yet this is the bare minimum of intelligibility – users may place much broader demands on the interpretability of information (Wang & Strong, 1996), ranging to practically requiring that “the thing speaks for itself” (Lieberman, 2000). If information is either unintelligible or meaningless to us, all its other qualities are irrelevant.

Unintelligible or meaningless information to one user may be intelligible or meaningful to another. The information embedded or created in its structure has not changed, but its quality differs according to user-determined criteria. For example, the same medical report of a patient’s blood chemistry could be written in either English or Japanese. To a physician who could not read it, the Japanese report would be unintelligible and meaningless. However, a physician fluent in both languages might find either report equally suitable. Intelligibility is a necessary but insufficient condition for interpretability. Consider the case in which a patient who wants to know the results of their medical check-up finds the clinical report to be intelligible (i.e. in readable English), but meaningless because they lack the ability to derive meaning from it.

---

<sup>3</sup> Some may treat time and cost as synonymous, but we contend that these instances are ones in which time is so dominant a factor that cost is disregarded, or the information is free. Nonetheless, the user is free to evaluate information sources for their quality of accessibility according to their needs.

Intelligibility and meaningfulness are user-defined IQ criteria. The actual content of the information does not depend on the user, nor on the quality ratings they assign. Thus interpretability is composed of both intelligibility and meaningfulness, with intelligibility the cusp of meaningfulness.

### ***Relevance***

Third, if we have information that we can understand and interpret, we want it to be relevant based on our user-specified criteria for the domain of interest and timely to our purpose within that domain. Of course, the user-specified criteria depend on the domain and purpose in mind. For example, if a surgeon performing a surgery wants to know about the patient's potential allergic reactions to anesthesia, a database providing all the information on the patient except that would be of no use. The information may be 'Accessible' and 'Interpretable' but not relevant in terms of user-specified criteria. *Relevance* has many possible domain- and purpose-related criteria, but if the information is outdated it is useless. Thus, timeliness is an important element of 'Relevance' as discussed below.

Wang, Reddy & Kon (1995) subsume relevance under the dimension of usefulness and treat timeliness as a separate usefulness criterion. However, it seems unlikely that information could be inaccessible or unintelligible, but still useful. Also, fitness for use is the global quality evaluation being made and decomposed by the model into specific criteria. Therefore usefulness is an inappropriate label, or is placed at the wrong level in the model. Also, while information could certainly be timely but irrelevant, the reverse seems unlikely, thus the criteria are not separable.

Timeliness has two components: *age* and *volatility* of the information. Age, or 'currency' of information is simply a measure of how old the information is based on how long ago it was recorded. All other things being equal, the more recently the information was collected, the more likely it is to be relevant. For example a medical report containing a patient's blood pressure values measured at their annual physical can be considered a current measurement for purposes of evaluating long-term health status. However, if a physician were to want to know the patient's blood pressure now, a more recent measurement is preferable. Volatility of information is a measure of information instability – the frequency of change of the value for an entity attribute of interest (the 'source value'). The more volatile information is the more rapidly *any* recorded values<sup>4</sup> become outdated. Non-volatile information is stable; it does not change nor become outdated. Again, for annual physical exams the information remains valid for one year, and for routine check-ups such periodic measures of blood pressure are satisfactory. But, during surgery, blood pressure values are much more briefly valid, more volatile, and must be monitored continuously to provide information on the patient's moment-to-moment status. Annual values are, of course, irrelevant in this context.

The datedness of information varies directly with its age and inversely with its volatility. Information must be updated as frequently as the source value changes or else become outdated. However, information that is updated as frequently as the source value changes may not be necessary for the user's purpose, nor practical, feasible or cost-effective. Thus, a relative measure of outdatedness – *timeliness* – becomes an important IQ sub-element. *Timeliness* is a

---

<sup>4</sup> Other than continuously recorded real-time information, which is the opposite extreme to non-volatile information.

judgment by the user of whether information is recent enough to be relevant, given the rate of change of the source value, and the domain and purpose of interest.

If information is updated frequently enough for the user's purposes then it is timely. If not, it may be irrelevant. The less timely information is, the less likely it is to be relevant to the user. For example, a doctor may require their recovering surgery patient to only have twice-daily blood pressure measurement, even though the underlying value varies continuously. Every twelve hours, the prior blood pressure measurement becomes outdated information and becomes less timely<sup>5</sup>. If the next measurement is not made on time, the most recent (i.e. least outdated) may suffice. Measurements from a week ago, however, are certainly no longer timely at all and therefore of unacceptable quality.

Users of historical information may need information from a specific point or period in time; this is different from timeliness. One can require relevant blood pressure information to include measurements from surgeries during a specific week last year *and* that were timely when recorded.

Since information may be relevant, but inaccessible or unintelligible, we use relevance as the dimensional label, and timeliness as one specific user-determined criterion among the many possible. This matches the loading of relevance and timeliness as factors important to Contextual Quality in the empirical model by Wang & Strong (1996). Information that does not match the domain or purpose of the user is presumed useless, and information that does but is outdated is similarly useless.

### ***Credibility***

Last, given access to interpretable, relevant information we require it to also be credible. Credibility of information exists when the information is plausible, when there is sufficient reason for it to be believed by the user (American Heritage Dictionary, 1992; OED Online, 2001). This dimension corresponds most closely with aspects of information frequently considered for quality measures and thought of as intrinsic to the information itself (Wang, Storey & Firth, 1995; Wang, Reddy & Kon, 1995), or as stemming from the system design or processing of information (Wang & Wang, 1996). We consider information that is accurate, complete, consistent (Wang, Reddy & Kon, 1995) and non-fictitious (Mautz and Sharif, 1964) to be credible.

Several IQ models have categorized these criteria under dimensions other than the intrinsic nature of information (Wang, Strong & Lee, 1997; Wang & Strong, 1996). This may be the result of confusion due to the dominance of user-definitions for virtually all quality criteria once fitness for use is established as the global quality standard. Other IQ models subsume information source credibility under the dimension of believability (Wang, Reddy & Kon, 1995). As discussed earlier, credibility of an information source is evidence attesting to IQ, not an attribute of the information itself. Even though source credibility may be a criterion used by an information user (Wang and Strong, 1997), it seems more likely to be used as a heuristic or proxy for the global dimension of believability, not as a criterion for it. This can be seen upon

---

<sup>5</sup> We recognize that the meaningfulness of the information may, in part, be derived in context with other values in a time series. Thus the first of two serial measurements may actually derive more relevance after the second is obtained. However, with successive new measurements the earlier ones become more outdated, less timely and less relevant.

examining the definition for “credible” (American Heritage Dictionary, 1992). Given a credible source, other evidence – accuracy, completeness, consistency and non-fictitiousness of the information itself – may be assumed, not evaluated. Thus evaluations of source credibility should enter the model at the same level as the main dimension, as evidence in support of it rather than valuations of elements that compose it (e.g. ‘third party assurance’ in Figure 4).

To avoid a circular definition between attribute and element we substitute *Credibility* for *Believability* and leave evaluations of the source outside of the model for the time being. Information that is retrievable, intelligible and meaningful, and relevant, yet lacks all credibility, would be useless. Credibility has four elements: Accuracy, Completeness, Consistency, and Non-fictitiousness.

Accuracy deals with information being true or error free with respect to some known, designated or measured value. As part of a patient examination, the patient’s name is known and therefore comparable for accuracy to information that should contain it. The patient’s identification number is designated and may be checked for accuracy against the algorithm or context from which it was derived. Lastly, the patient’s blood pressure can be measured directly to determine if the recorded value and the measurement are the same or sufficiently close for the user’s purposes. Accuracy plays a major role in most models of IQ (Wang, Storey & Firth, 1995) as an intrinsic attribute of the information itself. Yet establishing accuracy is difficult if not impossible in many circumstances, and what is acceptable or desirable information accuracy still requires judgment on the part of the user.

Completeness deals with information having all required parts of an entity’s information present (Wang, Reddy & Kon, 1995; Wang, Storey & Firth, 1995). A patient examination report example typically requires descriptive patient information such as name, age, sex, treatment and payment details, plus the results of various visit-specific tests and any pertinent diagnoses. Absence of any of these renders the report incomplete, unless there is tolerance for missing values for some attributes. In a database environment, completeness can be in violation if a patient or patients’ records are missing or certain field values are missing.

Consistency of information requires that multiple recordings of the value(s) for an entity’s attribute(s) be consistent across time or space (Wang, Reddy & Kon, 1995; Wang, Strong & Lee, 1997). To be consistent these values must be the same in all cases (for discrete values) or closely grouped in dispersion (for continuous values). Although consistency appears frequently as a proposed quality dimension (Wang, Storey & Firth, 1995; Wand & Wang, 1996), it does not appear as a prominent feature of empirically assessed user models of IQ (Wang and Strong, 1996).

Hospitals often store information for different departments separately, and the patient records for a male admitted in one department and tested in another should both have the discrete value “Male” recorded for his gender. Having “Female” recorded in one would be both inaccurate in the single case, and inconsistent with all other sources. If this patient’s blood pressure was measured once and recorded several places, it should be the same in all instances. The patient’s blood pressure measurements taken several times at a single visit, or multiple times across departments on the same day, should be tightly dispersed.

Lastly, non-fictitiousness is an important intrinsic attribute of information as used in auditing (e.g., see Mautz and Sharaf, 1964). Non-fictitious information has no false or redundant entities, fields, or attribute values. As mentioned earlier, the ‘Non-fictitiousness’ attribute would



be in violation if the database contains: 1) one or more records for patients record(s) that does (do) not exist, 2) redundant records for certain patients, i.e., certain patient records are repeated, or 3) fictitious value(s) in certain field(s). No IQ model directly addresses all aspects of this problem. Wand & Wang (1997) present a system-oriented model that most closely approximates this, discussing meaningless combinations of information (information not corresponding to the real world) and incorrect information (information wrongly mapping to the real world). However, fictitious information is not necessarily meaningless and can correspond to the real world. In fact, a goal of deliberately falsifying information is to undetectably simulate a real-world state that *could* occur, but did not. Another type of fictitiousness is redundancy. Redundant information is permissible in some systems models of IQ (Wand and Wang, 1997), yet leads to ambiguity wherein at least one item of information should not exist but it may be difficult to discern which is false. Establishing non-fictitiousness as a measure of credibility is an important auditing process.

Thus, our conceptual model of IQ (Figure 1) consists of three essential extrinsic attributes (or assertions): 'Accessibility', 'Interpretability', and 'Relevance', and one intrinsic attribute (or assertion): 'Credibility.' The extrinsic attributes determine the user perceived quality attributes and the intrinsic attribute, "Credibility", determines the internal aspect of quality of information, which consists of five elements (or sub-assertions): 'Accuracy', 'Completeness', 'Consistency', and 'Non-fictitiousness'.

#### **IV. Evidential Network for Assessing IQ**

Srivastava and Mock (2000) have developed an evidential network for WebTrust assurance services for the purpose of evaluating whether the Webtrust assurance criteria have been met. If the evidence gathered in the process provides a sufficiently high level of confidence (0.95 on a scale of 0-1) that the WebTrust criteria are met, then the assurance provider could issue an unqualified (i.e., clean) opinion on the service. A similar evidential network approach has been applied by Srivastava, Dutta and Johns (1996) in the audit process of a healthcare unit. There are basically three issues in such evidential network approaches. First is the relationship among the variables (i.e., assertions or sub-assertions) in the network. Second is the structure of the evidential network, which in essence requires the knowledge of what piece of evidence relates to what assertion or assertions. The network structure arises due to the fact that one item of evidence may pertain to more than one assertion or sub-assertion. The third issue deals with the representation of uncertainty involved in the judgment of whether a certain variable or attribute is met, at what level of confidence, based on the evidence collected. The first issue really deals with understanding the problem at hand. In other words, one needs to know the main variables (assertions or attributes) of the network and their interrelationships. In our case, the attributes that determine the quality of information are given in Figure 3.

Srivastava and Mock (2000) and Srivastava et al. (1996) have used Dempster-Shafer Theory of belief functions (Shafer, 1976) to represent uncertainties in the evidence. They have argued (Srivastava and Shafer, 1992) that belief functions provide a better framework for representing uncertainties in the evidence encountered in the situations faced by auditors or assurance provided. A recent study by Harrison et al. (2001) in auditing and a study by Curly and Golden (1995) in psychology provide further evidence in support of using belief functions for representing uncertainty in evidential reasoning. We take the same view and argue that belief functions would better represent uncertainties associated in assessing the quality of information.

Figure 4 represents an evidential network for IQ measurement. The rounded nodes represent variables in the network. These variables are: “Information Quality” (IQ), the extrinsic and intrinsic attributes AIRC (Accessibility, Interpretability, Relevance, and Credibility), the components of relevance: ‘Timeliness’ and ‘User-specified criteria’, and the components of Credibility: ‘Accuracy’, ‘Completeness’, ‘Consistency’, and ‘Non-fictitious’. The circle with ‘&’ inside it represents an ‘and’ relationship<sup>6</sup> between the variable on the left of it with the variables on the right. For example, the main variable ‘IQ’ is connected to the four variables AIRC on the right through an ‘and’ relationship. This implies that IQ is met (i.e., IQ is high) if and only if all the variables on the right are met (i.e., each has a high level of confidence that it is met). If any one of them is not met (i.e., it takes a low values) then IQ is not met (i.e., IQ is low).

The rectangular boxes represent items of evidence pertinent to various attributes as represented by direct linkages between items of evidence and the attributes. In order to determine the overall quality of information, one needs to gather the relevant items of evidence as indicated in Figure 4, evaluate the level of support each item of evidence provides to the corresponding variable(s), and then aggregate these assessment of support in the network to determine the overall level of support for the value ‘high quality’ of IQ. Expert opinion regarding evidential inputs to the model will be gathered. We will then use a computer system known as Auditor’s Assistant developed by Shafer, Shenoy and Srivastava (1988) for combining items of evidence in a network of variables similar to Figure 4 where judgment of uncertainty is expressed under belief functions.

Using the above software, we plan to perform the following sensitivity analyses:

1. Determine how sensitive the output result is with regard to changes in the input values.
2. Determine whether one can use non-numerical inputs (e.g. very high, high, medium, low, very low) based on some range of numerical values and test the sensitivity of output values.
3. Determine which item of evidence is the most significant for the overall IQ.
4. Determine sensitivity of the relationships among variables on the overall IQ. We will use the following relationships: ‘and’, a combination of ‘and’ and ‘or’, and an averaging relationship.

## **V. Summary, Conclusion and Directions for Future Research**

The modified IQ model presented here extends and bridges previous models, resolving ambiguities in terminology and relationships of quality attributes. In particular: judgments of information source credibility exist independently of information attributes and must therefore enter from outside any information model; ‘believability’ and ‘credibility’ cannot be independent quality attributes nor an attribute and related element as they are circularly related; ‘credibility’ is a global assessment based on one or more judgments and belongs at a high level within the quality model; information timeliness is an element of relevance to the user, not independent from relevance; and, although aspects of it are found in systems-oriented data quality models, non-fictitiousness as found in auditing is an important concept absent in other IQ models. In the theoretical introduction to the model we have also clarified a potentially critical ambiguity in the

---

<sup>6</sup> At the moment we are only considering ‘and’ relationships among the variables as considered by Srivastava and Mock (2000) and Srivastava et al. (1996). Such a relationship makes sense, especially when all the attributes are essential in order for the main objective to be met. Other relationships will subsequently be tested.

definition of information by proposing that usefulness does transform data to information, nor a define a characteristic of information itself, but is a judgment of IQ. This and the clarifications above provide the theoretical foundation to permit our modified model, which forms the structure for evidential network, to then be used to evaluate overall IQ. Toward this end, we have proposed several evaluative steps to be taken in determining appropriate relationships among the network variables, including several different rules of combination.

Testing of the logical implementation and behavior of the network, however, needs to be supplemented with investigations of its applicability for information consumers (as it is designed as a *user-centric* model). We intend to empirically evaluate the network structure and attributes with information users' direct assessments of IQ. While the models that form the foundation for our modifications represent a broad range of approaches (and of users in one case), the needs or concerns of specific groups may not be properly represented by a general model. As evident through the examples and discussion, future research will focus on the perceived IQ needs of two related groups – clinical and Web information users.

In addition, applicability of the model requires evaluation through field-testing. Given the concerns with the Web information quality (health information in particular), the evidential network could be used for rating website IQ through an online interface and user feedback collected to evaluate the tool. While at least one such IQ rating tool is available (MITRETECH, 1999), it does not use belief functions for representing nor aggregating users' ratings. As discussed earlier, the belief function formalism appears to be the best way to represent such judgments.

Lastly, given the global explosion of information availability and the apparent concerns regarding online information quality, we see a need for a robust model of IQ expressed in XML. As bandwidth and processing speeds increase, a theoretically and practically proven model of IQ holds great promise as a taxonomy for metadata tags that do away with the need for manual user evaluations of IQ.

## **References**

1. Biermann, J.S., Golladay, G.J, Greenfield, M.L.V.H. and L.H.Baker. "Evaluation of Cancer Information on the Internet," *Cancer*, August 1999, 86(3):381-390.
2. Bovee, M.W. and Srivastava, R.P. "A Molecular Model of Information," working paper, School of Business, University of Kansas, 2001.
3. Choo, W.C. *Information Management for the Intelligent Organization*, Information Today, Inc., Medford, New Jersey, USA, 1995.
4. Curley, S. P., and Golden, J.I. 1994. Using belief functions to represent degrees of belief. *Organization Behavior and Human Decision Processes*: 271 – 303.
5. Davenport, T.H. and L. Prusak, *Working Knowledge: How Organizations Manage What They Know*, Harvard Business School Press, Boston, Massachusetts, USA, 1998.
6. Deming, W. E. *Quality, productivity and competitive position*. Massachusetts Institute of Technology, Center for Advanced Engineering Study, Cambridge, Massachusetts, 1982.

7. Flexner, S. B. and Hauck, L.C., Eds. *The Random House Dictionary of the English Language, 2<sup>nd</sup> Edition*, Random House, Inc., New York, New York, 1987, pg. 508 (data), pg 980 (info).
8. Fox, C., Levitin, A. and Redman, T. "The Notion of Data and Its Quality Dimensions." *Information Processing and Management*, 1994, 30(1):9-19.
9. Garvin, David A. "Competing on the eight dimensions of quality," *Harvard Business Review*, November-December, 1987, pp 101-109.
10. Harrison, K., R. P. Srivastava, and R. D. Plumlee, "Auditors' Evaluations of Uncertain Audit Evidence: Belief Functions versus Probabilities," in *Belief Functions in Business Decisions*, edited by R. P. Srivastava and T. Mock, Physica-Verlag, Heidelberg, Springer-Verlag Company (2001).
11. Huang, K-T., Lee, Y.W and Wang, R.Y. *Quality Information and Knowledge*, Prentice Hall, Upper Saddle River New Jersey, USA, 1999.
12. Juran, J. M. *Juran on leadership for quality*. Free Press, New York, 1989.
13. Kinney, *Information Quality Assurance and Internal Control For Management Decision Making*, McGraw-Hill Higher Education, Boston, Massachusetts, USA, 2000.
14. Lieberman, E. "*Tres Ipsa Loquitor*," keynote speech at MIT IQ Conference, Boston, Massachusetts, USA, 2000.
15. Levitin, A. and Redman, T. "Quality Dimensions of a Conceptual View." *Information Processing and Management*, 1995, 31(1):81-88.
16. Mautz and Sharif, *Philosophy of Auditing*, American Accounting Association, Sarasota, Florida, 1964.
17. MITRETEK, *Criteria for Assessing the Quality of Health Information on the Internet – Policy Paper*, <http://hitiweb.mitretek.org/docs/policy.html>, 1999
18. Neef, D.E., *The Knowledge Economy*, Butterworth-Heinemann, Boston, Massachusetts, USA, 1998.
19. Oxford English Dictionary Online, 2001, [www.oed.com](http://www.oed.com).
20. Redman, Thomas C., *Data Quality For The Information Age*, Artech House, Boston, 1996.
21. Shafer, G. 1976. *A Mathematical Theory of Evidence*. Princeton University Press.
22. Shafer, G., P.P. Shenoy and R. P. Srivastava, "AUDITOR'S ASSISTANT: A Knowledge Engineering Tool For Audit Decisions," *Proceedings of the 1988 Touche Ross University of Kansas Symposium on Auditing Problems*, May 1988, pp. 61-79.
23. Silberg, W.M., Lundberg, G.D., and R.A. Musacchio. "Assessing, Controlling, and Assuring the Quality of Medical Information on the Internet: Caveant Lector et Viewor – Let the Reader and Viewer Beware," *Aging and Information Technology*, Fall 1997, 53-55.
24. Smets, Ph. "The combination of evidence in the transferable belief model." *IEEE Pattern Analysis and Machine Intelligence*, 1990, 12:447-458
25. Smets, Ph. "The Transferable Belief Model for Quantified Belief Representation." In *Handbook of Defeasible Reasoning and Uncertainty Management Systems, Vol. 1: Quantified Representation of Uncertainty & Imprecision*, Gabbay D. and Smets Ph. (Series Eds). Ph. Smets (Vol. eds.), Kluwer, Dordrecht (1998) 267-301.
26. Srivastava, R.P. personal conversation, 2001.

27. Srivastava, R. P., S. K. Dutta, and R. Johns, "An Expert System Approach to Audit Planning and Evaluation in the Belief-Function Framework," *International Journal of Intelligent Systems in Accounting, Finance and Management*, Vol. 5, No. 3, 1996, pp. 165-183.
28. Srivastava, R. P., and G. Shafer, "Belief-Function Formulas for Audit Risk," *The Accounting Review*, April 1992, pp. 249-283.
29. Srivastava, R. P. and T. J. Mock, "Evidential Reasoning for WebTrust Assurance Services," *Journal of Management Information Systems*, Vol. 10, No. 3, Winter 1999-2000, pp. 11-32.
30. Stonier, T. "Toward a new theory of information," *Journal of Information Science*, 1991, 17(5):257-263.
31. Strong, D.M., Lee, Y.W., and Wang, R.Y. "Data Quality in Context," *Communications of the ACM*, May 1997, 40(5):103-110.
32. *The American Heritage Dictionary of the English Language, Third Edition*, Houghton Mifflin Company, 1992.
33. Wang, R.Y., Reddy, M.P., and Kon, H.B. "Toward quality data: An attribute-based approach." *Decision Support Systems*, 13(1995): 349-372.
34. Wang, R.Y. and D.M. Strong. "Beyond Accuracy, What Data Quality Means to Data Consumers." *Journal of Management Information Systems*, Spring 1996, Vol. 124, pp. 5-34.
35. Wang, R.Y., Storey, V.C. and Firth, C.P. "A Framework for Analysis of Data Quality Research," *IEEE Transactions on Knowledge and Data Engineering*, August 1995, 7(4):623-639.
36. Wand, Y. and R.Y. Wang. "Anchoring Data Quality Dimensions in Ontological Foundations," *Communications of the ACM*, November 1996, 39(11):86-95.

Figure 1: Wang, Reddy and Kon (1995) Model of Data Quality

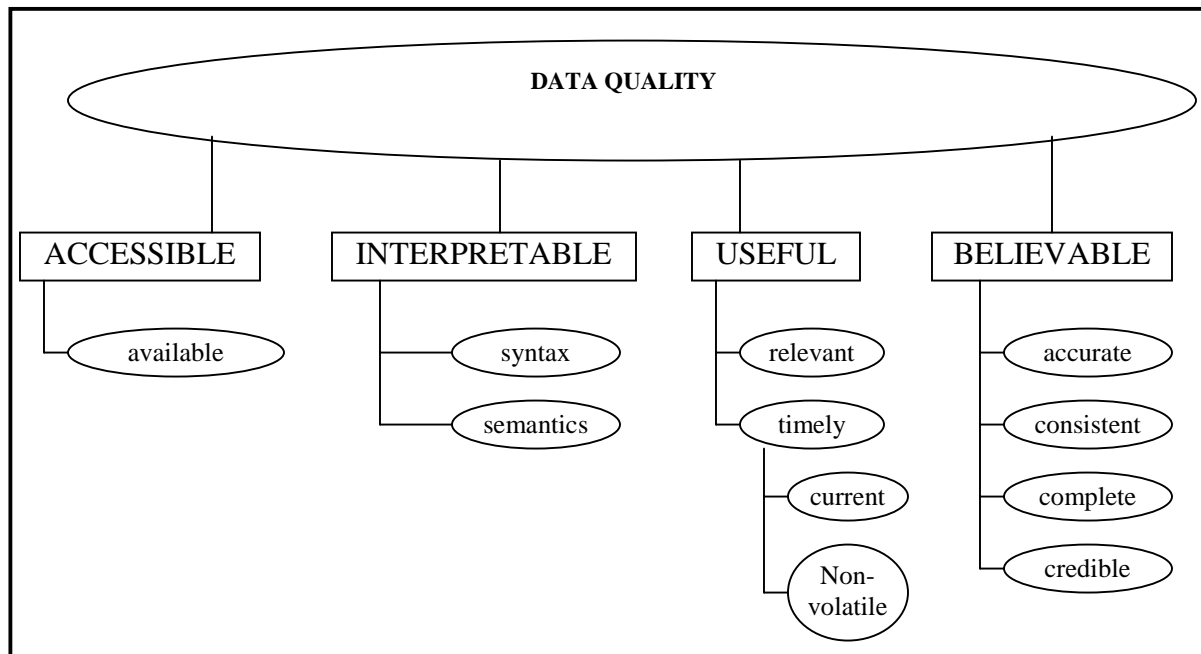


Figure 2: Wang and Strong (1996) Model of Data Quality

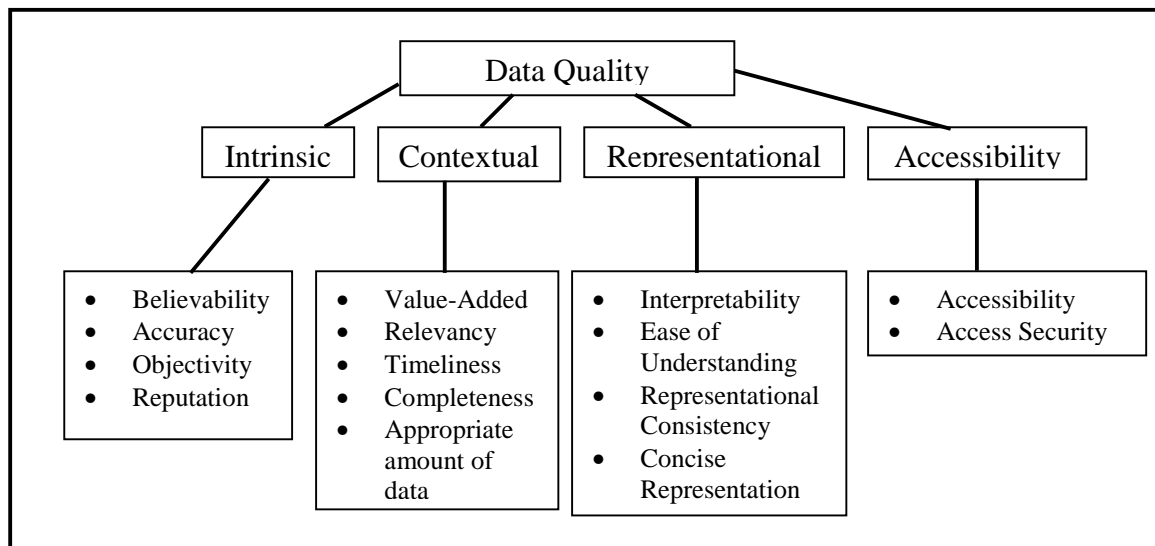


Figure 3: IQ Model Proposed in the Present Study

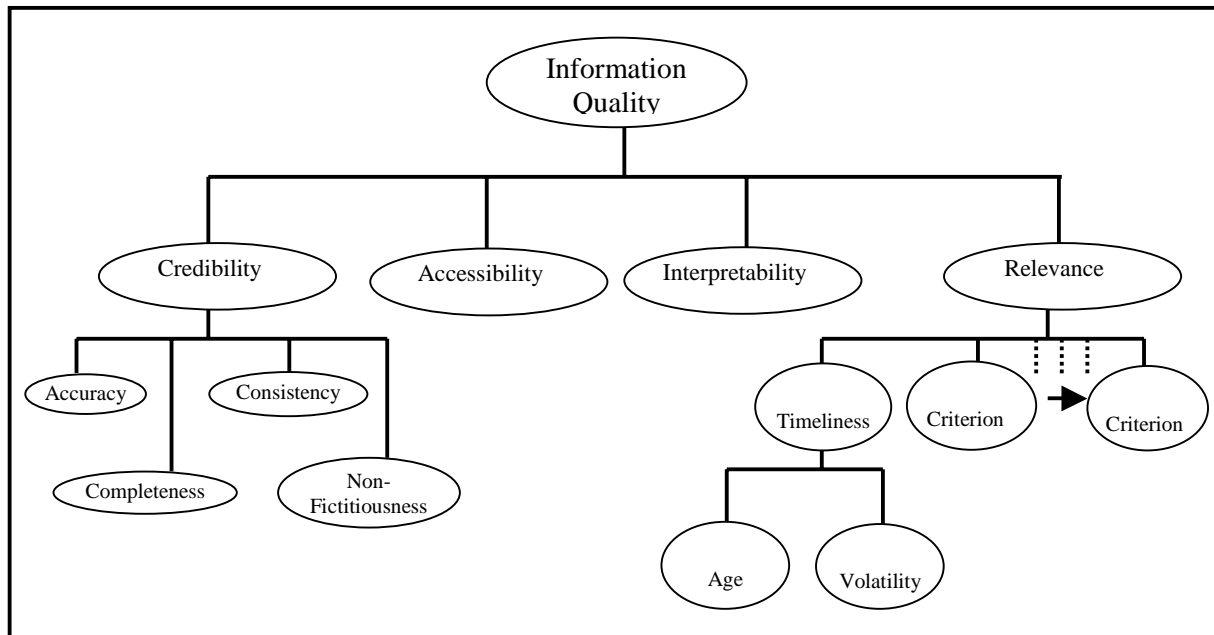
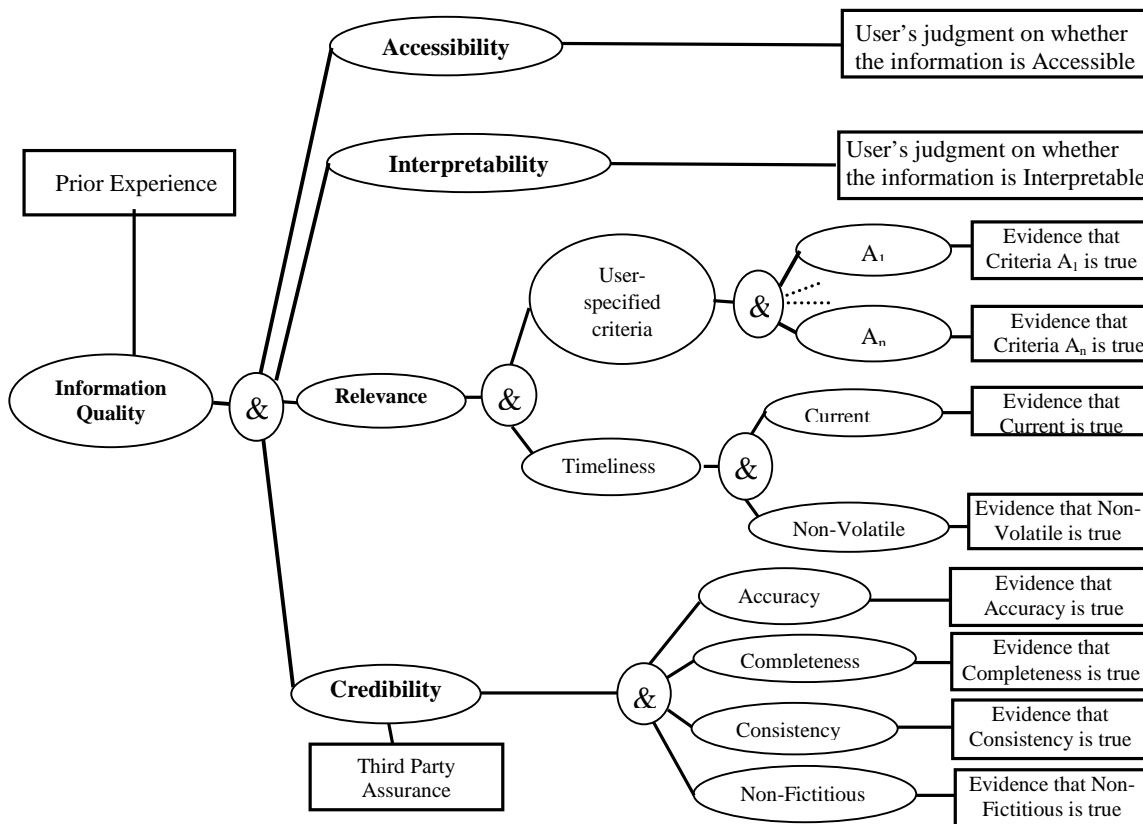


Figure 4. Evidential Network of Information Quality Attributes and Elements



**Table 4: IQ Attributes and Their Elements with Explanations and Examples**

Attribute	Elements	Sub-Elements or Cases	Explanation/ Definition	Example
Interpretability	Intelligibility	--	Capable of being understood, apprehended or comprehended.	A routine hospital report of the results of a patient's physical examination should be legible and intelligible. If, by accident, it were printed in ASCII code it would not be, even though it still contained the same information.
	Meaningfulness	--	If intelligible, the information has some minimum level of meaning <i>to the user</i> . The meaning content may be increased by adding structure or organization.	A patient examination report printed as a continuous string of words and values may barely be meaningful, but organized into tabular format it becomes more easily interpretable and meaningful.
Accessibility (retrievable)	Time		How long it takes to retrieve the information	Time needed to assemble in-house patient test information; lag-time for Internet replies to search queries; download time for files
	Cost		How the user measures the cost of retrieving the information.	Manpower needed to gather and assemble the information; price charged for an information product or service



**Table 4, Continued: IQ Attributes and Their Elements with Explanations and Examples**

Credibility (plausible or believable)	Accuracy	Known	True or error-free w/respect to some known value	The recorded patient name matches the known patient name
		Assigned	True or error-free w/respect to some designated or assigned value	The recorded patient number matches the assigned patient number
		Measured	True or error-free w/respect to a measured value	The recorded patient blood pressure value is within plausible limits, normal ranges, or is corroborated by other patient information
	Completeness		All required parts present; all attributes needed are present; no missing records; some tolerance for missing values	Patient information typically includes name, age, sex, treatment, and payment details, plus the results of various visit-specific tests. Interpretation of the results may be impaired if any are missing.
	Consistency	Discrete	Same value across all cases	A male patient should be recorded as a male in all departments and for all tests within a hospital
		Continuous <sub>1</sub>	Same value across multiple occurrences	A single measurement of a patient's blood pressure, recorded in multiple places should be the same in all instances
		Continuous <sub>2</sub>	Tightly dispersed values across multiple measures	Blood pressure measured multiple times w/in a short time should be close to some average of the true value
	Non-Fictitiousness	Records	No false or redundant records exist	No patient record should be completely identical to any other; each patient record should represent an actual patient hospital visit
		Attributes	No false or redundant attributes exist	No patient attribute should be completely identical to another; each patient attribute should represent an actual patient attribute
		Values	No false values exist	All values for patient attributes should be actual
Relevance (user domain- and purpose)	User-specified	A <sub>1</sub>	User-specified attributes derived from domain- and purpose-specificity	OSHA
		A <sub>2</sub>		The hospital
		A <sub>3</sub>		The department
		A <sub>4</sub>		The doctor
		A <sub>n</sub>		The patient
	Timeliness	Currency	Recentness of collection	Blood pressure may be measured annually or continuously
Volatility		How long it remains valid	For general health check-ups, annual blood pressure readings are sufficient; for surgery it needs to be monitored continuously.	

## A Generic Framework for Information Quality in Knowledge-intensive Processes

Martin J. Eppler

Institute for Media and Communications Management (=mcm *institute*)

University of St. Gallen

Blumenbergplatz 9

9000 St. Gallen

CH Switzerland

Phone: +41 71 224 24 07

Fax: +41 71 224 27 71

E-mail: [Martin.Eppler@unisg.ch](mailto:Martin.Eppler@unisg.ch)

**Abstract:** Based on an evaluation of existing information quality frameworks, action research with seven partner companies, and empirical surveys among practitioners, this paper proposes an information quality framework that overcomes some of the analyzed deficits of current approaches. The proposed framework is especially apt for the context of knowledge-intensive processes, such as market research, product development or consulting. It consists of four quality levels, namely community or relevance criteria, product or soundness criteria, process criteria, and infrastructure criteria. A main element of the framework are information quality principles which indicate how the criteria contained in the four levels of the framework can be improved. The paper outlines the basic elements of the framework and indicates how it can be applied in various knowledge-intensive processes. Two short case studies are provided, one from a market research company, and one from a book abstraction service. The paper concludes by pointing out future research needs in the domain of information quality frameworks, such as finding measurable and instructive quality indicators on all four levels.

### 1. Introduction: The Rationale for a New Information Quality Framework

*Organizational scientists should be viewed not as engineers offering technical advice to managers but as providers of conceptual and symbolic language for use in organizational discourse.*

*(Astley, Zammuto, 1992, p.443)*

In a prior study related to information quality frameworks, we have analyzed information quality frameworks from the last twelve years and found that they required further development in five areas in order to be useful for practitioners and researchers alike (Eppler, 2001). These improvement areas are:

1. The applicability of the frameworks in more than just one area, e.g. improving their *generic* nature.
2. The development of information quality frameworks that show *interdependencies* between different quality criteria (such as the accuracy timeliness tradeoff).
3. The inclusion of *problem areas* and *indicators* into these frameworks, as well as the inclusion of possible *solution* elements.

4. The development of *tools* which are based on the framework.
5. The development of frameworks that are at the same time *theoretical and practical* (that are at the same time rigorous and relevant, elaborate yet concise).

With the following information quality (IQ) framework, we try to overcome these shortcomings. We strive for a *generic* framework (in the sense of wide applicability) that can be used for any kind of knowledge-intensive process that has information both as an input and an output factor (this means information for direct human – and not direct machine – use; for a definition of knowledge-intensive processes see Eppler et al.)<sup>1</sup>. Generic also means that the framework can be used for various purposes, such as evaluation and assessment, improvement, management or monitoring. The framework will explicitly show tradeoffs between specific criteria (drawn as arrows that connect two criteria). It can be used to position information quality problems, and it consists of four principles which help to find solutions to IQ-problems (the integration, validation, contextualization, and activation principle). It will be shown that the framework is rooted in an existing theory (the media reference model) and that it can be used to provide tools such as checklists, diagnostic questionnaires or information quality guidelines. The framework is both tested against existing information quality theories and in terms of its practical applicability in the information management domain.

The goal of the present framework is neither prognosis nor precise description. It should help (as Porter points out in his discussion on the use of conceptual frameworks) to better *think through a problem* and select among strategic alternatives. Frameworks in this understanding identify the relevant variables and the questions which an analyst must answer in order to develop conclusions tailored to a particular company (see Porter, 1991, p. 955). Frameworks thus provide a *conceptual language* which practitioners can use to facilitate their mutual problem understanding and coordinate their collaborative actions (for this point, see Astley & Zammuto, 1992, p. 443). In the information quality context, a framework should thus provide a systematic and concise set of terms which practitioners and researchers can use to analyze and resolve information quality issues. The existing information quality frameworks do not provide that conceptual language or terminology since they focus on data problems in the data warehousing or information systems context, and not on the use of information by knowledge workers. This crucial distinction is illustrated in the table below. Table 1 compares typical data quality problems with those that are addressed in the context of knowledge-intensive processes.

---

<sup>1</sup> We define a knowledge-intensive process as a productive series of activities that involve information transformation and require specialized professional knowledge. Knowledge-intensive processes can be characterized by their often non-routine nature (unclear problem space, multiple decision options), the high requirements in terms of continuous learning and innovation, and the crucial importance of interpersonal communication on the one side, and of documentation (or codification) of information on the other.

Data Quality Problems	Information Quality Problems
Duplicates, multiple data sources	Conflicting recommendations or expert opinions in a study or analysis
Missing data relationships	Unclear causal effects in a diagnosis
Garbling (meaningless entries)	Wordy reports that have no logical flow
Spelling Errors	Cluttered language that contains grammatical errors
Obsolete or outdated entries	An analysis is not updated according to recent discoveries or changes in the organizational context
Inconsistent data formats or naming conventions	Inconsistent layout conventions or navigation structures
Misplaced data	Lost or 'buried' evidence
Complicated query procedures for a database	Difficult information navigation and retrieval in a knowledge base
Wrong data coding or tagging (adding wrong meta-data)	Inadequate or insufficient categorization (insufficient meta-information or contextual attributes) due to a lacking taxonomy
Incorrect data entries because of lacking source validation	Unsubstantiated conclusions with inadequate evidence
Manipulation of stored data (unauthorized deletion or modification of entries)	Manipulation of decision processes (overloading, confusing, diverting attention)

Table 1: Data Quality versus Information Quality Problems

Whereas data quality problems can be resolved through data cleansing algorithms, data profiling programs, stabilization algorithms (e.g., phonetic manipulation and error correction), statistical process control, or dictionary matching routines (see Strong et al., 1997, or Agosta, 2000), information quality problems can often not be solved through automated processes. They require (as do some data quality problems) fundamental analysis of business issues or questions, a change in work practices or process designs, an analysis of the involved information community and its expectations and skills, an evaluation of the relevant knowledge domains and its attributes, as well as an evaluation of the content management process and infrastructure. Typical remedies for information quality problems may include information design guidelines, publishing policies, authoring training, source validation rules, the purchase of additional information services and infrastructures, a re-design of the review and feedback process, etc.

Having outlined *why* a new information quality framework is necessary for the described context, we can now analyze *how* such a framework can be established.

## 2. Legitimacy of the Framework: Six Validation Areas

We believe that a *prescriptive* or normative framework such as this one, which does not describe how something does work or will work, but how it *should* work, can be legitimized or validated in the following six ways. They are at the same time the six empirical and theoretical bases of the framework presented in this study.

1. Through a solid existing *theory* which is used to generate the framework and in-/exclude certain elements: in this case the four levels or views of the framework are based on the

knowledge media theory of BEAT SCHMID, chair of communication management at the University of St. Gallen (see Schmid 2000).

2. Through *feedback from practitioners* about its usefulness: the framework presented in this paper has been discussed with practitioners in the fields of consulting, market research, corporate communications, and product development over the course of three years. It has been used to analyze real-life information quality problems and improve information products. The practitioners were given the chance to provide feedback on the framework's design and its components.

3. Through *comparisons with other frameworks* and their deficits: the current framework has been developed based on an analysis of more than twenty information quality frameworks from the last twelve years (see Eppler, 2001).

4. Through the evaluation of the framework through *meta criteria* (see Roehl 2000): The criteria which were used to evaluate the current framework are the following six: 1. Precision (all included terms are clearly defined) 2. Positioning (the context of the framework is made clear) 3. Consistency (the elements of the framework are mutually exclusive and collectively exhaustive) 4. Conciseness or parsimony (the framework uses a relatively small amount of elements) 5. Illustration (the framework can be illustrated through examples) 6. Practicality (the framework can be used as a tool to improve real-life problems) (see Eppler, 2001).

5. Through the use of *empirical surveys*:<sup>2</sup> the relevance of the included criteria has been tested by several surveys among employees of different companies in the consulting and market research context (see Brocks, 2000). In these surveys, the importance of the included criteria has been rated by the practitioners in relation to the process steps that they are involved in. As a result of these surveys, some information quality criteria were no longer included in the framework (such as believability), while others were added to it (such as applicability).

6. Through multiple case studies: As Eisenhardt (1989, p. 545) has pointed out, a conceptual framework can often be the result of case study research. Hence, case studies are a feasible way to see whether a framework does in fact fit with the reality of the corporate world or not. Two such case studies are presented in the last section of this paper. Those and others from the domains of consulting and financial services have contributed to the present framework.

Having stated the six validation possibilities that exist for a conceptual (prescriptive) framework, we can now turn to the framework itself and its elements.

### **3. Elements of the Framework: Views, Phases, and Principles**

The present framework consists of three major elements: The first element is the framework's vertical structure. It consists of *four views* on information quality that categorize crucial information quality criteria according to their relation to the target community, the information product, the information process, and to its infrastructure. The second element of the framework is the horizontal structure, which is divided into *four phases*. The four phases represent the life cycle of information from a user's point of view: it is searched and found, evaluated, adapted to a new context, and applied. The third major element of the framework are the *management*

---

<sup>2</sup> See for example the empirical research referenced in: Huang, Lee, Wang, 1999.

*principles*. They help to improve the quality of information in every phase. Below, these three major elements of the framework are described in more detail.

The overall *vertical* structure of the present framework is derived from an implicit convention which most recent information quality frameworks seem to follow, namely that they are divided into two sections: a category section, and a criteria (or dimensional) section. Thus, in most IQ-frameworks, the individual quality criteria are grouped into fewer information quality categories. These categories typically do not include qualifiers, but often have standard names such as intrinsic IQ (Wang & Strong, 1994) or quality of structuring (Königer & Reithmayer, 1998). In the current framework this twofold structure is also used, but with qualifying category names that already include a quality criteria on a higher level. The four IQ-categories or views are:

1. *Relevant information*: This category relates to whether the information is comprehensive enough, accurate enough, clear enough for the intended use, and whether it is easily applicable for the problem at hand. This category is also called *community view* since the relevance of a piece of information depends on the expectations and needs of a certain (writer-, administrator-, or user-) community.
2. *Sound information*: This second category contains criteria which describe the intrinsic or *product* characteristics of information, such as whether it is concise or not, consistent or not, correct or not, and current or not. Whereas the criteria in the first category (relevance) are subjective (indicated through the term “enough”) these should be relatively independent of the targeted community (indicated through the term “or not”).
3. *Optimized Process*: The third category contains criteria which relate to the content management process through which the information is created and distributed and whether that process is convenient (for writers, administrators, and users), and whether it provides the information in a timely, traceable (or attributable), and interactive manner.
4. *Reliable Infrastructure*: The fourth and final category contains criteria which relate to the infrastructure on which the content management process runs and through which the information is actually provided. Reliability in this context refers to a system’s easy and persistent accessibility, its security, its maintainability over time (including aspects of cost-efficiency), and its high (and continuous) speed or performance.

The logic behind these four categories or IQ-views is based on the knowledge media theory of BEAT SCHMID (see Schmid 2000). It states that any knowledge media (in the sense of a platform that enables the transfer of knowledge) design must begin with the analysis of the community of people who need to share knowledge, and analyze their needs, activities, and work practices. Then, the services and information objects which have to be provided to that (and by that) community need to be analyzed and a process has to be designed in order to deliver these information services or information objects. Only then can the infrastructure requirements and parameters be determined. Thus, the following framework is usually used in a top-down approach.

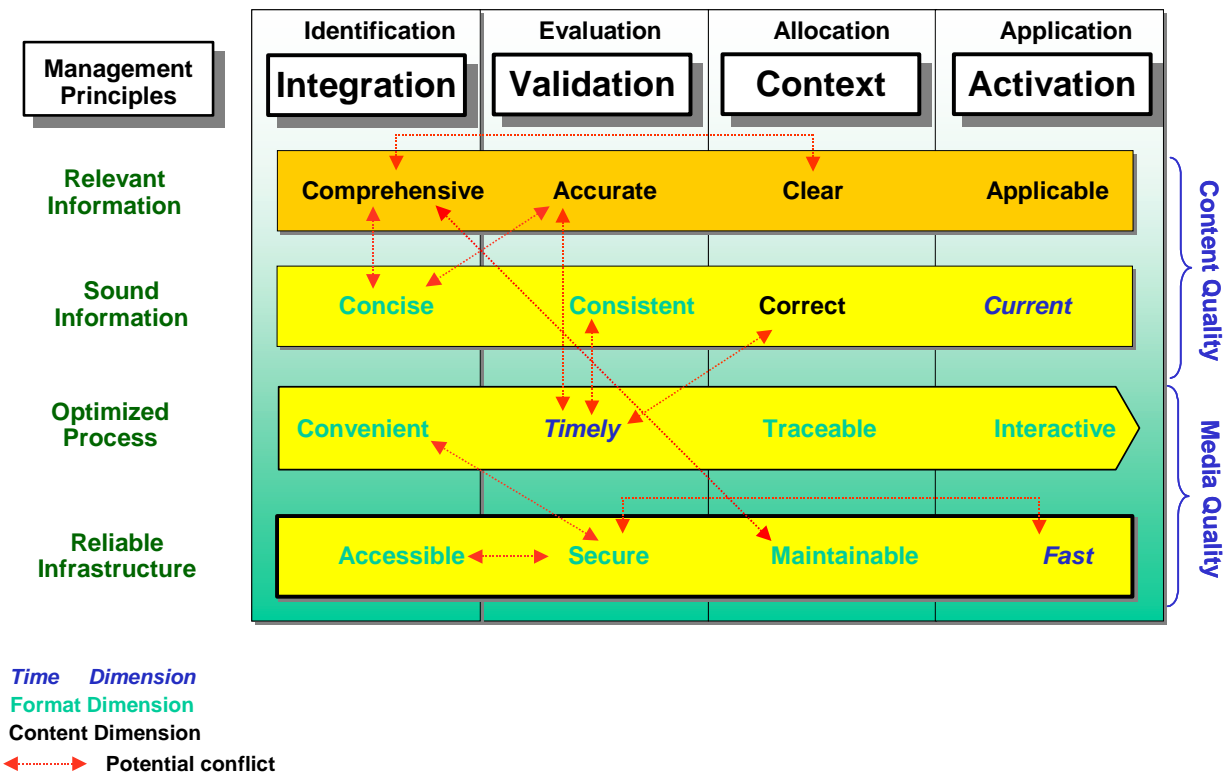


Figure 1: The Information Quality Framework

As figure one shows, the upper two levels of the framework are labeled as content quality, while the lower two are referred to as media quality. The first two categories, relevance and soundness, relate to the actual information itself, hence the term *content quality*. The second two categories, process and infrastructure, relate to the management of that information, and whether the delivery process and –infrastructure are of adequate quality, hence the term *media quality* which stresses the channel in which information is transported. For the end-user, both segments, media and content quality, may be perceived as one final product – the information and its various characteristics. For the information producers and administrators, however, this difference is crucial, since the authors usually cannot influence the media quality, and the administrators only have limited possibilities of influencing the content quality.

The *horizontal* structure of the framework incorporates a chronological sequence (or phases) from the user’s point of view. For him (or her) information may be the answer he needs to find, understand and evaluate, adapt to his context and apply in the right way. Thus, a knowledge media should assist him in identifying relevant and sound information. It should help him to evaluate whether the information is adequate for his purpose. It should assist him in re-contextualizing the information, that is to say understand its background and adapt it accordingly to the new situation. Finally, the knowledge media should provide assistance in making the found, evaluated, and allocated information actionable, e.g., use it effectively for decision making. In terms of the key questions of an information consumer which are answered in each phase, they can be described as follows: 1. Where is the information I need? (identification) 2. Can I trust it (evaluation) 3. Can I adapt it (allocation) 4. How should I best use it (application)?

As stated earlier, the current framework cannot only be used as a systematic arrangement of crucial information quality criteria, or as the key questions of users, but also as a systematic problem lens. The four phases of the framework can be used to designate four dominant information quality problems, namely: information overload (information is not integrated), information misjudgment (information is not validated), information misinterpretation (information is not seen in context or contextualized), and information misuse (information is not made actionable).

The third and final element of the framework are the management principles.<sup>3</sup> As mentioned, they provide pragmatic help in implementing the framework and achieving the quality criteria contained in it. The principles are also placed *vertically* along the framework since they follow the same step-by-step logic as the four phases discussed above. Every principle relates to the criteria that are found in the same column as the principle.

The *integration principle* states that high-quality information has to be aggregated or compressed (made comprehensive, concise, convenient, and accessible) in order to give the information consumer an overview before details are presented. The application of this principle should make it easier to identify relevant and sound information quickly because information is no longer distributed in various sources and formats. Means of applying this principle are abstracts (content summaries), visualization (e.g. maps or matrices), categorization or taxonomies (e.g., hierarchical content trees), prioritization, or personalization (as in an intranet portal, where information sources are integrated based on a personal user profile). The main IQ-problem that is resolved through this principle is *information overload* or the fact that information is no longer acknowledged but ignored or stored away (and thus loses relevance or impact) when it is fragmented, prolix, inconvenient, or inaccessible.

The *validation principle* states that high-quality information has to be validated (in terms of correctness, consistency, timeliness, and security) in order to present only justified information to the information consumer and that the validation mechanisms that lie behind a piece of information be made visible. Means of applying this principle are consistency-checks on the information itself, comparisons with other sources ('second opinions'), an analysis of the primary source of the information (its reputation and competence), and a rating mechanism and rating scale that makes the degree of validation of the information visible (and gives information consumers the chance to provide feedback on the perceived quality of the information). The main IQ-problem that is resolved through this principle is *misjudgment* of (incorrect, inconsistent, late, or manipulated) information.

The *context principle* states that high-quality information is always presented with its context of origination and its context of use (where did it come from, why is it important and to whom is it important, how should it be used). Through this, the information should become clearer for the target group, because it can understand the information's background. The target group can also better assess whether the information holds true for the new context and if it is correct even under different circumstances. The context principle should also assure that the information is traceable, that is to say that its various origination steps can be traced back to the original source (this

---

<sup>3</sup> According to Merrill "a principle is a proven, enduring guideline for human behavior. It is a relationship that is always true under appropriate conditions regardless of program or practice. See: Merrill (1987).



criteria is also known as attributability). Finally, the context principle refers to the infrastructure in which information is stored. This infrastructure should not be neglected, but maintained to serve in future contexts. Means of applying this principle are adding meta-information<sup>4</sup> (such as author, reviewer, origination and expiration dates, target group, etc.), referring to similar pieces of information or to people who have used the information, and referring to prior information of the same kind. The main IQ-problem that is resolved through this principle is the *misinterpretation* (and hence misallocation) of information.

The *activation principle* states that high-quality information provides means of activating the information in the mind of the information consumer and thus renders it memorable and consequently easily applicable for later use. The activation principle strives for greater user acceptance by making the information as applicable and current as possible and by providing it in an interactive and fast manner. Specific means of applying this principle are repetitions of crucial information elements, mnemonics (cognitive shortcuts such as abbreviations), stories (vivid plots which make the information more memorable), metaphoric language and metaphoric visualizations, check questions for the information user, simulations or animations that make the information come alive and motivate the information consumer to actively explore and use it, etc. The main IQ-problem that is resolved through this principle is often referred to as *paralysis by analysis* or the fact that information is often not stimulating or motivating actions or decisions, but rather delaying them. A generic term for this information quality problem is information *misuse*.

These four principles are an integral part of the framework. In consequence, the framework can help the analyst to not only think about the crucial information characteristics (the individual information quality criteria) and their inherent conflicts or tradeoffs, but also about how these characteristics can actually be improved. The figure below summarizes how the four principles can be applied in knowledge-intensive processes.

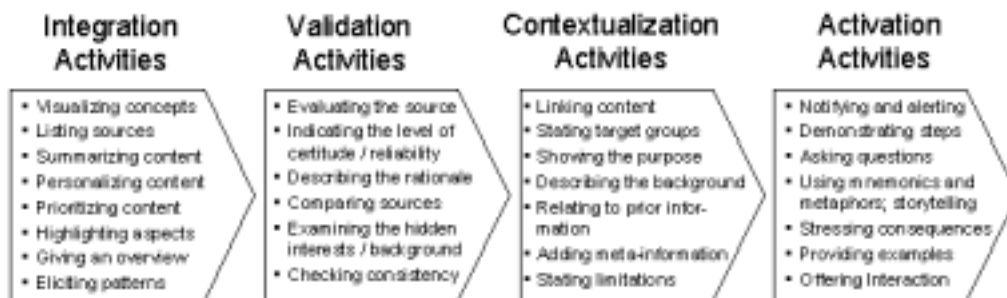


Figure 2: Ways of Implementing the Four Principles

Having described the main elements of the framework in overview, we can now turn to the specific information quality criteria and their relationships and see what they mean in specific knowledge-intensive processes.

<sup>4</sup> This improves the clarity of the information, its traceability, but also its maintainability for information administrators.

#### 4. Discussion of the Included Criteria and their Relationships

In this section, the logic of the individual criteria and their relation to the category names are explained. The potential conflicts between the individual criteria are also discussed in this segment.

The logic for the criteria contained in the first level is the following: *Relevant information* is information that is adequate for the community that requires it. Adequate in this context signifies that the scope (or breadth) of the information is right (comprehensive enough), that the precision and level of detail is sufficient (accurate enough), that the clarity of argumentation is sufficient (comprehensible, interpretable, or clear enough) and that the information is easily applicable for the target community.

The criteria of the second level follow this rationale: *Sound information* is information that has certain intrinsic (product) characteristics which make it of high quality independently of the community that deals with the information. The information can be said to be sound if it does not contain superfluous or non-related elements (conciseness), if it is internally consistent (does not contradict itself and uses the same format conventions)<sup>5</sup>, if it does not contain errors or false elements (correctness), and if it is not outdated by more recent information (currency).

The criteria of the third level all relate to information as a process: We refer to the information delivery process as an *optimized information process* if the following conditions are met: the information creation, administration, and delivery is as convenient as possible for the three information stakeholders (author, administrator, user), e.g., there are no superfluous or tedious steps; the access to the information is secure in the sense that both the information and the user are protected against unauthorized manipulations; the information is continuously maintained (cleansed, updated); and the way the information is accessed or retrieved can be adapted to one's personal preferences through interactive elements.

The criteria of the fourth level all deal with infrastructure requirements: For an information *infrastructure* to be *reliable*, it is important that it is always accessible (no down-times, otherwise the information itself is not accessible), that it is secure (protected against unauthorized access or information manipulation), that it is maintainable (that is to say that the information can also be accessed easily in the future), and that it enables a fast interaction between the stored information and the users (or the authors or administrators). Infrastructure in this framework does only relate to the hardware and operating system software of information systems. An information infrastructure can be any kind of channel that allows for information to be shared, such as a paper archive, a library, a documentation center, an intranet, a war room like control center, a television studio etc.

As the framework in figure one indicates, all criteria of which the framework consists relate to at least one dimension of either *time*, *content* or *format*. The first level of information quality, the relevance criteria, contain subjective notions of information quality that mainly relate to the content of the information (although one could argue that applicability is also a question of

---

<sup>5</sup> For this understanding see also Kahn & Strong, 1998, which view soundness as an IQ-category for criteria such as free-of-error, concise representation, completeness, consistent representation.

format and timing). The second level, the soundness criteria, contain criteria of all three dimensions, since information has to be sound in terms of format, content, and time aspects. Again, one could argue that criteria such as consistency are not only related to a consistent format of the information, but also to a consistent content (that it is free of self-contradictions). While this is certainly true, it is often easier to detect inconsistencies in the format than in the content. For the other criteria in this group, the dimension seem obvious. The last two levels of the framework do not contain any content criteria since they relate only to the media quality and not the content quality of the information. The process and the infrastructure can directly influence the format (the presentation) of the information and its timely delivery. The process can only indirectly affect the content quality of information, for example through rating and validation workflows that double-check the information before it is published or distributed.

As far as possible *tradeoffs* between individual criteria are concerned, one can argue that the most critical criteria are comprehensiveness, timeliness, security, conciseness and accuracy, since they provide the most potential conflicts with other criteria. A tradeoff in this context refers to a possible goal conflict, that is to say when the increase in quality in one criteria leads to a decreasing value in another. One tradeoff that has been discussed in the information quality literature is the accuracy-timeliness tradeoff (see Ballou & Pazer, 1987), which consists of a choice for either accurate or timely information, since the very fact of a timely delivery often impedes the careful consideration of accuracy issues. A similar tradeoff exists between timeliness and correctness: the faster the information has to be delivered, the less time is available to fully check its correctness for a given context. The same may be said for consistency and timeliness: the faster information has to be delivered, the less time can be spent on improving its format and content consistency. Another tradeoff that may exist is between accuracy (in the sense of precision or level of detail) and conciseness: the more accurate information is, the less concise is its presentation. This tradeoff is similar to the one between conciseness and comprehensiveness: the greater the scope of the information, the more difficult is its presentation in a concise format. The quest for comprehensive information may lead to less clarity, since the increased scope decreases the clear distinction between central and peripheral information and thus makes information more difficult to comprehend. A high level of comprehensiveness also makes the infrastructure more difficult to maintain, since more information objects need to be updated (or removed) on the infrastructure. The tradeoffs related to the security criteria are threefold: First, there is a potential conflict between convenience and security, since many security measures lead to inconvenient authorization procedures for information producers, administrators, or consumers. Second, there is a clear conflict between providing quick access to an information infrastructure and keeping the infrastructure secure. A typical example of this tradeoff in the computer context is the number of times one has to enter passwords to access a certain information system. Third, there may be a tradeoff between the speed of an information infrastructure and its security, since security measures require additional resources which in turn may slow down the functioning of an information infrastructure.

Making these (and other) tradeoffs visible can help the designer of an information system or information product in his interaction with information consumers and authors, since it shows them the *constraints* under which one has to operate. In the context of consulting and market research, we have used the tradeoffs in the framework to show clients that is not possible to request a report that is delivered within two weeks (timeliness), contains no errors whatsoever (correctness and consistency), has a high level of accuracy and is very comprehensive and at the same time not more than fifteen pages (conciseness). Finally, the tradeoffs can also show

differences between various user groups of information: while one group may require information in a very comprehensive format, another information consumer group may require the same information in an extremely concise format (due to time constraints).

Table one summarizes the discussed categories or levels, the information quality criteria, as well as their antonyms.

<b>Information Quality Levels</b>	<b>Information Quality Criteria</b>	<b>Opposites</b>
Community Level (Relevance)	1. Comprehensiveness	Incompleteness
	2. Accuracy	Inaccuracy
	3. Clarity	Obscurity
	4. Applicability	Uselessness
Product Level (Soundness)	5. Conciseness	Prolixity
	6. Consistency	Inconsistency
	7. Correctness	Falsity
	8. Currency	Obsolescence
Process Level (Optimization)	9. Convenience	Inconvenience
	10. Timeliness	Lateness
	11. Traceability	Indeterminacy
	12. Interactivity	Rigidity
Infrastructure Level (Reliability)	13. Accessibility	Inaccessibility
	14. Security	Exposure
	15. Maintainability	Neglect
	16. Speed	Slowness

Table 2: Information Quality Criteria and their Opposites

As a result of this juxtaposition, we can define the antipode of quality information as follows:

Low quality information is incomplete, inaccurate, obscure, useless, prolix (or wordy), inconsistent, false, obsolete, delivered in an inconvenient, late, undeterminable and rigid way, on an infrastructure that is inaccessible, exposed to manipulation and other security risks, not maintainable, and slow.

Having described the framework and its logic, we can now turn to its application. For this purpose, we provide short case studies that show how two companies' efforts to improve information quality can be analyzed with the help of the information quality framework.

### **5. Case Studies on Information Quality Improvements: IHA-GfM Market Research and getAbstract**

In section two of this paper, we stated that one way of validating<sup>6</sup> a conceptual framework such as this one consists of applying it in real world cases. This should reveal the usefulness of the framework as a systematic lens for information quality problems and solutions.

---

<sup>6</sup> The term validation in this context does not relate to the truth value of the framework, but rather to its applicability.

The first company case study deals with the knowledge-intensive process of **market research**. From 1998 to 2001, we have collaborated with a leading Swiss market research company called IHA-GfM (www.ihagfm.ch), a subsidiary of the international market research group GfK. The organization in Switzerland has over 300 employees and thus qualifies as a medium-sized company. The main products of this company are market research reports, market statistics, and market and (food-, non-food, near-food-, pharmaceutical, and media-) product analysis tools, such as media monitoring tools, category management tools or sales analyzers. In working together with the company, we have analyzed its *information process* (from the client briefing and first offer to survey construction, survey use, survey codification and analysis, to the final survey interpretation and client feedback) through workshops and interviews with the specialists of the company, as well as the final *information product* (the market reports) and its *infrastructure* (such as the company's client extranets). Since this company's main products are in fact information products, the quality of information is a crucial competitive component. Until our involvement with the company, however, the quality of information was mainly viewed as accuracy, consistency, correctness, timeliness and currency. But in 1999 (as market data became more and more of a commodity), the company understood that it could only enter a higher margin business if other quality criteria started to become relevant, such as applicability, convenience, conciseness, clarity, or maintainability. This, however, also meant a change in the qualification of its staff, who – up until then – were mostly trained in statistical analysis and not in information design and effective client communication (we will return to this important point when describing the specific improvement activities).

In six workshops with the company's project managers<sup>7</sup> we gathered and analyzed the challenges in the area of information quality that they saw (and realized because of their client satisfaction surveys) and we tried to find ways to improve the identified deficits. The reoccurring themes or challenges in terms of information quality were the following five issues:

1. The timeliness of the information that the company provided to its clients was seen as sometimes inadequate (still too many market research reports were not delivered on-time).
2. The accessibility and convenience of the information for clients was judged to be insufficient (it was argued that the new media were not yet fully used for the benefit of the clients).
3. The applicability of the information for clients was seen as a great improvement area (here it was argued that more added-value needed to be provided with the market data, such as benchmarks, comparisons, trend analyses, recommendations, consulting services etc.)
4. Finding the right scope or level of detail (for a market research report) in order not to overload clients was seen as a constant challenge.
5. Because of the relative autonomy of the various units of the company, the project managers considered it a major challenge to provide information in a consistent structure and layout.

Because of these problems, the company, decided to launch two projects to improve its competitiveness in this area. One project was launched to increase the quality of information and its sharing internally (labeled as *Knowledge Management*), the other project was launched to increase the value of information for clients (labeled as *Value 2000*).<sup>8</sup> The specific measures that

---

<sup>7</sup> The first such workshop was held in January of 1999 with about fifteen participants. The last was held in March 2001 with twelve participants.

<sup>8</sup> The knowledge management project was sponsored by the head of human resources, while the value 2000 project was directly sponsored by the CEO of the company.

were taken are summarized in the table below, where they are listed with the IQ-criteria that they affect most.

Information Quality Levels	Information Quality Criteria	Activities to improve the IQ-criteria
Community Level (Relevance)	Comprehensiveness	In order to increase the comprehensiveness of the information provided to clients, the company entered a joint-venture with Mediametrix to enlarge its scope to web user data.
	Accuracy	No specific measures were taken since the present label was seen as sufficient. The company had implemented a (certified) quality management system for its processes earlier.
	Clarity	Different layout templates were introduced that should make the information clearer and more easily interpretable. Presentations were introduced to make the information contained in market research reports clearer to the client.
	Applicability	In addition to just reporting the market data, reports now include interpretations of the data, further analysis and cross references, and recommendations for action. The reports are not only presented, but discussed with the client to determine its internal use. The project managers are trained in consulting tools in order to improve the impact of the gathered information.
Product Level (Soundness)	Conciseness	All reports now include executive summaries. Many reports have the statistical information in the appendix and focus on the key results.
	Consistency	All market reports that a client receives have a similar structure, layout and logic.
	Correctness	No specific measures were taken in regard to correctness.
	Currency	The use of adhoc on-line surveys was intensified to provide more up-to-date consumer data to clients.
Process Level (Optimization)	Convenience	The market report is not only delivered as a document, but also as PowerPoint slides, as a CD-ROM, and in the future also in an updated form on the client extranet.
	Timeliness	Pre-tests were intensified in order to eliminate time lags or possible errors early on in the process.
	Traceability	A knowledge map was developed and put on the company's intranet which makes it possible to trace back any tool or method to a tool owner or tool specialist.
	Interactivity	Clients are given more opportunities to provide input (via briefings, e-mails, presentations, telephone conferences etc.) during the information gathering process, before the report is finished.
Infrastructure Level (Reliability)	Accessibility	The client extranet can be accessed from any computer with internet access anytime of the day or night.
	Security	The client extranet is protected through a password and a hidden link. The intranet is protected through a firewall.
	Maintainability	Specialized key accountants are assigned to the client extranets where the market reports are updated or cleansed.
	Speed	No specific measures were taken since the available infrastructure was seen as fast enough.

Table 3 : Implemented Information Quality Improvement Activities at IHA-GfM

One key insight (regarding the information quality framework) from this implementation period at IHA-GfM emerged as more people were involved in the endeavor: As far as the momentum of

implementation was concerned, the great number of crucial information quality criteria was a disadvantage. Even a framework that would consist of only seven or eight core criteria would still make it difficult to focus the workforce on the necessary improvements and to drive the change process at a company. Because of this insight, we started to rely on the aforementioned four IQ-principles that are aimed at improving the IQ-criteria, rather than on the criteria themselves. Applying this insight to the situation at IHA-GfM, the following results were achieved:

- *Integration*: almost every major market report now contains a concise and systematic executive summary and an on-site presentation where the main consequences of a market study are presented in overview.
- *Validation*: almost every market research report contains an appendix that explains (in detail) how the data was gathered, analyzed and condensed.
- *Contextualization*: most market reports now refer to related reports or provide links to other available information or benchmarks that may render the information more meaningful by putting it into perspective and enabling comparisons.
- *Activation*: many market reports are now not just sent to the client or just presented in a management meeting, but actually discussed and analyzed in a workshop-like setting where clients can ask questions or probe deeper together with the consultants of the market research company. Most resources were spent on this aspect, especially in the area of training and tools.

The last issue from the above list points at a second key insight that emerged from the experience at IHA-GfM, namely that information quality improvements do not always require major information system changes. One of the key activities to improve the relevance dimension of information at IHA-GfM was **training** project managers on how to tailor and activate information for their clients through better presentations and better information visualization.

The second company case study deals with the knowledge-intensive process of **keeping up with relevant management know-how**. It was not gathered through participatory research like the case study on market research, but through a four hour interview with the CEO of the company, through being a client of the company for four months, and through a document and website analysis. The analyzed company is getAbstract.com Inc., the analyzed information products are book abstracts.

GetAbstract is a knowledge compression and rating company based in Lucerne, Switzerland with additional offices in Fort Lauderdale, Paris, Hamburg, Beijing and Hong Kong. It has about twenty full-time employees, and a network of 120 part-time collaborators. It is, according to the company, a “leading provider of compressed knowledge” mainly in the area of business books. GetAbstract states its mission as follows:

To get the latest business trends and knowledge into the hands and heads of executives, managers and business students worldwide through concise Abstracts (summaries) of the newest and most important books on the market.

The company was founded in 1998 and incorporated in 1999. It has received major funding from two Swiss banks and various institutional and private investors. The company provides its (about

3000) private and more than sixty corporate clients and subscribers with concise, five-pages abstracts, available in four languages and across multiple platforms (as Adobe PDF files or PalmPilot files sent by e-mail, as audio abstracts, or in the form of a repository as a part of a company’s intranet). The abstracts of current business books are written by a network of 120 professional writers and edited by three professional editors in Switzerland and the United States. These editors also rate each book according to its overall appeal, applicability, innovation, and style. The revenues of the company are generated by annual subscriptions of 299 US\$ per client (this fee provides access to all available abstracts and to one new abstract every week automatically sent by e-mail) or through its corporate clients who integrate the service (with its library of over 1200 abstracted books) into their intranets.

Every book abstract has the same consistent structure: A thumbnail of the book next to the title and publishing information, a half-page of key take-aways, a rating on a scale of five to ten (books which receive a rating below five are not summarized), a one paragraph long review and recommendation, the actual abstract itself (about three pages long), and information about the author(s). The last two lines of every abstract provide a list of buzz-words used in the book.

On its web-page, GetAbstract provides full access to all abstracts which are categorized in so called knowledge channels, such as leadership, strategy, or technology. One can also search the database of summaries through a keyword search or browse a list of top downloads or new abstracts. Table three lists the main functionalities of the getAbstract service and shows which information quality criteria are influenced by it.

<b>Information Quality Levels</b>	<b>Information Quality Criteria</b>	<b>Functions of GetAbstract.com</b>
Community Level  (Relevance)	Comprehensiveness	At the book level: the most important elements of the book are represented in the abstract. At the portal level: most general management bestsellers are summarized and about 8000 business books are screened per years.
	Accuracy	The accuracy of the provided information is determined by the editorial guidelines that are provided to the writers and by the quality of the writers who are mostly professional journalists.
	Clarity	Only professional writers are hired to write abstracts. Professional editors review the abstracts for clarity and style.
	Applicability	Regular feedback from abstract users is acknowledged and incorporated. The abstracts focus on take-outs and main new terms.
Product Level  (Soundness)	Conciseness	Every book abstract is limited to five pages. Reviews are limited to one paragraph per book.
	Consistency	Every book abstract has the same structure: take-away, rating, author, buzz words. Authentic quotes from the book are included on the side.
	Correctness	The book abstracts are corrected by a team of editors.
	Currency	New book summaries are added every week, about five hundred new books are summarized every year.
Process Level  (Optimization)	Convenience	The book abstract can be simply clicked at and is directly mailed to the inbox of the client where it can be read as a PDF-file or on the palm-pilot.
	Timeliness	New abstracts that fit the profile of a client are automatically sent out by e-mail.
	Traceability	Author and publisher information is always given. However, the reviewer’s name is not disclosed.



	Interactivity	The getAbstract client can interactively edit his account and his interest profile. He can browse various book categories or do a key word search.
Infrastructure Level (Reliability)	Accessibility	All abstracts are accessible all of the time from any computer with Internet-access.
	Security	The getAbstract site is protected by a firewall. The client account is protected through a password.
	Maintainability	The site is continually updated and improved by a team of technical experts.
	Speed	The response time of the server seems immediate.

Table 4: Characteristics of GetAbstract Services that Increase Information Quality

Major competitors of getAbstract are Summaries.com (which offers the most inexpensive book summaries and delivers them over the Internet) Soundview's summary.com – whose thirty summaries per year are longer than getAbstracts but also available on tape and in print - and meansbusiness.com which has a significantly lower number of book abstracts in total, but also provides concise summaries of content across various books in its so called concept suites. A concept suite is a summary of book chapters from various books that deal with the same topic. The biggest entry barriers for this type of service are the legal obstacles (getting publishers to agree to the book abstraction) and finding qualified writers who can provide consistently instructive abstracts over a long period of time.

Again, we can summarize GetAbstract's major benefits or innovations with the four information quality principles:

- *Integration*: GetAbstract integrates on two levels: every book is reduced to five pages, and a great number of books is integrated in one web site.
- *Validation*: GetAbstract pre-selects and filters new business books and rates them according to a defined set of criteria.
- *Contextualization*: GetAbstract provides information on the author and his or her background, it states possible target groups of a book, and it will add references to similar books or to books others have found to be useful.
- *Activation*: GetAbstract stresses the key take-outs of every book and provides them through a push-mechanism to the reader (based on a user-defined profile).

The getAbstract case study has shown that high-quality information (in this case especially in the area of conciseness, comprehensiveness, and convenience) may be considered as a growing industry in its own right. The case study has also shown that a framework like the one presented in this paper can indicate possible future market niches, such as the one discovered by getAbstract to increase the conciseness of business knowledge in the forms of books.

## 6. Conclusion and Outlook

In this last section of the paper, we summarize the central findings in the first paragraph and provide an outlook to future research steps in the second one.

To increase the quality of information, it has to be targeted at a specific community in order to be relevant, it has to be managed as a product (with intrinsic qualities that we call soundness) and as a (continually optimized) content management process, and the platform on which information is provided has to be managed in order to be reliable. We refer to this reasoning as the four views or levels of information quality: the *community* level, the *product* level, the *process* level, and the *infrastructure* level. To manage the quality of information one has to pay attention to the author's, administrator's, and user's point of view and to their specific needs. One has to be aware of the potential conflicts between various information quality criteria and make these constraints visible for information consumers, authors, and administrators. The information user needs to be able to find and access the information (*identification* phase), he needs to be able to assess the information (*evaluation* phase), he has to be able to see the information in context and adapt it to his specific situation (*allocation* phase), and he has to be able to use the information for decision making or other applications (*application* phase). In order to assure that this is possible, certain management activities must take place at every one of the four described levels and in every one of the described phases. One can summarize these management or value-adding activities with the help of four principles: the identification-, validation-, contextualization-, and activation-principle. These principles make it easier to communicate and implement an information quality improvement program (versus having to explain a great number of criteria).

This paper has to be seen as an element of a larger research project. This project consists of finding an adequate application context for information quality, evaluating existing information quality frameworks, finding their improvement areas for the examined context, devising a modified framework for the application context, and illustrating and applying the framework with the help of documented case studies. As of now, the first four steps of this research project have been completed: 1. An application context has been defined, namely knowledge-intensive processes. 2. Existing information quality frameworks have been screened and evaluated according to specific meta-criteria. 3. Five deficits have been identified. 4. A new framework has been proposed. What remains to be done at this point, is to fully document the researched case studies and show how the framework's application can actually improve information quality in real-life situations. A specific challenge in this endeavor will be the compilation of a realistic set of indicators that make information quality improvements in knowledge-intensive processes measurable.

## **References**

- Agosta, L. (2000) Data Quality Methodologies and Technologies, Giga Information Group, Planning Assumption, August 30, 2000. URL: [www.gigaweb.com](http://www.gigaweb.com) [14.08.2001]
- Astley, W.G.; Zammuto, R.F. (1992) Organization Science, Managers, and Language Games, in: *Organization Science*, 3: 4, pp. 443-460.
- Ballou, D. P.; Pazer, H. L. (1987) Designing Information Systems to Optimize the Accuracy-Timeliness Tradeoff, in: *Information Systems Research*, 6(1), pp.509-521.
- Brocks, M.: *Wissensarbeit im Consultingcontext: Die Rolle der Informationsqualität*, Diplomarbeit MBE-HSG, St. Gallen: Universität St. Gallen, 2000.
- Eppler, M., Röpnack A., Seifried P.: Improving Knowledge Intensive Processes through an Enterprise Knowledge Medium, in Prasad, J. (Ed.): PROCEEDINGS OF THE 1999 ACM SIGCPR CONFERENCE Managing

Organizational Knowledge for Strategic Advantage: The Key Role of Information Technology and Personnel, 1999, pp. 222-230.

Eppler, M.: The Concept of Information Quality: An Interdisciplinary Evaluation of Recent Information Quality Frameworks, in: *Studies in Communication Sciences*, No. 1, 2001, pp. 167 -182.

Eisenhardt, K.M. (1989) Building Theories from Case Study Research, in: *Academy of Management Review*, Vol. 14, No. 4 (October), pp. 532-550.

Huang, K.-T.; Lee, Y.W.; Wang, R.Y. (1999) *Quality Information and Knowledge*. New Jersey: Prentice Hall.

Kahn, B. K.; Strong, D. M. (1998) Product and Service Performance Model for Information Quality: An Update, in: Chengalur-Smith, I.; Pipino, L. L. (1998) *Proceedings of the 1998 Conference on Information Quality*, Cambridge, MA: Massachusetts Institute of Technology.

Königer, P.; Reithmayer, W. (1998): *Management unstrukturierter Informationen*, Frankfurt.

Merrill, M. D. (1987) The New Component Design Theory: Instructional design for Courseware Authoring in: *Instructional Science*, 16, pp. 19-34.

Porter, M.E. (1991) Towards A Dynamic Theory of Strategy, in: *Strategic Management Journal*, Vol. 12, pp. 954-117.

Roehl, H. (2000) *Instrumente der Wissensorganisation*, Wiesbaden: Gabler.

Schmid, B. (2000) *Knowledge Media*, St. Gallen: NetAcademy Press.

Strong, D. M., Lee, Y. W., Wang, R. Y. (1997) 10 Potholes in the Road to Information Quality, in: *Computer IEEE*, 30(8), pp. 38-46.

## **A College Course: Data Quality in Information Systems**

Craig W. Fisher  
Marist College  
Poughkeepsie, NY 12601  
(845) 575-3000 x2621  
[Craig.Fisher@Marist.edu](mailto:Craig.Fisher@Marist.edu)

**Abstract:** Information Systems (IS) college students are not prepared for the demands of improving information quality in our databases and data warehouses. Information Quality (IQ)/Data Quality (DQ) in IS curricula tends to be subject to individual faculty preferences. Given the significance and pervasiveness of DQ and the lack of focused attention to DQ/IQ in college curricula, I created a *Data Quality in Information Systems* course. The purpose of the course is to alert our would-be-IS-professionals to the pervasiveness and criticality of data quality problems. The secondary agenda is to begin to arm the students with approaches and the commitment to overcome these problems.

Four major exercises were quite powerful in helping the students understand and assimilate concepts of Data and Information Quality. The cases included Total Quality Management analysis, Data Warehouse Cleansing, Data Quality Information (data tags), exercise and an Information Quality Assessment project. The purpose of this paper is to share the essence of the new course.

### **Introduction**

Khalil et al (1999) states that Information Systems (IS) college students are ". . . not equipped with a broad understanding of the principles behind measuring, analyzing, and improving IQ in an organization." Their finding that Information Quality (IQ)/Data Quality (DQ) in IS curricula tends to be subject to individual faculty preferences seems to hold true at Marist College. At Marist College, our IS program contains a number of courses that touch on various aspects of data quality. Data quality is a topic of at least one complete (75 minutes) lecture in half of our ten IS courses, and is mentioned in all ten courses. However, the five courses that mention data quality do not test on data quality beyond simple definitions. The other five courses that lecture on data quality may or may not directly test on data quality. Typical test questions might indirectly reference data quality in topics such as software engineering, data validation methods, types of data errors, anomalies, testing, requirements documents, normalization and referential integrity.

Given the significance and pervasiveness of DQ problems (Redman, 1996, 1998, Tayi and Ballou, 1998; Orr, 1998) and the lack of focused attention to DQ/IQ in college curricula (Khalil, et al., 1999), I created a *Data Quality and Information Systems* course at Marist College. The purpose of the course is to alert our would-be-IS-professionals to the pervasiveness and criticality of data quality problems. The secondary agenda is to begin to arm the students with approaches and the commitment to overcome these problems. The purpose of this paper is to share the essence of the new course.

## The Students

The course is aimed at first semester seniors majoring in Information Systems who have completed eight of the ten required IS courses, including systems analysis, systems design, data management, data communication, and problem solving and programming. These students should be able to analyze an end-user procedural or process problem, design a system solution, and develop, test, and implement a solution. They have the knowledge and skills to clean up anomalies and ensure *third normal form* of databases. They are aware of, but not experts in, data warehouses. Ten seniors took the course in the Fall 2000 semester and fifteen are registered for the Fall 2001 semester.

## The Course

### Course Objectives

There were six basic objectives to the course.

1. Develop in-depth understanding of Data and Information Quality (DQ and IQ).
2. Understand DQ and IQ Concepts in Information Systems projects.
3. Recognize various patterns of data and design deficiencies in systems.
4. Suggest DQ and IQ improvement plans in light of known deficiencies in systems.
5. Understand of the role and importance of DQ and IQ in data warehouses.
6. Discuss the role and importance of DQ in Decision Support Systems.

### Course Approach

The class met once a week for 2 hours and 45 minutes including a 15-minute break. Students were asked to read and study text and journal articles to learn the fundamental concepts of Data and Information Quality. There were four key exercises to explore Total Quality Management (TQM), use of data tags, data warehouse cleaning, and information quality assessments. These four IS and DQ exercises were designed to engage the students' new knowledge of the concepts with their analytical abilities so that they can develop meaningful solutions to real problems.

Evaluation of students consisted of two major exams, the midterm and final exams, with each exam counting for one quarter of their grade. The four exercises counted at 10% each, for a total of 40% of the grade. Classroom participation accounted for the final 10%.

The required textbook was Quality Information and Knowledge by Huang, Lee, and Wang, published in 1999 by Prentice Hall. The complete list of recommended journal articles appears in the bibliography of this paper.

### Semester Plan

**SESSION A**  
Classes 1,2 & 3

Motivation & Concepts  
Manage Information as Product  
TQM Case Study

**SESSION B**  
Classes 4 & 5

Measure, Analyze, Improve IQ  
But Who Will Use Measurements?  
DQI & DSS Exercise

<b>SESSION C</b> Class 6	Midterm
<b>SESSION D</b> Classes 7,8, 9 & 10	DQ & Data Warehouse (DW) DW Exercise Knowledge as Assets
<b>SESSION E</b> Classes 11,12, 13 & 14	IQA Study "Sell" Clients, Conduct Surveys, Analyze Data & Debate Results Present & Discuss with Clients
<b>SESSION F</b> Class 15	Future Organizational Knowledge Conclusions, Feedback
<b>SESSION G</b> Class 16	Final Exam

### ***Elaboration and Feedback of Sessions***

#### **Session A: Introduction to Quality**

This session was scheduled to require three weeks of class, but consumed four weeks.

**Class 1** cannot be pro forma since it is a 2 hour and 45 minute time slot and should not be wasted; yet the students have not prepared any readings. I gave more pure lecture than is normally desirable. References to a variety of specific data quality issues as raised by Redman (1996, 1998), Kingma (1996), Wilson (1992), Tayi and Ballou, (1998), Orr (1998), Huang, Lee & Wang (1999), and Celko (1995) provided much of the motivational statement. In addition, we spent at least half an hour on the Space Shuttle *Challenger* and another half an hour on the USS *Vincennes* shutdown of a passenger jet (Fisher, 2001). A key aspect of this introduction was asking the students to provide examples of data quality problems from their daily lives. They were most enthusiastic about credit card charges being in error, their names being misspelled, college registration errors, and the like. The Class 1 lecture proved to be successful as the students were astounded at the significance and pervasiveness of data quality problems. They seemed motivated to study further.

**Class 2** focused on Chapters 1 and 2 of the text (Huang, Lee & Wang, 1999) in detail and covered *The Malcolm Baldrige National Quality Improvement Act of 1987 - Application Guidelines 1988* (Public Law 100 – 107, 1987). Students readily grasped Huang's two basic

propositions<sup>1</sup> and several concepts such as "best practice" and "core competency." The students struggled to grasp the concept of managing information as a product due to limited experience in that area. We spent a profitable amount of time discussing the text section *Establish an Information Quality Program* (p. 27 - 28), and we debated whether or not "information quality is 'everybody's responsibility'" (p. 28).

The TQM Case Exercise<sup>2</sup> describes a corporation that is having trouble integrating information systems properly. However, the basis of the problem may be the company's approach to, or lack of an approach to, TQM. Critical ingredients included environment, goals, standards, responsibilities, manufacturing processes & steps, and relationships with customers and vendors. The assignment was to read and analyze the case (16 pages), answer a series of questions, develop and analyze alternative solutions, and recommend a solution. The students enjoyed discussing the facts, symptoms, and problems of the case in Class 2, and the alternative solutions in Class 3.

In **Class 3**, we continued the discussion of the TQM case for close to one hour, about 30 minutes longer than I planned. We established two teams of three and one team of four to provide informal presentations to the class. For the second half of the class we discussed Chapter 3 of the text. I underestimated the time needed for this useful chapter. The students learned the concept of recognizing data, design, and operational deficiencies in mapping real world scenarios to information systems. While some examples are given in the text, I believe that more complete examples or a full case study would be useful.

I next introduced the 16 dimensions of data quality. These are grouped into four categories as follows:

- *Intrinsic IQ* – Accuracy, Objectivity, Believability, Reputation
- *Contextual IQ* – Relevancy, Value Added, Timeliness, Completeness, Amount of Information
- *Accessibility IQ* – Access, Security
- *Representational IQ* – Interpretability, Ease of Understanding, Ease of Manipulation, Concise Representation, Consistent Representation

In addition to these groupings, I may consider additional groupings in the future. For example, the model of product and service presented by Khalil et al (1999) groups the dimensions by specification and expectations. I believe the students would gain more insight by analyzing the two different groupings. The students also read Strong et al's 1997 article on *Data Quality in Context*.

## Session B: Measure, Analyze and Improve Data Quality

This session was planned to cover two weeks but consumed three weeks.

The concept of placing additional fields (data tags) in a database to carry data quality information (DQI) was foreign to the students. Thus the discussion and elaboration of DQI concepts consumed more time than planned. Recent research has demonstrated that simply telling people the quality of their data doesn't predict whether they will use that information

---

<sup>1</sup> Proposition 1: Firms must create a reservoir of quality information. Proposition 2: Firms must create a wealth of organizational knowledge. (Huang, et al. 1999. p. 4).

<sup>2</sup> Found in the casebook for McLeod, R J., *Management Information Systems: A Study of Computer-Based Information Systems*. 6 ed. 1995, Englewood Cliffs, NJ: Prentice Hall.

about their data quality (Chengalur-Smith, et al, 1999; Fisher, 1999). The students were asked to complete the Apartment Selection Task (Appendix A) to determine the amount of use, if any, that a person makes of data quality information<sup>3</sup> (DQI) when it is provided (Chengalur-Smith, et al, 1999; Fisher, 1999).

I gave the task with DQI to half the class and the task without DQI to the other half of the class. In a discussion before the task activity, many students said that they believed that if people were told the quality of their data that it would influence their decisions. Thus, it was very surprising to them that they did not use the reliability data in the task. I presented results from other studies that indicate experience, manager status, and time pressure influence the amount and type of use made of DQI. The students' first hand experiences with the task made it easy for them to follow.

Most of the students were now aware of the problem that some people may not use DQI if provided, and those who do use DQI may vary in their usage from others who use it. There are many implications here that we alluded to but did not investigate. Some examples of these implications are the "business case" for DQI data, training employees as to the purpose and uses of DQI, format of DQI, and the need to explore characteristics of decision makers and their familiarity with the data.

The three classes of metrics (perception, application independent, application dependent) were clear and readily grasped by the students. We continued studying the 16 dimensions of data quality (Huang, et al., 1999). I found that I made too many assumptions about how much the students would grasp. A more thoroughly planned set of examples and preplanned questions would have made this session more productive.

### Session C: Midterm

### Session D: Data Quality and Data Warehouses

Since data warehouses are becoming critical in corporations and government and are so fraught with data quality problems, the next logical step in the course was to walk through a data warehouse cleaning exercise. The students read Fayyad's *Data Mining and Knowledge Discovery* (1996); Golfarelli and Rizzi's *A Methodological Framework for Data Warehouse Design* (1999); and Ballou and Tayi's *Enhancing Data Quality in Data Warehouses* (1999).

Pamela Neely (2000) has been doing research on applying financial auditing processes to processes for cleaning data warehouses. She agreed to participate in two of my classes. At the first class, she presented a data warehouse problem with which she was involved. In the second class, the students built a dictionary and merged individual files into a data warehouse.

This exercise is an actual current project in a major city in New York state where a number of homeless shelters are receiving financial support. Each homeless shelter collects its own data, but recently the shelters have been asked to provide data for a comprehensive data warehouse. We gave the students definitions of five different databases that were designed to collect data on people<sup>4</sup> who frequent specific homeless shelters. The assignment was to build a common data dictionary and then build an integrated data warehouse. We asked the students to identify and resolve problems and issues in building the data dictionary. Finally, the students were asked to document errors found in the merging of the databases into the data warehouse (Neely, 2000).

---

<sup>3</sup> Data tags

<sup>4</sup> All names were changed for these exercises



The reality of the possible errors that occurred and the trouble that the students had in merging first the dictionaries and then the data was extremely educational. This is a must-do exercise for a course in data quality.

The article by Golfarelli and Rizzi (1999) proved to be too technical and abstract for our purposes. However, the Ballou article was particularly well received by the class.

#### Session D: Real Client Information Quality Assessment (IQA)

The students were asked to work in teams to perform an Information Quality Assessment study at Marist College (Appendix B). I prearranged these studies with the Vice President of Business/Financial Affairs, the Registrar, and the Director of Information Technology. I had three teams and I asked each team to give a kick-off presentation to one of the three stakeholders. The presentation described the value that the organization would receive from an IQA study and what it would require to complete the study. I underestimated the effort to use the IQA system, so the students prepared hardcopy survey forms. They then conducted the IQA Survey, collected the data, put it into an Excel database, analyzed it, and prepared a final report for presentation to the client (Huang, et al., 1999). We used two class periods to analyze, discuss, and debate various findings, and a third class for the final presentation to the clients.

The students gave a final presentation to the clients. The students compared and contrasted IQ perceptions from each of the different stakeholder organizations and from managerial versus worker perspectives. Finally, the students presented their conclusions and recommendations.

Each client agreed to come to a class session for the final presentation. Thus all students saw all teams present and discuss the project with their clients. These sessions were the high points of the course. The clients were interested in the results and also elaborated on various problems that they experienced with data flows and information quality. For example, the VP of Business/Financial Affairs explained the data dependencies between Housing, Registration, and Billing.

While the purpose of this paper is not to elaborate on specific conclusions, I mention some of the specific findings to give the reader a flavor of the students' results.

The IQA study included 38 Information Technology people, 14 Business Office people and 12 Registrar people. Chart 1 illustrates a comparison of the perceptions of all employees of the IT function versus all employees of the Business Office for a specific set of business databases. High numbers represent more satisfaction while low numbers represent more dissatisfaction. It is clear from Chart 1 that the IT people have a much more positive view of the quality of the system than do the Business Office people. IT people rated timeliness a very high 8.2, while the Business Office people rated timeliness at a mediocre 5.5. Differences were also found in ease-of-manipulation, consistency, accuracy, and value-add. Our class study identified potential problem areas and opened up discussion between the two organizations. The students recommended that SQL projects be started to address manipulation and timeliness. They recommended Quality Circles be started to address consistency, accuracy, and value-add.

Chart 1 is just one example of our findings. We also discussed managers versus non-managers, all users-all systems versus all IT-all systems, and all registrar versus all IT. Each group for each set of application databases ranked the dimensions on importance. These findings, coupled with the user executives' feedback, were the highlight of the course.

## Summary/Conclusion

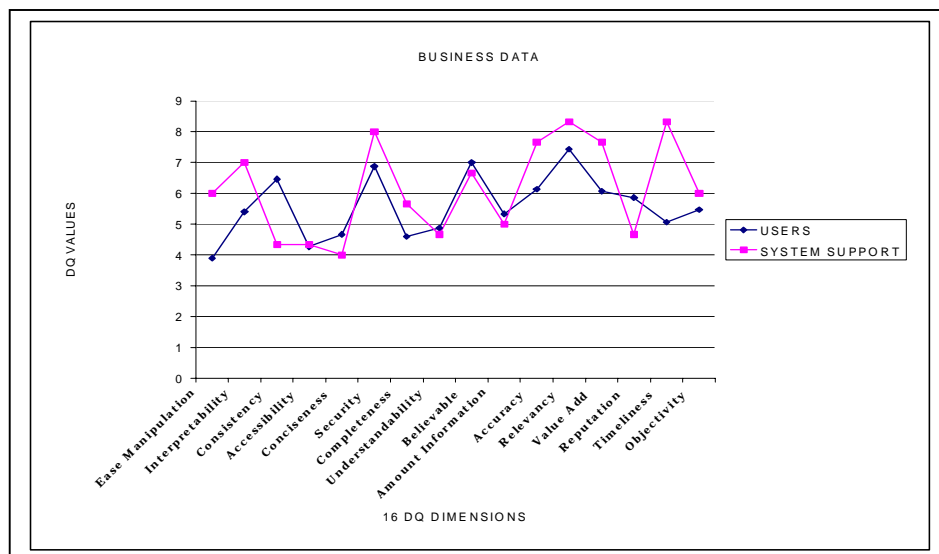
This first course in *Data Quality and Information Systems* met five of its six objectives. The combination of the text, the articles Strong et al (1997), Ballou & Tayi (1999), Redman (1998), Orr (1998), and the four major exercises were quite powerful in helping the students understand and assimilate concepts of Data and Information Quality. The first TQM case set the stage for understanding quality in context of information systems. Professor Neely's class exercise building a real dictionary and data warehouse from five fragmented and disperse databases was very exciting and informative. The IQA work was the highlight of the course. I cannot overstate the value of having the students develop and give the presentation that "sells" the value of doing an IQA study to the user executive. This was followed by conducting the survey, deciding how to enter data, analyzing the data, and presenting findings to the user with subsequent discussion.

The course partially met the last objective, which was to be able to discuss the role and importance of DQ in Decision Support Systems. We completed an exercise and had an interesting discussion about the use of DQI in decision-making. However, this is such a small part of DSS that I cannot claim we made much of an impact in this arena. We did not discuss decision theory, methods of decision-making, group decision-making, nor other factors in decision-making.

Of the ten students that took the course, nine completed the anonymous course/instructor evaluation form at the end of the semester. In the handwritten comment section, five of the students said that they appreciated the real-life exercises. They especially enjoyed the executives coming to class to discuss data quality issues. There was only one negative comment, and that related to the Knowledge Management section of the text. I feel that this was more an issue of time than a problem with the text; we fell behind in the schedule, but I still required that the students read all chapters. The students read the last two assignments without benefit of introduction from their teacher.

While the primary objectives were met in this course the real issue is whether they were the right objectives. This should be a subject for further research and discussion in IQ200x.

CHART 1



## Recommended Readings

1. Ballou, D. P., and Pazer, H. L. (1985). "Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems." *Management Science*, 31(2), 150 - 162.
2. Ballou, D. P., & Pazer, H. L. (1995). "Designing Information Systems to Optimize the Accuracy-timeliness Tradeoff." *Information Systems Research*, 6(1), 51-72.
3. Ballou, D. P., and Tayi, Giri K. (1999). "Enhancing Data Quality in Data Warehouse Environments." *Communications of the ACM*, 42(1), 73 - 78.
4. Bontempo and Zagelow, (1998), "The IBM Data Warehouse Architecture." *Communications of the ACM*, 41(9), 38 - 48.
5. Brachman, R., Khabaza, T., Kloesgen, W., Piatetsky-Shapiro, and Simoudis, E. (1996). "Mining Business Databases." *Communications of the ACM*, 39(11), 42 - 48.
6. Celko, J., and McDonald, J. (1995). "Don't Warehouse Dirty Data." *Datamation*, Oct.15.
7. Chengalur-Smith, I., Ballou, D. P., Pazer, H. (1999). "The Impact of Data Quality Information on Decision-Making: An Exploratory Analysis." *IEEE Transactions on Knowledge and Data Engineering*.
8. Davenport, T. H. (1997). *Information Ecology*. N. Y., NY: Oxford University Press.
9. Fayyad, U., and Uthurusamy, R. (1996). "Data Mining and Knowledge Discovery in Databases." *Communications of the ACM*, 39(11), 24 - 26.
10. Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). "The KDD Process for Extracting Useful Knowledge from Volumes of Data." *Communications of the ACM*, 39(11), 27 - 34.
11. Fisher, C. W. & Kingma, B. (2001). "Criticality of Data Quality as Exemplified in Two Disasters." *Information & Management*. Elsevier, Netherlands. (Note: paper accepted; proofs completed; publication date pending.)
12. Fisher, C.W., (1999). *An Empirically Based Exploration of the Interaction of Time Constraints and Experience Levels on the Data Quality Information (DQI) Factor in Decision-Making*, University at Albany: Albany, NY. 222.
13. Golfarelli and Rizzi. (1999). "A Methodological Framework for Data Warehouse Design." *DOLAP - ACM 1999*, 3 - 9.
14. Huang, Lee, and Wang. (1999). *Quality Information and Knowledge*. Englewood Cliffs, NJ: Prentice Hall.
15. Kaplan, D., Krishnan, R., Padman, R. and Peters, J. (1998). "Assessing Data Quality in Accounting Information Systems." *Communications of the ACM*, 41(2), 72 - 78.
16. Klein, B. D., Goodhue, D.L., and Davis, G.B. (1997). "Can Humans Detect Errors in Data? Impact of Base Rates, Incentives and Goals." *MIS Quarterly*, 21(June), 169 - 194.
17. Klein, B. D. (1998). *User Detection of Errors in Data: Learning through Direct and Indirect Experience*. Paper presented at the AIS98.
18. Khalil, O. E. M., Strong, D. M., Kahn, B. K., and Pipino, L. L. (1999). "Teaching Information Quality in Information systems Undergraduate Education." *Informing Science*, 2(3). p. 53 - 59.
19. McLeod, R. J. (1995). *Management Information Systems: A Study of Computer-Based Information Systems*. 6 ed. Englewood Cliffs, NJ: Prentice Hall.
20. Morrissey, J. M. (1990). "Imprecise Information and Uncertainty in Information Systems." *ACM Transactions on Information Systems*, 8(2), 159 - 180.
21. Motro, A., and Smets, P. (Ed.). (1996). *Uncertainty Management in Information Systems: From Needs to Solutions*. Kluwer Academic Publishers.{{missing address info (see

Prentice Hall in # 19}}

22. Neely, M. Pamela. (2000). "A Process for Auditing Source Data Quality in an integrated Data Repository: Development and Testing." *Dissertation Proposal*, University at Albany, Albany, NY.
23. Orr, K. (1998). "Data Quality and Systems Theory." *Communications of the ACM*, 41(2), 66 - 71.
24. Redman, T. C. (1995). "Improve Data Quality for Competitive Advantage." *Sloan Management Review* (Winter), 99-107.
25. Redman, T. C. (1996). *Data Quality for the Information Age*. Norwood, MA: Artech House, Inc.
26. Redman, T. C. (1998). "The Impact of Poor Data Quality on the Typical Enterprise." *Communications of the ACM*, 41(2), 79 - 82.
27. Sanbonmatsu, D., M., Kardes, Frank R., and Herr, Paul M. (1992). "The Role of Prior Knowledge and Missing Information in Multiattribute Evaluation." *Organizational Behavior and Human Decision Processes*, 51(1), 76 - 91.
28. Simpson, C., and Prusak, Laurence. (1995). "Troubles with Information Overload—Moving from Quantity to Quality in Information Provision." *International Journal of Information Management*, 15(6), 413 - 425.
29. Strong, D. M., Lee, Y. W., and Wang, R. Y. (1997). "Data Quality in Context." *Communications of the ACM*, 40(5), 103 - 110.
30. Tayi, G., and Ballou, D. P. (1998). "Examining Data Quality." *Communications of the ACM*, 41(2), 54 - 57.
31. Wand, Y., and Wang, Richard Y. (1996). "Anchoring Data Quality Dimensions in Ontological Foundations." *Communications of the ACM*, 39(11), 86 - 95.
32. Wang, R. Y., Storey, Veda C., and Firth, Christopher P. (1995). "A Framework for Analysis of Data Quality Research." *IEEE Transactions on Knowledge and Data Engineering*, 7(4), 623 - 639.
33. Wang, R. Y., & Strong, D. (1996). "Beyond Accuracy: What Data Quality Means to Data Consumers." *Journal of Management Information Systems*, 12(4), 5 - 34.
34. Wang, R. P. (1998). "A Product Perspective on Total Data Quality Management." *Communications of the ACM*, 41(2), 58 - 65.
35. Watson, R. T., Pitt, Leyland F., Kavan, C. Bruce. (1998). "Measuring Information Systems Service Quality: Lessons From Two Longitudinal Case Studies." *MIS Quarterly*, 22(1), 61
36. Public Law 100 - 107. (1987). The Malcolm Baldrige National Quality Improvement Act of 1987. Application Guidelines 1988.

## **APPENDIX A**

### **Apartment Selection Task**

**Number:** \_\_\_\_\_

#### **Task**

Your job requires you to move to a new city. You have a friend who lives there and you request her help in locating an apartment. You provide a list of criteria that are important to you and you have weights in mind for each criteria, which reflect their relative importance to you. Your friend gathers information about four potential apartment complexes and passes it along to you. She scores each apartment complex on each factor on a 50-point scale, such that a higher number is always more desirable. For example, a rating of 40 for rent expense is more desirable than a rating of 30.

Your objective is to choose the complex that overall performs the best. However, you realize that the data she obtained may not be completely accurate. For instance, she estimated commuting time by looking at the map. Also, the complex managers indicated that the rent quoted could increase at any time.

You decide to incorporate this uncertainty into your decision-making process by using a reliability measure where a score of 100 indicates perfectly reliable data and 0 scores completely unreliable data. The following table displays the reliability of the information provided about each criterion and the weights you assigned to each criterion.

<i>Criterion</i>	<i>Reliability</i>	<i>Weight</i>
Parking facilities	57	1
Commuting time to work	23	2.5
Floor space	76	2
Number of bedrooms	68	1.5
Rent expense	44	3

After multiplying the ratings by the weights for each criterion, you obtained the following results. The weighted scores are shown on the next page (for example, a score of 70 for commuting time is the result of multiplying its rating of 28 by its weight of 2.5).

Given the weighted scores, the objective is to choose the apartment complex that overall is the best (has the highest overall sum). Rank the apartment complexes in order of preference with 1 corresponding to the complex you would most prefer and 4 to the one you would least prefer. (Use all the information given to break any ties.) Recall that the reliability refers to the data and not to the weights. Next to each apartment complex write its rank, along with a brief explanation of how you arrived at the rank.

A	<i>Criterion</i>	<i>Reliability</i>	<i>Rating</i>	<i>Weight</i>	<i>Weighted scores</i>	<b>Rank = ____</b> <i>Explanation:</i>
	Parking facilities	57	22	1	22	
	Commuting time	23	28	2.5	70	
	Floor space	76	20	2	40	
	# of bedrooms	68	32	1.5	48	
	Rent expense	44	40	3	120	

B	<i>Criterion</i>	<i>Reliability</i>	<i>Rating</i>	<i>Weight</i>	<i>Weighted scores</i>	<b>Rank = ____</b> <i>Explanation:</i>
	Parking facilities	57	25	1	25	
	Commuting time	23	32	2.5	80	
	Floor space	76	31	2	62	
	# of bedrooms	68	36	1.5	54	
	Rent expense	44	36	3	108	

C	<i>Criterion</i>	<i>Reliability</i>	<i>Rating</i>	<i>Weight</i>	<i>Weighted scores</i>	<b>Rank = ____</b> <i>Explanation:</i>
	Parking facilities	57	28	1	28	
	Commuting time	23	27	2.5	67.5	
	Floor space	76	33	2	66	
	# of bedrooms	68	29	1.5	43.5	
	Rent expense	44	26	3	78	

D	<i>Criterion</i>	<i>Reliability</i>	<i>Rating</i>	<i>Weight</i>	<i>Weighted scores</i>	<b>Rank = ____</b> <i>Explanation:</i>
	Parking facilities	57	27	1	27	
	Commuting time	23	25	2.5	62.5	
	Floor space	76	35	2	70	
	# of bedrooms	68	38	1.5	57	
	Rent expense	44	37	3	111	

## **APPENDIX B**

### **Assignment: Information Quality Assessment**

**Objective:** Perform an assessment of Information Quality in certain systems at Marist College.

**Systems:** Various Business and Registration Systems.

**Personnel:** Management and staff from the Business Office, Registration Office and the entire Information Technology department.

**General Statement:**

Study and apply IQ Assessment concepts; key references include:

Text: *Quality Information and Knowledge* (Huang, Lee and Wang, 1999); and

Article: "Data Quality in Context" (Strong, Lee, & Wang, 1996).

**Teams:** You will work in three teams of three or four members. However, you will all need all of the data. I recommend that you divide the data entry and then simply merge the data into one large file and make copies of the file for each team to use.

**Steps:**

1. Prepare and administer IQA Survey.
2. Design Excel or Access DB to store the data. You must agree on the database design and I have supplied a suggested one. Feel free to modify it, but to arrive at one complete report you will need one database.
3. Populate your data storage with the actual data results.
4. Design and apply statistical evaluations of the data.
  - For example (but not limited to):
    - Averages, ranges etc for entire population on the 16 DQ dimensions
    - Rank the dimensions
    - Averages, ranges by subset of the population on the 16 DQ dimensions where subsets are organized by system and by type of respondent:
      - System:* Registration, Billing, etc.
      - Respondent:* Consumers, Custodians, Providers, Managers, etc.
    - Note: An analysis of the data may lead to different groupings, etc.
  - Rank the dimensions by subset(s)
  - Statistically compare various subsets (e.g., t-tests)
  - Use correlations to determine strength of relationships
  - Perform IQ Context Assessment (Sec 3 of IQA Survey)
5. Develop graphs to illustrate your findings.
6. Develop recommendations where possible from IQA and IQ context.
7. Reach conclusions and prepare professional write-up of those conclusions.
8. Prepare presentations of conclusions.
9. Give presentations to the "stakeholder" groups (or management).

## **Overview of an Approach to Data Quality**

**John Gimpert**  
Deloitte & Touche  
**Tim Krick**  
Deloitte & Touche

### **Executive Summary**

Data quality issues have been faced by a number of Deloitte & Touche (D&T) clients in a variety of contexts. Practical challenges arising out of these experiences motivated the development of an integrated approach to data quality. This approach includes three key phases: 1) Build a Foundation 2) Transform 3) Sustain.

Client examples (with names changed) are used to illustrate key points of each of these phases:

- 1) Build a Foundation -- Data definitions and business rules are discussed in the context of the real estate industry.
- 2) Transform -- Data cleansing can require a massive work effort, but risk assessment can be used to better prioritize work and allocate resources.
- 3) Sustain -- Planning and effort is required to support continued return on investments in data quality.

Lessons learned from these and other client experiences are discussed, including insights into business decisions around people, processes, and technologies.



## Overview of an Approach to Data Quality

**John Gimpert**  
**Tim Krick**

**Deloitte & Touche**

© 2001 Deloitte & Touche LLP. All rights reserved. Proprietary information. This document contains concepts, ideas and materials that are proprietary to Deloitte & Touche LLP and may not be used, copied, provided to others, or referred to without the express permission of Deloitte & Touche LLP.

## Discussion Topics

- Background and History
- Challenges Experienced in Practice
- An Integrated Approach
  - ❖ Building the Foundation
  - ❖ Data Quality Transformation
  - ❖ Sustaining Data Quality
- Lessons Learned

**Deloitte & Touche**

© 2001 Deloitte & Touche LLP. All rights reserved. Proprietary information. This document contains concepts, ideas and materials that are proprietary to Deloitte & Touche LLP and may not be used, copied, provided to others, or referred to without the express permission of Deloitte & Touche LLP.

## D&T's Data Quality and Integrity (DQI) Practice

- Our Genesis - supporting the complex data needs of audit teams
  - ❖ Data extraction
  - ❖ Data analysis
- Growth through support of system integration projects
- Broad client experience has confirmed:
  - ❖ Complexity of data quality issues
  - ❖ More than just supporting system integration
  - ❖ Need for a broad, integrated approach
  - ❖ Need to mobilize people with a variety of competencies

**Deloitte & Touche**

© 2001 Deloitte & Touche LLP. All rights reserved. Proprietary information. This document contains concepts, ideas and materials that are proprietary to Deloitte & Touche LLP and may not be used, copied, provided to others, or referred to without the express permission of Deloitte & Touche LLP.

## Clients

Recent clients served include:









**LEHMAN BROTHERS, INC.**

*Other:*

*Major Real Estate Organization*

*Major Securities Firm*

*Child Support State Disbursement Unit*



**Deloitte & Touche**

© 2001 Deloitte & Touche LLP. All rights reserved. Proprietary information. This document contains concepts, ideas and materials that are proprietary to Deloitte & Touche LLP and may not be used, copied, provided to others, or referred to without the express permission of Deloitte & Touche LLP.


## Challenges Experienced in Practice

- Need to address data quality issues at an enterprise level and at a detailed level
- Data quality issues cross organizational boundaries
- Executive commitment
- Data definitions
- Business rules – early and often
- Difficulty integrating disparate systems
- Prioritization of work efforts / addressing resource constraints
- Sustaining ROI

**Deloitte & Touche**

© 2001 Deloitte & Touche LLP. All rights reserved. Proprietary information. This document contains concepts, ideas and materials that are proprietary to Deloitte & Touche LLP and may not be used, copied, provided to others, or referred to without the express permission of Deloitte & Touche LLP.

## Approach to Data Quality



<div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;">Governance</div> <div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;">Build Business Case for DQ</div> <div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;">Define DQ Requirements</div> <div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;">Understand Business Process</div> <div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;">Understand Data Model</div> <div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;">Data Definitions and Business Rules</div> <div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;">Change Management</div>	<div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;">Sourcing/Mapping</div> <div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;">Data Risk Assessment</div> <div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;">Data Analysis</div> <div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;">Corrective Actions &amp; Cleansing</div> <div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;">Testing</div> <div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;">Conversion/Transformation</div> <div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;">Optimization</div>	<div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;">Design Continuous Monitoring Process</div> <div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;">Construct Continuous Data Monitoring Process</div> <div style="border: 1px solid black; padding: 5px; margin-bottom: 5px;">Implement Continuous Data Monitoring Process</div>
--	---	--

*Prior implementation challenges have led us to develop an integrated approach to addressing data quality issues.*

**Deloitte & Touche**

© 2001 Deloitte & Touche LLP. All rights reserved. Proprietary information. This document contains concepts, ideas and materials that are proprietary to Deloitte & Touche LLP and may not be used, copied, provided to others, or referred to without the express permission of Deloitte & Touche LLP.

### Building a Foundation: Importance of Data Definitions

**Example: Real Estate Company Implementing a Data Warehouse**

Governance
Build Business Case for DQ
Define DQ Requirements
Understand Business Process
Understand Data Model
Data Definitions and Business Rules
Change Management

**Issue:** Redundant and conflicting data definitions (e.g., several different definitions of "square footage")

- Different systems
- Different internal users and external requirements

**Impact:** Difficulty comparing & consolidating information

- Comparing information across enterprise
- Consolidating information within data warehouse

**Solution:** Business rule definition and analysis

- Rationalize data definitions around property maintenance and portfolio development (from 500+ to 150)
- Expand data definitions for square footage

© 2001 Deloitte & Touche LLP. All rights reserved. Proprietary information. This document contains concepts, ideas and materials that are proprietary to Deloitte & Touche LLP and may not be used, copied, provided to others, or referred to without the express permission of Deloitte & Touche LLP.

**Deloitte & Touche**

### Building a Foundation: Importance of Business Rules

**Example: Real Estate Company Implementing a Data Warehouse**

Governance
Build Business Case for DQ
Define DQ Requirements
Understand Business Process
Understand Data Model
Data Definitions and Business Rules
Change Management

**Issue:** Difficulty populating certain data fields, when there are more than 5 different levels of ownership. For example, need to report "primary" geographic location of a property.

**Impact:** Inefficiency without automated business rules

- Users making manual, time-consuming, and inconsistent decisions
- Rules can be complex

**Solution:** Business Rule analysis

- Test potential business rules & analyze impact
- Implement business rules within application

© 2001 Deloitte & Touche LLP. All rights reserved. Proprietary information. This document contains concepts, ideas and materials that are proprietary to Deloitte & Touche LLP and may not be used, copied, provided to others, or referred to without the express permission of Deloitte & Touche LLP.

**Deloitte & Touche**

### Transform: Importance of Risk Based Approach

**Example: Fortune 500 Client Implementing a New System**

Sourcing/Mapping
Data Risk Assessment
Data Analysis
Corrective Actions & Cleansing
Testing
Conversion/Transformation
Optimization

**Issue:** Client sourcing data from over 200 data sources and over 15 functional areas of the company

**Impact:** Difficult to prioritize and assess data quality risks

**Solution:** Using interviews and risk assessment templates, assessed risk for conversion & interface files based on key drivers

- Impact of Errors
- Likelihood of Errors
- Perceived Level of Control in Source System

© 2001 Deloitte & Touche LLP. All rights reserved. Proprietary information. This document contains concepts, ideas and materials that are proprietary to Deloitte & Touche LLP and may not be used, copied, provided to others, or referred to without the express permission of Deloitte & Touche LLP.

**Deloitte & Touche**

### Sustain: Importance of Continuous Monitoring

**Example: Major Securities Company concerned about ongoing Data quality of a CRM system**

Design Continuous Monitoring Process
Construct Continuous Data Monitoring Process
Implement Continuous Data Monitoring Process

**Issue:** Recent CRM implementation did not meet the users' needs, because system did not deliver reliable information.

**Impact:** Users lost faith in system, and the investment in the system was compromised.

**Solution:** Implement a new system, based on lessons learned.

- Begin with Foundation issues, such as buy-in and ownership
- Design infrastructure for maintaining high data quality
  - ❖ People & processes
  - ❖ System edits
  - ❖ Data quality dashboard for data quality group monitoring (Generates emails to users related to exceptions / fixes)
- Also perform detailed data analysis prior to implementation

© 2001 Deloitte & Touche LLP. All rights reserved. Proprietary information. This document contains concepts, ideas and materials that are proprietary to Deloitte & Touche LLP and may not be used, copied, provided to others, or referred to without the express permission of Deloitte & Touche LLP.

**Deloitte & Touche**

### Lessons Learned

- People
  - ❖ Broad spectrum of skillsets
  - ❖ Sustained focus important
- Process
  - ❖ Prioritize and focus on problem areas
  - ❖ Maintain broad perspective while working with detailed data
  - ❖ Both point-in-time view and maintaining ongoing quality important
- Tools
  - ❖ Essential for efficiency
  - ❖ Should not drive the approach

© 2001 Deloitte & Touche LLP. All rights reserved. Proprietary information. This document contains concepts, ideas and materials that are proprietary to Deloitte & Touche LLP and may not be used, copied, provided to others, or referred to without the express permission of Deloitte & Touche LLP.

**Deloitte & Touche**

### Questions?

© 2001 Deloitte & Touche LLP. All rights reserved. Proprietary information. This document contains concepts, ideas and materials that are proprietary to Deloitte & Touche LLP and may not be used, copied, provided to others, or referred to without the express permission of Deloitte & Touche LLP.

**Deloitte & Touche**

# **Data Quality and Medical Record Abstraction in the Veterans Health Administration's External Peer Review Program**

(Practice-oriented paper)

James H. Forsythe, PhD

Epidemiologist, West Virginia Medical Institute

Jonathan B. Perlin, MD, PhD

Chief, VHA Office of Quality and Performance

John Brehm, MD

Chief Medical Officer, West Virginia Medical Institute

## **Executive Summary**

Under the Veterans Health Administration's External Peer Review Program, the West Virginia Medical Institute (WVMI) conducts monthly medical record abstractions in over 150 VA Medical Centers throughout the United States and Puerto Rico. The abstractions are performed by approximately 90 highly trained abstractors and are used to assess VHA clinical performance for: in-patient and out-patient encounters, JCAHO ORYX measures, and ad hoc studies on topics such as management of low back pain, spinal chord injury, and diabetic foot care. To help improve the validity and reliability of the abstracted medical data, WVMI has implemented a multi-method approach to monitoring abstracted data quality. The approach includes five major components:

- Bi-weekly computer-aided screening to detect anomalous performance (e.g., leading and terminal digit distributions of continuous variables);
- On-site interrater reliability assessments and calculation of prevalence adjusted Kappa agreement between abstractors and supervising Network Coordinators;
- Random and special assignment audits by one or more trained auditors;
- Analyses using SAS Enterprise Miner (including runs and randomness testing, hierarchal modeling (decision tree and cluster analysis) and neural network programming for assessing performance;
- Statistical process control for tracking and trending performance of abstractors, VAMCs, and items over time.

In addition, WVMI has created web-enabled feedback capabilities so that key administrators can rapidly access and report on performance impacting data quality. This paper will outline the data quality techniques and results that have enhanced the use of medical record data for assessing clinical performance throughout the VHA system.

### Data Quality and Medical Record Abstraction in the Veterans Health Administration's External Peer Review Program

James H. Forsythe, Ph.D.  
West Virginia Medical Institute  
Jonathan B. Perlin, M.D., Ph.D.  
Chief, Office of Performance and Quality  
Veterans Health Administration  
John G. Brehm, M.D.  
West Virginia Medical Institute



### West Virginia Medical Institute

- WVMI is one of 37 designated Peer Review Organizations serving Medicare and Medicaid beneficiaries
- Staffed by 200 employees located in six offices in WV, VA, DE, & MD
- WVMI conducts medical record review for the Veterans Health Administration and the Department of Defense



### Veterans Health Administration External Peer Review Program

- “EPRP” began in 1992; WVMI has been prime contractor for both of the 5 year cycles
- EPRP assesses clinical guideline performance using third party medical record abstraction
- EPRP is used for comparing performance among VHA hospitals, clinics, and across the 22 administrative regions



### Medical Record Review

- WVMI has a Nation-wide network of 95 certified medical record abstractors
- Records are abstracted throughout the year at 170 hospitals
- In FY 2001 over 350,000 records were abstracted in hospitals, out-patient clinics, and other care delivery settings
- Records are transmitted electronically to Charleston, WV and compiled and analyzed for quarterly reports



### Objectives for Abstractor Monitoring and Data Quality Assessment

- Measure abstractor performance and detect anomalous behavior
- Use “real-time” surveillance & analytical techniques to more quickly identify and correct substandard abstractor performance
- Rule out abstractor “error” and focus on other sources of variation
- Use surveillance for quality control *and* quality improvement



### Techniques used to Build the Monitoring and Assessment Model

- Data Entry Error Detection
- Leading & Terminal Digit Analysis
- Pattern Analysis
- Cluster Analysis
- AI-aided Profiling

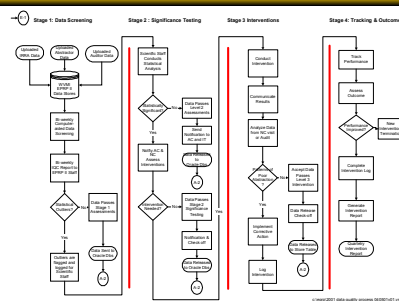


### Steps in Monitoring and Improving Abstractor Performance

- Screen up-loaded medical records
- Identify abstractors (and records) with unexpected results
- Analyze results to determine source and extent of the anomalous performance
- Conduct interventions and field audit where needed
- Use results for quality improvement training



### Abstractor Monitoring and Data Quality Assessment Flow Diagram

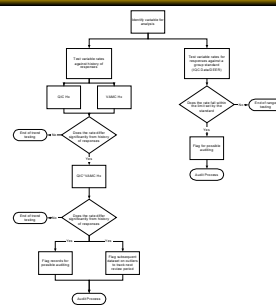


### Stage 1 "Real-time" Screening Techniques

- Data Entry Error Rates
- Leading & Terminal digit analysis
- Disease discrepancy rates
- Diabetic abnormal foot rates
- % Dates filled
- Do not review rates



### IQC Data Screening Process

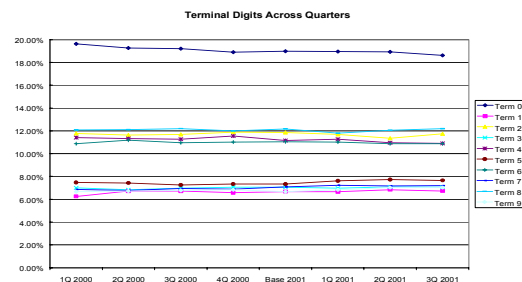


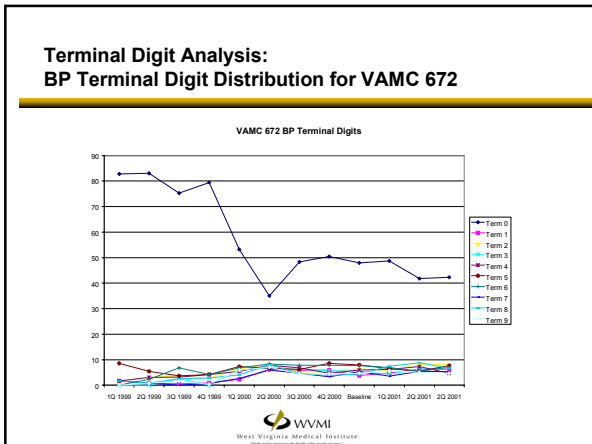
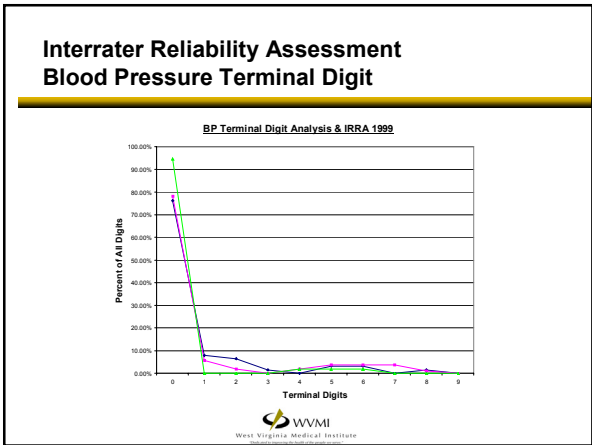
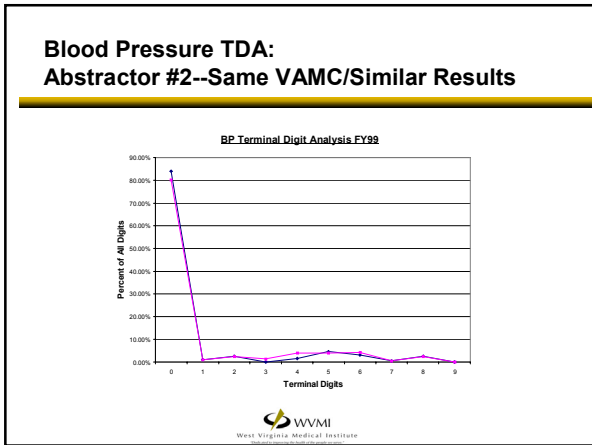
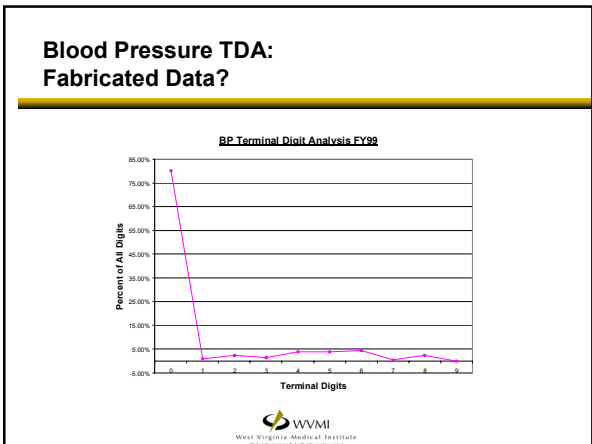
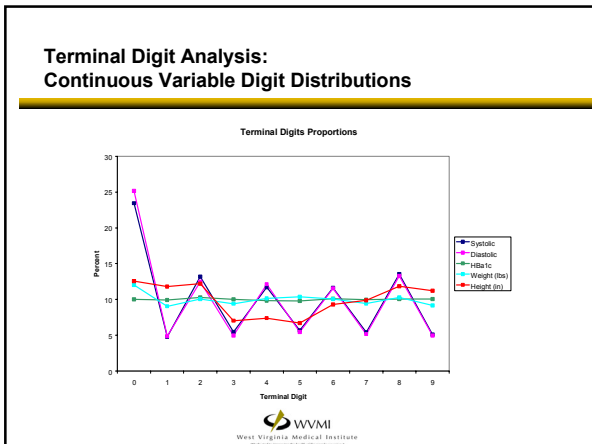
### Examples of Data Entry Error Reduction

	4Q98	3Q00
<b>HT</b>		
Max	762	82
Min	-70	49
<b>WT</b>		
Max	2,891	515
Min	-15	47
<b>BPs</b>		
Max	662	253
Min	70	54
<b>BPd</b>		
Max	150	126
Min	1	32
<b>HbA1c</b>		
Max	98.7	18.4
Min	0	3.5



### Terminal Digit Analysis: Continuous Variables





### Data Screening Abstractor Anomaly Report

#### 3rd Quarter 2001 - Anomaly Report

Number of Anomalies	5
QIC	Area
120	Abnormal Foot Exams: Pnumovac Contraindicated: HTN Discrepancy: DM Discrepancy: COPD
Number of Anomalies	3
QIC	Area
204	Abnormal Foot Exams: Hospice/Terminal: Pnumovac Contraindicated:
185	Pnumovac Contraindicated: HTN Discrepancy: COPD Discrepancy:
138	Terminal Digit 0: Pnumovac Contraindicated: Do not review:
128	Pnumovac Contraindicated: Do not review: COPD Discrepancy:
102	Terminal Digit 0: Pnumovac Contraindicated: HTN Discrepancy:

WVMI  
West Virginia Medical Institute

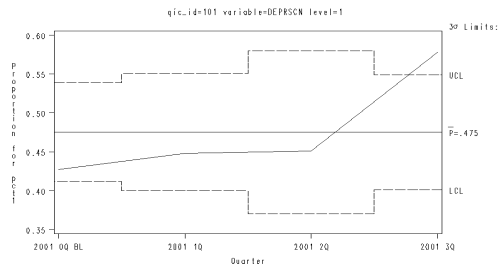
## Abstractor Outlier Report

2001 3rd Quarter - QIC Outlier Report: Key Performance Indicators for QIC 199

variable	level	description	Chg from history	Chg from previous	2001 3Q	2001 2Q	2001 1Q	2001 0Q BL
HMOGVD	1	Yes	extreme increase	significant increase	26%	17%	0.6%	4.3%
	2	No	extreme decrease	significant decrease	74%	83%	100%	96%
TOBSTATUS	1	Current user	increased	increased	19%	19%	18%	18%
	2	Former user	increased	increased	23%	21%	17%	16%
	3	Decline current user not further info.	decreased	decreased	20%	22%	20%	20%
	4	No use in past 7 years	increased	increased	20%	31%	28%	28%
	5	No documentation	extreme decrease	significant decrease	0.2%	0.2%	0.8%	10%



## SPC Analysis of Increase in Depression Screening



## Techniques used to Assess Data Reliability

- Interrater Reliability Assessment
- Intrarater Reliability Assessment
- False Negative & False Positive Rates
- Service/Clinical Indicator Date Variance
- Item Reliability Assessment



## Interrater Reliability Assessments

- "IRRAs" occur between abstractors and either their field supervisor or an auditor
- Attempt to interrater between 20 and 25 records
- Calculation of agreement using weighted percent agreement and Kappa "beyond chance" agreement
- Abstractors (and items) yielding low Kappa agreement (< .85) are identified for QI training



## Problems with Kappa in Contexts of High Goal Attainment

- *Prevalence of an observed trait:*  
100% Agreement that a service was provided = No Kappa Score

Example:

	Yes	No
Yes	20	0
No	0	0

% agreement = 100  
Kappa can not be calculated



## One disagreement can yield a Kappa Score of Zero

- 95%+ Agreement that a service was/was not provided can yield a **zero or negative** Kappa Score

Example:

	Yes	No
Yes	19	1
No	0	0

% agreement = 95  
Kappa = 0




### High agreement yielding a negative Kappa

- 90%+ Agreement that a service was/was not provided can yield a zero or negative Kappa Score
- Example 5:
 

	Yes	No
Yes	18	1
No	1	0

% agreement = 90  
Kappa = -.05


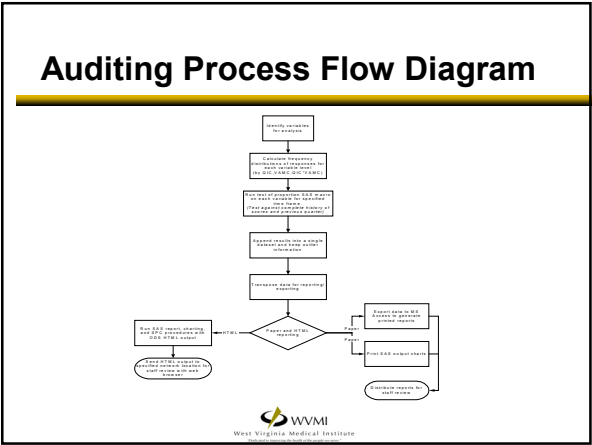
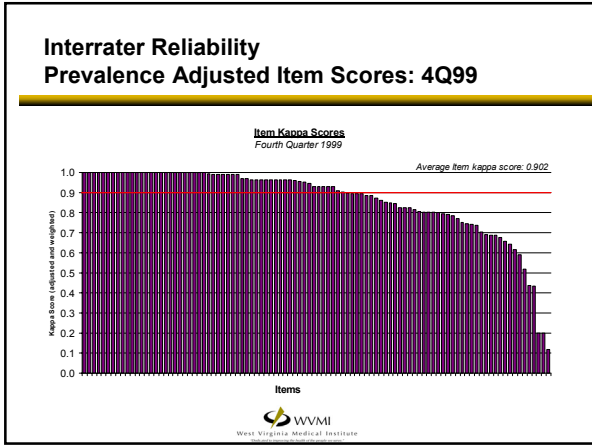


### High agreement yielding an "unacceptable" Kappa Score

- 95% Agreement that a service was/was not provided can yield a low Kappa Score
- Example:
 


	Yes	No
Yes	18	0
No	1	1

% agreement = 95  
Kappa = .64


### 3Q01 Abstractor Assessment & Audit Kappas

Abstractor Assessment		Audit	
Kappas	#Records	Kappas	#Records
0.83	32	0.87	24
0.88	16	0.84	23
0.94	15	0.88	19
0.85	12	0.89	24
0.96	13	0.87	17
Overall: 0.892		0.92	24
		0.84	21
		0.84	22
		0.86	28
		0.87	25
		Overall: 0.868	

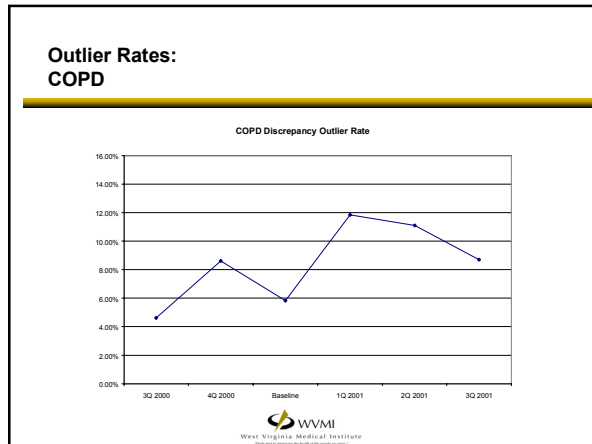
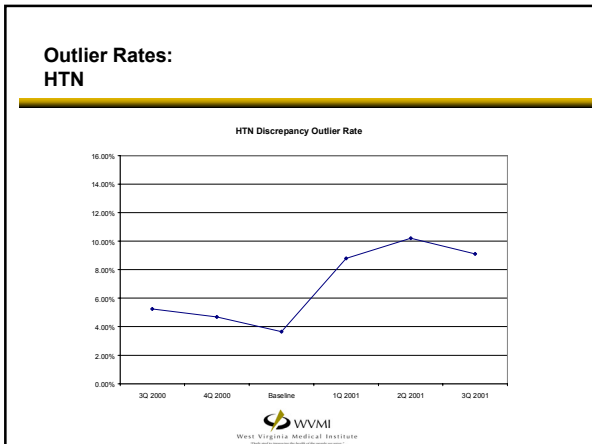
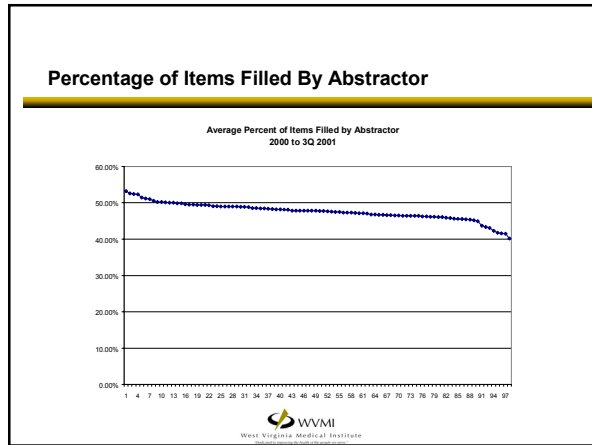
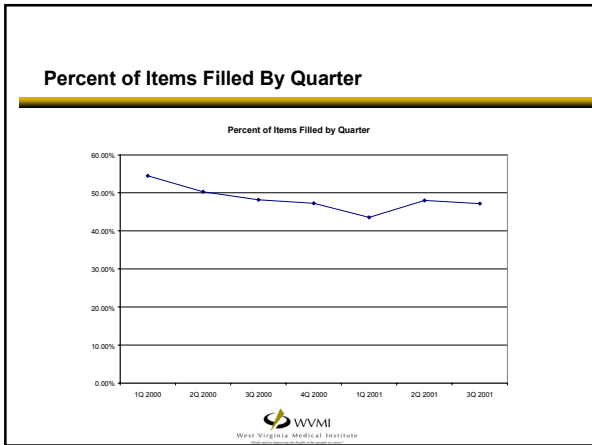
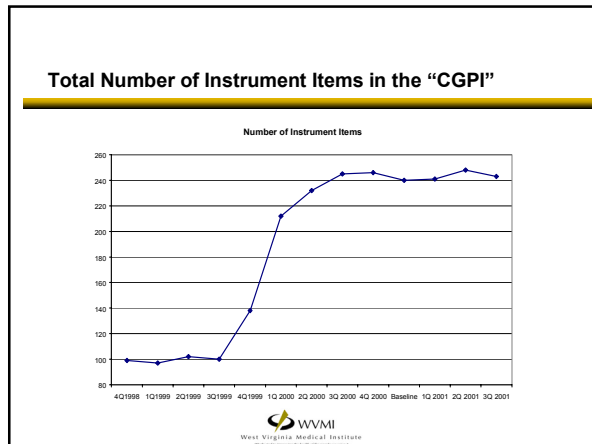
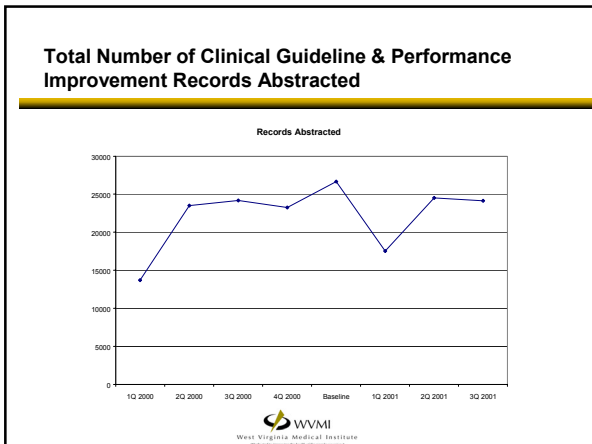


### Abstractor Performance with Increases in Record Volume

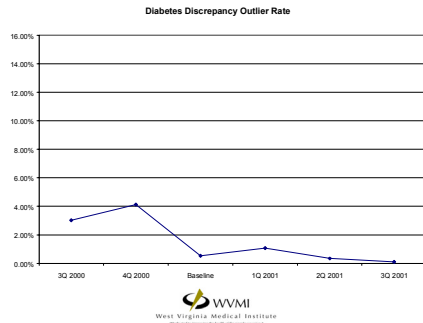
- Over an 18 month period, the number of required abstractions nearly tripled
- WVMI increased the number of abstractors from 35 to 95
- How has performance been impacted with increases in record volume?
- How has adding items to the instrument impacted medical record abstraction?



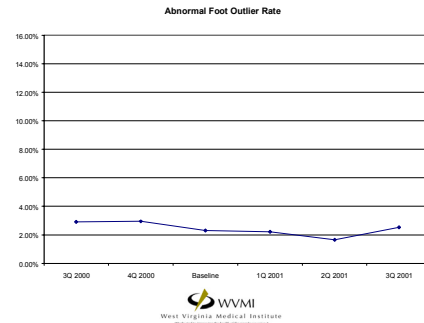




### Outlier Rates: Diabetes



### Outlier Rates: Abnormal Foot



### Current Status of the Assessment and Data Quality Model

- Demonstrated ability to detect negligent or fabricated data
- Rate of agreement among abstractors is approximately 90%
- Agreement rates are impacted by quality in, and types of, record keeping (paper, electronic, and both together) and, the item needing abstraction

**The Canadian Institute for Health Information (CIHI)  
Data Quality Framework, Version 1:  
A Meta-Evaluation and Future Directions**

(Practice Oriented Paper)

J.A. Long  
Consultant, Data Quality  
Canadian Institute for Health Information  
Phone: 416-481-1616 Ext. 3579  
Fax: 416-481-2950  
Email: [jlong@cihi.ca](mailto:jlong@cihi.ca)  
[www.cihi.ca](http://www.cihi.ca)

J.A. Richards  
Manager, Data Quality  
Canadian Institute for Health Information  
Phone: 416-481-2002  
Fax: 416-481-2950  
Email: [jrichards@cihi.ca](mailto:jrichards@cihi.ca)  
[www.cihi.ca](http://www.cihi.ca)

C.E. Seko  
Senior Analyst, Health and Outcomes Statistics  
Health Statistics Division, Statistics Canada  
Phone: 613-951-4931  
Fax: 613-951-4251  
Email: [sekocet@statcan.ca](mailto:sekocet@statcan.ca)  
[www.statcan.ca](http://www.statcan.ca)

**Abstract.** Information quality (IQ) problems can have severe consequences in the health care sector. Since its inception, the Canadian Institute for Health Information (CIHI) has recognized the importance of information quality and has implemented a framework designed to evaluate the data quality of the numerous CIHI data holdings. After one year of implementation, a meta-evaluation of the CIHI data quality framework evaluation process was conducted and the framework was found to be both relatively strong theoretically as well as practical. Despite a relatively favourable meta-evaluation, several aspects of the framework are scheduled for improvement. It is recommended that data quality framework and meta-evaluation development be recognized as crucial with the ultimate objective of improving information in the health care sector.

**Keywords:** Data quality, framework, information quality research, meta-evaluation

**Introduction.** Information quality (IQ) problems can have severe financial and operational consequences for organizations<sup>i</sup>, and in the health care sector, can impact life and death decisions. Since its inception, the Canadian Institute for Health Information (CIHI) has

recognized the importance of data quality. Although data quality has been a priority since the establishment of CIHI, a new framework, designed to evaluate the data quality of the numerous CIHI data holdings, has recently been implemented. The following is a meta-evaluation of the current CIHI data quality framework evaluation process.

**Background.** The CIHI Data Quality Framework, Version 1 (CIHI-DQF, v1) is a four-level conceptual model designed to standardize and facilitate the systematic quantification and measurement of data quality at CIHI<sup>ii</sup>. Incorporated in 1993, CIHI is a federally chartered, yet independent, not-for-profit organization. The Institute was established through the amalgamation of two non-governmental organizations, i.e., the Hospital Medical Records Institute (HMRI) and the MIS Group, along with selected databases and functions from the Health Information Division of Health Canada and the Health Statistics Division of Statistics Canada (STC)<sup>iii,iv</sup>.

CIHI has taken a central role in the development of Canada's health information system and is mandated to “serve as the national mechanism to coordinate the development and maintenance of a comprehensive and integrated health information system for Canada” and “to provide and co-ordinate the provision of accurate and timely information required for: the establishment of sound health policy, the effective management of the Canadian Health System, and for generating public awareness about factors affecting good health. Consistent with its mandate, an important role taken on at CIHI is the collection, processing, and maintenance of a growing number of clinical databases or registries, as well as, health human resources, health services, and health expenditures databases. To date, the CIHI data holdings include 22 databases and registries, many of which are national in scope<sup>v</sup>.

As a result of CIHI being an amalgamation of several programs, the data quality methods were initially non-standardized and database or registry specific. In recognition of the vital importance of data quality, as well as, due to the responsibility of maintaining 22 databases or registries, the need for a standard strategy to identify data quality problems, to enable senior CIHI management to allocate finite resources across 22 data holdings, and to solve unforeseen problems is fast becoming imperative.

In response to the identified need for a standard, organized, and systematic approach to data quality, one of the authors (Seko) was seconded from STC in 1999 to develop a data quality strategy in collaboration with CIHI senior management. The cornerstone of the resulting strategy is the CIHI Data Quality Framework, Version 1 (CIHI-DQF, v1). The CIHI-DQF, v1 draws on the statistical literature<sup>vi,vii,viii,ix,x</sup> the STC guidelines and methods<sup>xi</sup>, the Information Quality literature<sup>xii</sup>, the CIHI mandate, as well as the principle of Continuous Quality Improvement (CQI). The first version of the framework was ready and implemented in April 2000. To date, two major database evaluations have been conducted and another three major database evaluations are taking place at the time of this study.

**Rationale.** Consistent with the principle of CQI for ongoing data quality measurement, evaluation, and improvement, efforts must be made to ensure the integrity and relevance of the evaluation process itself. This can be achieved by requiring that the process itself be systematically assessed and improved on a continuing basis. That is, not only should the principle of CQI be applied to the data holdings, it should also be applied to the *methods* used to

evaluate the data. As CIHI-DQF, v1 has been successfully implemented for over one year now, in an ongoing endeavor to improve its effectiveness, it is an opportune time to conduct a meta-evaluation in the effort to assess its performance.

While the importance of applying CQI to any data quality framework or evaluation process should be apparent, there appears to be little evidence in the literature that this idea has been considered. In fact, it appears that the only relevant work available is a research-in-progress conducted by Eppler and Wittig<sup>xiii</sup>.

Eppler and Wittig argue that an IQ framework should be practical in addition to being theoretically strong. They reason that an IQ framework should be theory driven and a theory based conceptual map upon which a framework is based should be available to the research community. Moreover, they suggest that a systematic and concise set of measurement and evaluation criteria, a scheme to analyze and solve quality problems, and a plan to facilitate proactive management should be available.

In order to assess whether some of the leading IQ frameworks are academically rigorous, as well as practical, Eppler and Wittig put forward a basic method for evaluation. The aim of this paper is to conduct a meta-evaluation of the CIHI-DQF, v1 according to the method developed by Eppler and Wittig. The meta-evaluation results integrated with the preliminary framework implementation experience will together direct future framework improvement.

**Methods.** A basic descriptive meta-evaluation was conducted of the CIHI Data Quality Framework, Version 1 (CIHI-DQF, v1). The evaluation methodology was simply a rating of whether the CIHI-DQF, v1 addressed the Eppler and Wittig categories or not; and if so, a simple qualitative statement was included as to how well the category was addressed. Applied findings based on the first year of the framework's implementation were integrated into the meta-evaluation.

### **The CIHI Data Quality Framework, Version 1 (CIHI-DQF, v1)**

More specifically, the CIHI-DQF, v1 was designed to operationalize, measure, and evaluate the quality of the CIHI data holdings using a standard and systematic approach. The objectives of the CIHI-DQF, v1 are: 1) to standardize information on data quality, both for internal and external users; 2) to provide a common strategy for assessing data quality; and 3) to define a work process for CIHI's data holdings that identifies data quality priorities and produces continuous improvement in data quality.

According to the CIHI-DQF, v1, 'quality' is defined as 'fitness for use'<sup>2</sup> and 'data quality' is operationally defined and measured along five common and widely used quality dimensions. Consequently the framework was designed to facilitate the evaluation of these dimensions, as well as, to provide a single overall evaluation of a data holding based on the five dimensions.

Specifically, the CIHI-DQF, v1 is organized as a four-level conceptual model. At the foundation of the model are 86 basic unit items that are known as *criteria*. The 86 criteria can be collapsed into the second level of 24 *characteristics* (e.g., under-coverage, reliability, and interpretability)

that in turn, can be collapsed into 5 *dimensions* of data quality (i.e., i. accuracy, ii. timeliness, iii. comparability, iv. usability, and v. relevance). Finally, the 5 dimensions can be collapsed into one overall evaluation of the database. Figure 1 below provides a summary of the CIHI-DQF, v1.

**Figure 1. The CIHI Data Quality Framework Evaluation Instrument, Version 1 (CIHI-DQF, v1)**

Dimension	Characteristics	Criteria
I. Accuracy	i.1. Over-coverage	1-6
	i.2. Under-coverage	7-12
	i.3. Simple response bias	13
	i.4. Reliability	14-15
	i.5. Correlated response bias	16
	i.6. Collection and capture	17-24
	i.7. Unit non-response	25-26
	i.8. Item (partial) non-response	27-30
	i.9. Edit and imputation	31-37
	i.10. Processing	38
	i.11. Estimation	39-41
II. Timeliness	ii.1. Timeliness	actual release-planned release 42-45
III. Comparability	iii.1. Comprehensiveness	46-49
	iii.2. Integration	50-53
	iii.3. Standardization	54-57
	iii.4. Equivalency	58-59
	iii.5. Linkage-ability	60-64
	iii.6. Product comparability	65
	iii.7. Historical comparability	66-69
IV. Usability	iv.1. Accessibility	70-75
	iv.2. Documentation	77-78
	iv.3. Interpretability	79-81
V. Relevance	v.1. Adaptability	82-84
	v.2. Value	85-86

Whereas the CIHI-DQF, v1 provides standard definitions and a common strategy, the framework itself is designed to be a part of a work process that identifies data quality priorities and produces continuous improvement in data quality. Once the measurement of data quality is achieved, then it must be improved, then measured again, improved, and so on. While the framework quantifies the concepts of data quality and enables measurement, the evaluation process, based on the CIHI-DQF, v1, puts the principle of CQI into action.

### **The CIHI Data Quality Framework, Version 1 (CIHI-DQF, v1) Database Evaluation**

While the objectives of the CIHI-DQF, v1 are essentially to standardize efforts and to define a work process, the objectives of the data holdings evaluation process based on the CIHI-DQF, v1 are: 1) to identify and rank aspects of data quality needing improvement; and 2) to produce information on data quality that feeds into the creation of data quality documentation for users<sup>xiv</sup>. In fact, the facilitation of data quality documentation for users is key to the database evaluation process.

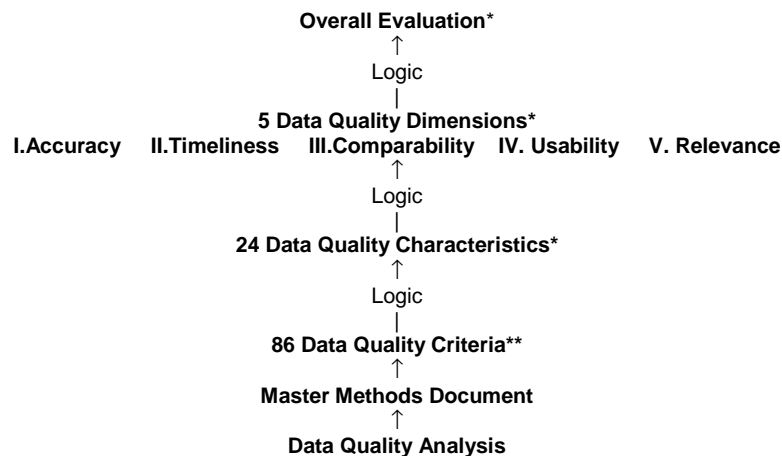
In addition to the CIHI-DQF, v1, a data quality manual that itemizes the step-by-step evaluation process is also provided to those responsible for CIHI data holdings. In fact, the manual was designed to complement, and is integral, to the framework. Part 1 of the manual provides instruction for analysing data quality, describing data quality, and for improving data quality, and

Parts 2 and 3 contain the CIHI-DQF, v1 and an instrument based on the framework, respectively. Direction on recommended analyses and detailed instruction on how to prepare an evaluation document are provided in the manual. Alongside assessments, evaluators are also instructed to include short and long term recommendations. Moreover, upon assessment completion evaluators are instructed to rank and summarize the recommendations at the beginning of the evaluation document. Assignments are made in consultation with the staff responsible for a given database and in tandem with the Data Quality Section. In addition, ongoing Data Quality Workshops designed to educate staff with respect to the database evaluation process are offered annually. In fact, training is considered to be central to the entire data quality strategy.

Besides the preparation of an evaluation document, the evaluation process includes the completion of the CIHI Data Quality Framework Evaluation Instrument, Version 1 which is designed to facilitate the computation of an evaluation and to enable the Data Quality Section track progress within and across databases. Specifically, the 86 criteria are scored in a consistent fashion so that a low value indicates a less favorable evaluation and a score of 0 indicates that a criterion is 'not applicable'. The criteria evaluations are: 0) not applicable; 1) unknown; 2) not met; and 3) met.

Likewise, the characteristic, dimension, and overall evaluations are scored such that low values indicate less favorable evaluations (not including 'not applicable'). The characteristic, dimension, and overall evaluations are: 1) unknown; 2) not acceptable; 3) marginal; and 4) appropriate. Once categories are assigned at the criteria level, the characteristic, dimension, and overall database evaluations can be easily computed based on the framework algorithm. SAS code, based on the CIHI-DQF, v1 algorithm for reading and scoring the data is currently under development. Figure 2 below provides an outline of how the CIHI Data Quality Framework Evaluation Instrument, Version 1 is scored.

Figure 2. The CIHI Data Quality Framework, Instrument, Version1 (CIHI-DQF, v1) Algorithm



\*1. Appropriate, 2. Marginal, 3. Not acceptable, and 4. Unknown  
 \*\*0. Not Available, 1. Unknown, 2. Not Met, and 3. Met

In other words, at the bottom of the four-level model are 86 criteria that roll up to 24 data quality characteristics. Each of the 86 criteria, and hence the 24 data quality characteristics, can be regularly evaluated for a database or registry. Combined evaluations of constituent

characteristics define the assessment of a dimension. Combining dimensions gives an overall impression of the data quality for a database. All levels can also be combined across databases to summarize a dimension (or characteristic) for the entire set of CIHI data holdings. The aim is to identify and rank aspects of data quality in need of improvement in a comparable way such that resources can be optimally allocated across competing data holdings. The result is a comprehensive and integrated picture of the data quality within and across databases.

Again, the purpose of collecting and scoring evaluation data is to help identify and prioritize data quality improvement tasks<sup>xv</sup>. Lastly, evaluators are instructed to think of the evaluation process as ongoing and to decide on a fixed time period (e.g., annually) for continuous data quality evaluations and improvement, with the ultimate objective of improving data quality and, by extension, critical health information based on the data. Lastly, the database evaluation process also includes a careful consideration of confidentiality, privacy, and security issues and all concerns are addressed.

### The CIHI Data Quality Framework, Version 1 (CIHI-DQF, v1) Meta-Evaluation

The evaluation of the CIHI-DQF, v1 evaluation process, or meta-evaluation, was conducted according to Eppler and Wittig. Briefly, Eppler and Wittig put forward a basic method to assess the academic rigour as well as the practicality of an information quality framework. Their method entails assessing a framework according to the following six evaluation categories or ‘meta-criteria’: 1) definitions; 2) positioning; and 3) consistency, to assess theoretical robustness, as well as, 1) conciseness; 2) examples; and 3) tools, and to assess practicality. Each category or criterion is associated with a key question that is used to evaluate a framework. Figure 3 below illustrates the Eppler and Wittig meta-criteria and associated evaluation questions.

Figure 3: The Eppler and Wittig (2000) Meta-Criteria for the Evaluation of IQ Frameworks

Meta-Criteria	Evaluation Questions
I. Analytic	
I.1 Definitions	Are all individual information quality criteria clearly defined and explained? Are all the dimensions to which the individual criteria are grouped (if existing) defined and explained?
I.2 Positioning	Is the context of the framework’s application (and its limits) clear? Is the framework positioned within existing literature?
I.3 Consistency	Are the individual criteria mutually exclusive and collectively exhaustive? Is the framework overall divided into systematic dimensions that are also mutually exclusive and collectively exhaustive? Is it clear why a group of criteria belongs to the same dimension?
II. Practical	
II.1 Conciseness	Is the framework concise in the sense that it can be easily remembered? Are there (as a minimal rule of thumb) less than seven dimensions and less than seven criteria per dimension?
II.2 Examples	Are specific and illustrative examples given to explain the various criteria (e.g., case studies)?
II.3 Tools	Is the framework accompanied by a tool that can be used to put it into practice, such as a questionnaire, a software application, or a step-by-step implementation guide or methodology?

Hence, the CIHI-DQF, v1 was evaluated according to the six Eppler and Wittig theoretical and practical evaluation questions. The evaluation methodology was simply a brief description of if the meta-criteria were addressed, and if so, how each question was addressed by the CIHI-DQF, v1. Where applicable, findings from applied experience were included and possible future improvements were also itemized.



In terms of the applied experience, while the CIHI-DQF, v1 was officially released in April 2001, its test phase spanned April 2000-April 2001. Though applied experience involving all of the CIHI databases and registries was not available at the time of the study, two major database evaluations using the framework were conducted prior to the framework's official release and another three major evaluations were underway at the time of the study. A summary of evaluator feedback as well as notes from those involved in the development of version 1 was carried out and integrated with the meta-evaluation. Future directions were itemized based on the meta-evaluation results and the applied experience.

**Results.** Included for evaluation was the CIHI Data Quality Framework, Version 1 (CIHI-DQF, v1) which was developed in 1999 and tested from April 2000-April 2001. The following represents the results of a meta-evaluation of the CIHI-DQF, v1 conducted according to the method put forward by Eppler and Wittig. Applied findings and test phase notes were integrated along with the meta-evaluation results and future directions are presented.

Table 1 provides a summary of the CIHI-DQF, v1 meta-evaluation findings for the analytic or theoretical performance of the CIHI-DQF, v1. Regarding the first meta-criterion known as 'definitions', characteristic and dimension definitions are provided, however, applied findings suggest that the definitions are not detailed enough. Moreover, references are not provided. Certain dimensions (e.g., relevancy) and characteristics (e.g., equivalency) have been found to be unclear. The level 1 criteria, on the other hand, have been found to be clear. Improving the framework definitions and referencing have been targeted in version 2.

**Table 1. A Meta-Evaluation of the Theoretical Foundation of the CIHI Data Quality Framework, Version 1 (CIHI-DQF, v1)**

Meta-Criteria*	The CIHI-DQF, v1 response	Applied findings Year 1	Summary	Future Development
I. Analytic I.1. Definitions	Definitions of the dimensions and the characteristics exist and are provided.	Overall, definitions are not detailed enough and references are not provided.  Certain dimensions (e.g., relevancy) and characteristics (e.g., equivalency) have been found to be particularly unclear.	Definitions exist and are provided but, similar to other frameworks, are in need of development.	Improve and develop definitions. As well, include examples and references.
I.2. Positioning	The framework is based on STC DQ Guidelines and the CQI literature as well as the CIHI mandate.	Not raised as an issue. Those applying the framework focused on the practical aspects of the framework.	More information regarding how the framework fits with respect to the field was nevertheless flagged by the authors as necessary.  The context of the framework's application and limits could be clearer.	Despite not being raised as an issue and although a brief summary of the literature is provided, the position of the framework within the literature will be expanded and referenced in order to contribute to the field.
I.3. Consistency	Individual dimensions, characteristics, and criteria are reasonably mutually exclusive and collectively exhaustive.	While most dimensions were found to be reasonably distinct, the characteristics within Comparability (e.g., integration, standardization, and linkage) and Relevance were reported to be confusing.  The framework is generic and was found to be applicable for several different data CIHI sources.	The concepts underpinning the Comparability and Relevance dimensions need further clarification.  Interdependency across the dimensions (e.g., the tradeoff between accuracy and timeliness ) is not available.  The framework was found to be generic.	The concepts underpinning the Comparability and Relevance dimensions will be clarified for version2.  The discussion of the Interdependency across dimensions or characteristics will be explored.

\*Eppler and Wittig (2000)

In terms of positioning, while the CIHI-DQF, v1 is based on the Statistics Canada (STC) framework, the STC framework in turn is based on the literature. Nonetheless, the context of the framework's application and its limits have been found to be unclear in the documentation and, again, references are not included. Whereas those who applied the framework did not raise these issues as a problem, those involved with the design of version 1 flagged the background description of the framework as cursory and an area in need of expansion when time and resources permit. Version 2 will include an expanded background section that describes how the CIHI-DQF, v1 fits with the literature and all references will be included.

For consistency, the framework's individual components were, for the most part, found to be mutually exclusive and collectively exhaustive. The framework overall is generally thought to be divided into systematic dimensions that make sense and the numerous characteristics are thought to be logically assigned to their dimensions. While most dimensions were found to be reasonably distinct, the characteristics within 'comparability' (e.g., integration, standardization, and linkage)

were flagged as unclear. Although no comments were made by the test phase evaluators, consistent with Eppler and Wittig we also targeted the relevance dimension as a possible area for clarification. Furthermore, an explanation of the interdependency across the dimensions (e.g., the tradeoff between accuracy and timeliness or linkage-ability and privacy) was not well developed. These tradeoffs will need to be further explored in version 2. Also in terms of consistency, the framework was found to be applicable for different data sources (e.g., clinical data or health human resources data). The applicability of the framework outside of the collection of official health care sector statistics is unknown.

Table 2 summarizes how pragmatic the CIHI-DQF, v1 is. In terms of practicality, test phase participants initially felt overwhelmed and where to start was not clear to them. As well, they found that the framework evaluation document was not user-friendly. In response, an instrument based verbatim on the framework was developed and initial feedback indicates that user-friendliness may have been improved. Despite not knowing where to start, once underway evaluators indicated that the framework was concise and practical. Also of note, the framework contains less than seven dimensions and, for the most part, less than seven characteristics per dimension. Version 2 will include clearer step-by-step directions in the manual and the manual will become more centered on the framework.

**Table 2. A Meta-Evaluation of the Pragmatic Aspects of the CIHI Data Quality Framework, Version 1 (CIHI-DQF, v1)**

Meta-Criteria*	The CIHI-DQF, v1 response	Applied findings Year 1	Summary	Future Development
II. Pragmatic II.1. Conciseness	The framework has less than 7 dimensions and, for the most part, less than 7 characteristics per dimension. For the most part, each characteristic is based on fewer than 7 criteria.	Length has not been raised as an issue.  Where to start and user-friendliness have been flagged as areas in need of improvement.	Good start.	Conciseness has not been targeted in version2.  The manual will become more framework centered and the step-by-step evaluation instructions will be expanded.
II.2. Examples	Of the 86 criteria, only 6 specific examples are provided. No case studies are provided.	Examples were requested for all of the components as well as for the entire process of the data quality evaluation.	More criteria or characteristic examples should be included as well as entire evaluations as they become available.	Available examples will be incorporated and, as they become available, entire hardcopy and online evaluations will be included.
II.3. Tools	Tools include ongoing training, a manual, an instrument version of the framework, and an evaluation algorithm.	Although several tools exist, where to start has been raised as a consistent concern.  The connection between the manual and the framework was found to be unclear.	The value and number of tools is sufficient however the tie between them is not clear enough.	The connection between the framework and the manual will be improved. Clearer step-by-step instructions will be provided.  Framework algorithm SAS code will be developed.  A software application has been suggested, however, this suggestion may not be addressed in version2.

Of the 86 criteria included in the CIHI-DQF, v1, only 6 specific examples are provided and no case studies are provided. Congruent with the poor evaluation for the ‘examples’ meta-criterion, the authors, as well as the test phase participants, indicated that more examples were necessary in version 2. Available examples will be included and as new examples become available they will also be incorporated.

The data quality training, manual, framework, and instrument were found to be valuable tools and the number of tools was found to be sufficient. A software version of the instrument was suggested however operational restraints might prevent software development in the near future. Although the existence and number of tools was found to be sufficient, how they related was identified as an area for improvement. Of specific concern, was the relation between the manual and framework documents. The connection between these documents will be clarified in version 2.

**Discussion.** After one year of implementation, the time had come to evaluate the performance of the CIHI Data Quality Framework, Version 1 (CIHI-DQF, v1). Although the importance of quality data is unchallenged and frameworks designed to operationalize or quantify data quality, with the ultimate objective of improving it, are crucial, few meta-evaluation protocols are readily available. This is troubling for those who maintain health data, as poor quality data can be dangerous and has the potential to cause great harm. The approach used by Eppler and Wittig in their review of seven principal frameworks was applied to the CIHI-DQF, v1 and integrated with preliminary implementation findings in order to guide improvement efforts.

### **The CIHI Data Quality Framework Evaluation, Version 1 (CIHI-DQF, v1) Meta-Evaluation Findings**

In terms of theoretical robustness, the CIHI-DQF, v1 concepts were found overall to be reasonably well defined, well positioned in the literature, and consistent. However, certain weaknesses were detected. While definitions of the framework characteristics and dimensions are provided in the CIHI-DQF, v1, the definitions may not be detailed enough and references need to be provided. Consistent with the findings of Eppler and Wittig, certain dimensions (e.g., relevancy) or characteristics have been found to be unclear.

While the framework is based on the literature, the context of the framework’s application and its limits are not clearly stated in the documentation. The framework was also found to be consistent within its dimensions and characteristics, as well as, with other frameworks. Similar to the seven other frameworks reviewed by Eppler and Wittig, the CIHI-DQF, v1 includes the concepts of timeliness, accessibility (as a characteristic within usability), relevance, and accuracy. While no dimensions, characteristics, or criteria in the CIHI-DQF, v1 were specifically referred to as ‘objectivity’, ‘consistency’, or ‘completeness’, there may be significant overlap with some of the categories included in the CIHI-DQF, v1. For example, ‘completeness’ might be similar to the framework concept of ‘comprehensiveness’.

Consistent with other work, however, the interdependency across the dimensions (e.g., the tradeoff between accuracy and timeliness or the tradeoff between linkage-ability and privacy) was not well developed and was identified as an area for expansion in version 2. While Eppler

and Wittig found that many of the frameworks they reviewed were domain-specific, it's believed that the CIHI-DQF, v1 is also fairly generic within the realm of official health statistics collection and was found to be applicable across several diverse data sources at CIHI. In fact, given that much of the health care sector population data is 'administrative data', it is possible that the framework has not considered fully this type of specific data, along with its unique challenges<sup>xvi</sup>. One solution might be to maintain applicability as well as to include more components specific to administrative data.

In terms of practicality, test phase participants tended to feel overwhelmed. Once underway, however, evaluators indicated that the framework was concise and practical. The inclusion of illustrative examples on the other hand was flagged as an area for improvement. Lastly, unlike other frameworks the CIHI-DQF, v1 is accompanied by several tools including a workshop, a manual, framework documentation, and an evaluation instrument. Although the value and the number of tools were found to be sufficient, how the tools related was found to be unclear.

### **CIHI Data Quality Framework, Version 2 (CIHI-DQF, v2) Development Plans**

Based on both the meta-evaluation results and the applied experience, improvement plans for the framework include:

- expanding the framework definitions;
- expanding the background sections of the CIHI-DQF, v1 documentation so that the conceptual map is clear for evaluators as well as for the research community;
- including all references;
- making the documentation more framework centered and user-friendly (e.g., clarifying the step-by-step instructions as well as the relationship between the manual and framework);
- redesigning the manual so that it will better support the framework and instrument (e.g., an analogy could be a tax guide and tax form)
- expanding the explanation of the interdependency across the dimensions (e.g., the tradeoff between accuracy and timeliness); and
- examples will be included.

Although the CIHI-DQF, v1 appears to compare favorably in many respects to other frameworks, certain 'big picture' changes, consistent with Eppler and Wittig's recommended future directions, are being considered. First, at present an evaluation based on the CIHI-DQF, v1 is more representative of whether an important list of criteria have been considered rather than an actual qualitative evaluation of the data quality or, by extension information quality, of a database. More qualitative versions are currently under development.

Second, future development could also involve the incorporation of the entire error model. For example, the CIHI-DQF, v1 does not cover possible errors made between the time of the clinical action (e.g., a clinical intervention) and the chart documentation, nor does it consider the critical importance of system design. While the CIHI-DQF, v1 has been found to be systematic and concise, it does not provide a scheme to solve detected problems and its ability to facilitate proactive data quality management needs enhancement. Suggested improvement schemes and proactive management approaches will be explored.

Finally, the cost of framework implementation must be addressed. Outside of the cost of a new data quality unit responsible for the development and support of the framework, within CIHI experience to date suggests that most of the resources, e.g., database managers, analysts, technical support, and documentation, required for framework implementation are already in place. Other than the initial study time required to learn the revised standard approach, it appears that no additional resources have been necessary. One preliminary observation, however, might be that the cost of implementation, as measured by the time required for existing staff to complete an evaluation, varies with the level of methodological expertise available.

The importance of establishing the cost of implementation within CIHI has resulted in an effort to track relevant time and resource use. Future plans include the development of an external version of the framework, hence an understanding of the issues and costs involved may be of interest to other health care sector settings, e.g., hospitals or clinical research institutions.

In addition to existing data quality efforts, a new CIHI data quality section (three methodologists, a classification expert, some administrative support, and a manager) has been put in place and is devoted to studying data quality issues, framework and methodological development, and support. While replicating such a unit in many external health care settings would be costly there may be no need to do so. Although the framework has not yet been implemented externally, some additional observations based on our experience can be made. Primarily what is required for implementing such a framework is methodological (i.e., statistical or epidemiological) expertise. However, what is fundamental, in addition to readily available methodological expertise, as well as the basic infrastructure required for a clinical or administrative database, is senior management commitment, active sponsorship for the idea, and an assurance of commitment and resources in all operational plans. Such factors have proven crucial for successful framework implementation at CIHI and thus should be considered for any external implementation.

### **Potential Limitations of the Study**

It is recognized that the main weakness of this study is that the meta-evaluation methodology was based on only one article (i.e., Eppler and Wittig (2000)). While the findings of the meta-evaluation were congruent with our applied experience, we acknowledge that the field of meta-evaluation is in its infancy. In fact, other than Eppler and Wittig, prior work in the field of data quality framework meta-evaluation appears to be unavailable. To interpret results with confidence, a meta-evaluation must be based on a solid body of literature.

In addition to a call for more meta-evaluation work, data quality practitioners would benefit from comparative analyses of the different types of meta-evaluation methodologies, i.e., meta-meta-evaluations or meta<sup>2</sup>-evaluations. Nevertheless, conducting the Eppler and Wittig meta-evaluation combined with applied results is a good start. It is hoped that this paper will not only elicit feedback regarding the CIHI-DQF, v1, but will also help to stimulate the field of framework meta-evaluation as well as meta<sup>2</sup>-evaluation. The need for more in-depth, rigorous, and complete meta-evaluation methodologies is obvious, especially in health care where quality information is crucial.

Another limitation of this study might be that the literature search was not comprehensive enough. That said, data quality research seems to be spread across numerous fields and the search and acquisition process was found to be challenging. An additional recommendation based on this study could be a reiteration of the importance of *The Data Quality Journal* and the annual MIT IQ conference and proceedings as centers of excellence for practitioners of data quality. Lastly, the omission of an explanation of the difference between the concepts of ‘information quality’ and ‘data quality’ might be interpreted as another limitation. Due to time and space constraints, and for the purposes of this study, no definitions were provided except to state that information quality follows from data quality.

### **Recommendations for Future Research**

Other recommendations based on this study echo the call sounded by Huang, Lee, and Wang (1999), as well as, by Eppler and Wittig (2000) for improved concept definition and standardization for the field. Even the definitions of framework components, such as the difference between a criterion and a characteristic, could benefit from standard definitions. One suggestion could be a dictionary for the field of data quality much like the *Dictionary of Epidemiology, 4<sup>th</sup> Edition*<sup>xvii</sup> is for the field of Epidemiology. Such a dictionary might draw on and collate several disciplines where work on the important concepts (e.g., accuracy) has also been conducted. Moreover, while Eppler and Wittig provide an excellent start, given the vital importance of quality information, especially in health care where lives can be affected, and consistent with the principle of CQI, meta-evaluation methodology development is necessary.

As a final point, whereas senior management at CIHI understands the importance of data quality, framework development, and meta-evaluation research, some in the health care field may not fully recognized the impact of adequate data quality or conversely the impact of poor quality data and by extension, information. The study and communication of the extent, impact, and resolution of data quality, and hence information quality, must be more forcefully pursued. Nowhere else may this be more pertinent than in the health care sector where critical decisions are being made and lives may be in the balance.

**Conclusion.** In summary, the CIHI Data Quality Framework, Version 1 (CIHI-DQF, v1) is a new framework and its evolution was anticipated. The CIHI-DQF, v1 was evaluated and found to rank relatively well when compared to other frameworks. The framework was found to be both relatively strong theoretically as well as practical and reasonably generic. Despite a relatively favorable meta-evaluation, several aspects of the CIHI-DQF, v1 are slated for improvement (e.g., more explanation of the trade-offs involved across quality dimensions). Given the importance of quality information, especially in health care where life and death decisions may be involved, it is surprising that so few generic frameworks seem to exist and meta-evaluation methodologies seem almost nonexistent. It is also recommended that framework and meta-evaluation development be flagged as crucial with the ultimate objective of improving information especially in the health care sector.

## References

- <sup>i</sup> Strong, D.M., Lee Y.W., & Wang, R.Y. (1997) 10 Potholes in the Road to Information quality, in: *Computer IEEE*, pp. 38-46.
- <sup>ii</sup> CIHI Data Quality Framework, Version 1, April 2001, CIHI.
- <sup>iii</sup> National Consensus Conference on Population Health Indicators, 1999, CIHI.
- <sup>iv</sup> [www.cihi.ca](http://www.cihi.ca)
- <sup>v</sup> CIHI Products and Services Catalogue 2001, CIHI.
- <sup>vi</sup> Deming, W.E., *The New Economics for Industry, Government, and Education*. W. Edwards Deming Institute, 1994.
- <sup>vii</sup> Deming, W.E., *Out of the Crisis*, W. Edwards Deming Institute, 1986.
- <sup>viii</sup> Brackstone, G. J. (1987) Issues in the Use of Administrative Records for Statistical Purposes, *Survey Methodology* 13.
- <sup>ix</sup> Brackstone, G. (1999) Managing Data Quality in a Statistical Agency. *Survey Methodology* Vol 25(2) pp, 139-149.
- <sup>x</sup> U.S Federal Committee on Statistical Methodology, *Statistical Policy Working Paper 4 – Glossary of Nonsampling Error Terms: An Illustration of a Semantic Problem in Statistics*, 1987.
- <sup>xi</sup> Statistics Canada Quality Guidelines, 3<sup>rd</sup> Edition, October 1998, Statistics Canada.
- <sup>xii</sup> Huang, K.T., Lee, W. L., and Wang, R. Y., *Quality Information and Knowledge*, Prentice-Hall, October 1998.
- <sup>xiii</sup> Eppler, M.J. and Wittig, D. (2000) Conceptualizing Information Quality: A Review of Information Quality Frameworks from the last Ten Years. *Proceedings of the 2000 Conference on Information Quality*. Eds Klein D. and Rossin D. F. IQ-2000 MIT Cambridge, Massachusetts, USA. pp 83-91.
- <sup>xiv</sup> CIHI Data Quality Manual, Version 1, April 2001 p 48.
- <sup>xv</sup> CIHI Data Quality Framework Evaluation Instrument, Version 1 Users' Guide, CIHI 2001.
- <sup>xvi</sup> Iezzoni, L. I. (1997 ) Assessing Quality Using Administrative Data. *Annals of Internal Medicine*. Vol 127(8 part2), pp 666-674.
- <sup>xvii</sup> Last, J. M. *A Dictionary of Epidemiology*, 4<sup>th</sup> Edition. Oxford University Press, 2001.



## **Tracking the Physical and Information Product Flows in Mobile Patient Service Supply Chain: A Real-Vision Lab Approach<sup>1</sup>**

<b>P. Balasubramanian</b> Assistant Professor School of Management Boston University 595 Commonwealth Avenue Boston, MA 02215 email: <a href="mailto:bala@bu.edu">bala@bu.edu</a>	<b>G. Shankaranarayan</b> Assistant Professor School of Management Boston University 595 Commonwealth Avenue Boston, MA 02215 email: <a href="mailto:gshankar@bu.edu">gshankar@bu.edu</a>	<b>R. Wang</b> Associate Professor School of Management Boston University 595 Commonwealth Avenue Boston, MA 02215 email: <a href="mailto:rwang@bu.edu">rwang@bu.edu</a>
---	---	--

**Abstract:** In this research we have presented the concept of a Virtual Business Environment (VBE) that supports dynamic decision-making and examined the implications for data quality in such environments. We have motivated the need for such environments using the operations and patient flows in a hospital. We have further described the critical need for high quality information and the need to track and measure quality in such environments where real-time data is collected and used. An important issue here is the need to seamlessly integrate real-time data and data collected by other traditional means. We have proposed an architecture that addresses this requirement. Visualization is a technique that plays an important role in managing information quality. We have proposed the notion of information product maps (IPMAPs) as a modeling method for representing the creation, processing, and consumption of information products in these environments. Quality dimensions incorporated into the IPMAP permit the information manager to examine the quality of the product under different scenarios. This examination can be visually performed using a VBE for managing information quality.

---

<sup>1</sup> This research is sponsored by BUILDE. We would like to thank Shakoov Jilani, Theresa Meyer and Max Bessanov for their research work on this project.

## **1. Introduction**

Information technology (IT) experts have been pointing out that technology exists to bring the practice of healthcare into a better digital shape for the 21<sup>st</sup> century. They envision patients all over the country accessing secure medical records and setting up appointments, shopping for the best hospitals, looking up lab results, tracking claims, or consulting with a specialist electronically. From the provider's perspective, patients could be electronically monitored, and medical information shared with other care givers, public health threats identified before they spread, and medication errors reduced by automating order processing and data entry. While many individual components in health care systems are now computerized, techniques for aggregating and storing this data in a system, methods for guaranteeing the quality of the data, and systems permitting effective transmission of information to the right people and the right time lag far behind.

There are several reasons for the lag. With many hospitals struggling to make ends meet under managed care and insurance reimbursements for many treatments being cut in recent years, investing in new data and computing systems is often a low priority [14]. Secondly, complex legal, cultural, and social barriers impact the implementation of such systems. Finally, health organizations are facing difficulties in managing and controlling the quality of large volumes of medical data that is required to support their complex decision-making tasks. This is because data is captured by many different sources, stored in a variety of different systems, and transmitted over networks that transcend organizational and system boundaries. This is further complicated by the advances in wireless technology that now permits capture of real-time data. The data captured by such technologies as the 802.11B networks, radio frequency tags, and infra-red sensors provide richer content in that they can capture the location of the source (in the context of the network) and can potentially monitor the source on a 24x7 basis.

Advances in technology have helped to solve some of the traditional problems of information quality by obviating the need for data entry and by automating the transcription of information. These also decrease the time interval between capturing information and making it available for consumption. On the other hand, the very same technologies also pose some challenges. Clearly, the volume of data that can be captured will be considerably higher – some of this is critical and some not useful at all. What are the useful data elements that can be captured by these technologies? How do we know if we are capturing the “right” data? What new data quality standards are required to ensure high quality data? Understanding the implications of these and other such questions provide some interesting challenges to researchers. Addressing these issues is important because it is believed that the availability of real-time data and its seamless integration with data captured using traditional technologies may significantly impact decision-making in complex environments.

In our research we propose the use of a Virtual Business Environment (VBE) [1] to support dynamic decision-making in complex environments. Conceptually, a Virtual Business Environment can be defined as a suite of integrated applications (processes) and tool sets which support specific, major business capabilities or needs. The VBE provides decision-makers with integrated and seamless access to all the business capabilities required for analyzing and executing business decisions. It includes the required technology infrastructure for capturing and managing the data needs for decision-making. An important component of the VBE is visualization of real-time data. Visualizing data is increasingly becoming an accepted technique for assisting complex decision making processes. Visualization has gained acceptance because it provides a way for senior decision makers to understand and view the results from a complex

decision model (e.g. simulations) in a more intuitive fashion. Complex models and results from these have had to be interpreted by model builders and knowledgeable users for it to make sense. The data being visualized is often real-time data that is streamed in from production databases or directly from the source(s).

To successfully implement a VBE for decision-making using real-time data, three important issues need to be addressed. First, a clear determination must be made about the data to be captured and the technology used to capture it. In the case of remote data collection, we must decide on the kinds of sensors (802.11b, infra red, etc.) to be used and their deployment. Second, we must understand how to seamlessly integrate real-time data from wireless networks with data that is captured using traditional mechanisms and technology. This data must be managed in such a way that it can support the wide variety of decision models that are employed in these environments. Third, we need to examine the implications for information quality for dynamic decision-making in such complex environments. The data/information captured, processed, and managed must be of superior quality because it is used in dynamic decision environments with techniques like data visualization. An important objective in such environments is that decisions can be made with a better understanding of the complex decision model and that decisions can be made sooner. If the data streamed in does not conform to high quality standards it is bound to have a negative impact on the critical decisions made.

Our objectives in this paper are two-fold. First we propose a conceptual architecture that permits the collection, storage, and utilization of data in a complex decision environment. We describe an instance of this architecture for the hospital environment. We use this architecture to identify the implications for information quality in such environments. This is described by treating the information as a product, a methodology that has gained considerable acceptance in the recent past [5, 7, 10, 22]. Specifically, we develop an information product map for a sample of the information used here and to help us identify quality implications and measures to ensure high information quality. Second, we describe the VBE in the context of the Real-Vision Lab (RVL), a facility that would support data visualization and the creation of VBEs.

The rest of this paper is organized as follows. We present an overview of the information product map (IPMAP) and its implications for information quality in section 2. In section 3 we outline the mobile patient project. A brief description of the patient service chain and the information flow is in section 4 along with IPMAP representations of two sample products. The Real Vision Lab is presented in section 5. A information technology architecture for creating a VBE in this lab is also described along with the implications for information quality management. We conclude the paper in section 6 and suggest directions for further research.

## **2. Overview of Relevant Research**

The concept of managing information as a product as a method for improving information quality in decision-making environments has received considerable attention in the recent past. Several research papers that deal with various aspects of this concept have been presented. Significant among these are the definition of quality measures for information products [5, 7, 10], principles for managing information as a product [22], specifying practices for continuously improving the processes involved [9], and identifying benchmarks for information quality [13]. A fundamental notion that underlies all of the research in this area is that the final information product is a document/artifact that is sent to the consumer and examples of information products used include birth certificate, eyeglass prescription, student transcript, hospital bill, or bank statement. Analyzing information as product raises some interesting questions. For example, what if the information product is produced and consumed at the same time? What if the

information product is used in performing “what if” analysis that is typical in a decision-making environment? In such cases, as the product is being “consumed”, the consumer may decide to modify some inputs/processes to re-create the product with perhaps a different “flavor”? While we assume here that the consumer has some control over the manufacture of the product (not uncommon in such environments), we also need to understand that we are dealing with a dynamic environment in which the product is being created and more importantly, consumed simultaneously.

A method for representing the manufacture of the information product, the IP-MAP, has been described in [20]. It provides a systematic method for representing the processes involved in manufacturing (or creating) the IP by extending the information manufacturing system model proposed by Ballou et al. [5] to develop a formal modeling method for creating an IP-MAP. This representation offers several advantages. First, using this representation, the IP manager will be able to visualize the most important phases in the manufacture of an IP and identify the critical phases that affect its quality. Second, the conceptual representation would allow IP managers to pinpoint bottlenecks in the information manufacturing system and estimate the time to deliver the IP. Third, based on the principles of continuous improvement for the processes involved, the IP-MAP representation would not only help identify ownership of the processes at each of these phases but would also help in implementing quality-at-source. Fourth, the representation would permit IP managers to understand the organizational (business units) as well as information system boundaries spanned by the different processes used in the manufacture of the IP. Finally, it permits the measurement of the quality of the IP at the various stages of the manufacturing process using appropriate quality dimensions.

The information product that we described above will be produced and managed by an environment that we call a Virtual Business Environment (VBE). There are several reasons why it is important to understand the implications of information quality in such environments as the VBE: (1) the decision environment is complex and dynamic where critical decisions need to be made quickly (such as dynamic routing/scheduling in hospitals). (2) Real time data is streamed in and typically there is no time to examine and correct data errors and (3) data is collected and integrated from multiple sources that span organizational and system boundaries necessitating the need for tracking and identifying all data elements and the business units/organizations responsible for capturing / processing / storing each. The IPMAP representation is useful for understanding the information quality implications and controls needed in such environments. Using the IPMAP we can represent the sources, sequentially identify the processes that transform data from these sources into “components” that are assembled to create the final product, the storage mechanisms where the source data and/or components reside during the manufacture, and for each of the above identify the business unit(s) and information systems associated. This would help us “visualize” the creation of the information product, identify the critical points in the manufacture where quality measures/checks need to be performed, and define (subjectively) the quality standards required at each stage of the manufacture. This would ensure that the final product satisfies the specified quality standards defined for each product.

### **3. Mobile Patient Information Project**

The overall objective of the mobile patient information project (MPIP) is to understand the technology, examine its implications on information quality, and gain useful insights that will enable us to answer some or all of the questions listed earlier. Specifically, we will explore how a wireless, sensory network can track patient location and movement within a hospital, and how

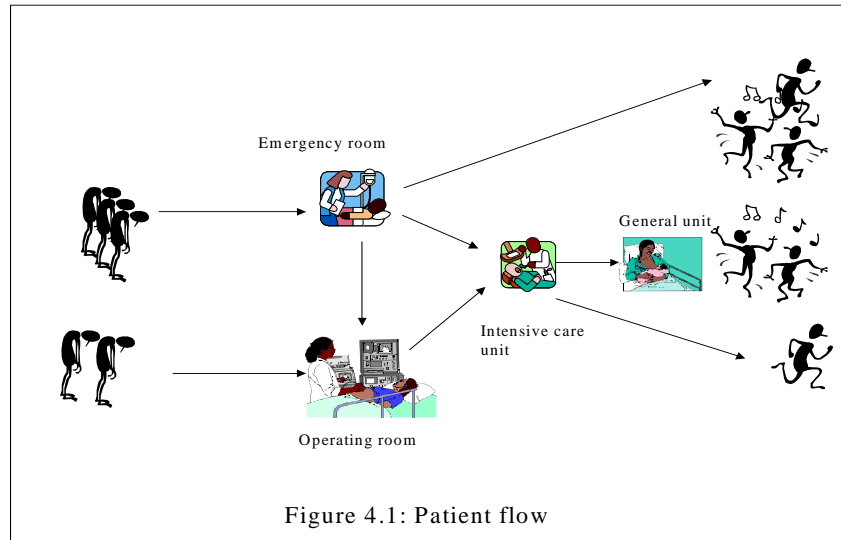
this information can be integrated with existing patient information to help improve hospital operations, achieve optimal utilization of resources, and ultimately provide superior patient care.

One motivating factor for this undertaking is the widespread use of wireless networks and technologies. We have briefly discussed this in section 1 and will take this issue up again in section 4. The other motivation stems from the complex and yet unsolved problems that hospitals and public health departments are facing today. Services provided by hospitals in the United States have come to be viewed as commodities. In an effort to keep costs down, many jobs and services deemed peripheral have been cut, often at the cost of quality care. Realizing the need to compete on a basis other than cost, and in the face of real or potential financial loss, hospitals are now starting to examine the potential of operational models for improving efficiencies in patient flow.

For example, one of the greatest sources of lost revenue in hospitals is the improper utilization of resources such as operating rooms and intensive care units (ICUs). This problem persists, even as waiting times rise and perceived quality of service drops. Currently, bottlenecks often result as semi-autonomous departments focus on optimizing local throughput without considering how their actions affect the performance of other departments [16]. Even the most efficient operations are subject to the woes of variability. A hospital environment is not only subject to the variability of patient types, arrivals and behavior, but also to a great degree of variability in the processes it uses to provide care. Poor management of resources can serve to intensify systematic variability, further straining the system during peak periods. These “artificial” sources of variability, however, can be eliminated through the use of effective cross-departmental scheduling [15]. With this, the focus can shift to resolving bottlenecks dynamically during those peak periods caused by systematic variability.

Perfect information has the potential of adding enormous value to a complex system such as that in a hospital. Complex decision-making models have been developed to manage the allocation of limited resources in the hospital environment. The accuracy of these models currently depends upon the quality of simulated data. It has been impossible up until now to obtain a comprehensive record of patient movement and location within a hospital, therefore the results of these decision-making models have been based upon an incomplete picture. Access to the total stream of patient movement allows us to explore in more detail the intricacies of the system and may reveal bottlenecks of which we were previously unaware. This information can then be used to coordinate the use of shared resources and ultimately to transform operations, not only in the predictive sense, but also in real time.

Real time data that is directly captured off the network (wireless or otherwise) has two inherent advantages. First, as manual processes are not part of the data capture or data entry and hence the data is more “clean” than in cases where manual transcription or data entry is required. Second, the data is available for use virtually immediately with no delays. This has an important implication for hospital administrators attempting to understand the causes for “traffic jams” in operating rooms, overcrowding of ICUs, and diversion of ambulances from ER due to the operating at maximum capacity. They can use the real-time data captured to replay scenarios and over time identify “events” using which they can predict when such a situation is going to occur and take precautionary measures to resolve it or to avoid it altogether. Visualization can play a vital role here by allowing the administrators to view a “map” of the hospital layout, examine events as they occur by tracking the movement of patients on the map, and visualize the circumstances that triggered overcrowding and poor utilization.



Wireless technology can be used to “tag” the patient possibly at the time the patient check-in or registers into the hospital. For very critical patients brought into the ER this may not happen until after the initial examination or emergency treatments. The “tag” identifies the patient uniquely and may be worn on the patient’s person, for example, on his/her wrist similar to what is in practice today. The other data about the patient such as the patient history (prior admissions/treatment records) can be linked with this identifier tag. As the patient is examined by doctors/interns, the data diagnostic information, procedures to be scheduled, and treatment recommendations can be captured and linked with the patient identifier. These data (history, treatment, etc.) are currently being captured in electronic systems using manual data entry (nurses or by physicians themselves). The proliferation of wireless palm tops and PDAs may very soon have this data capture performed wirelessly at the patient’s bedside. Monitoring devices that track patient’s vital signs can also be linked to the patient identifier resulting in all of the patient’s information being tracked with the patient seamlessly.

This collection of information has several advantages to hospital administrators and physicians alike. Accurate information about patient flow through multiple departments in the hospital environment could allow for more precise analyses of demand, and therefore more effective scheduling of shared resources. It could also allow for immediate action in a dynamic scheduling environment. If the system reveals that there is a queue of patients waiting for x-rays, but no wait for MRI, patients who need both may be diverted to the unused resource to balance utilization across departments. From the physician’s perspective, the benefits of having an integrated collection of “clean and accurate” information ranging from patient history all the way to current conditions are easily recognizable and needs no further explanation.

#### 4. Patient Services and Information Flows

A mobile patient service chain is the set of activities that are performed to serve a patient. To understand the information and architecture needs (positions for sensors, meta-data definitions, etc), we must understand the physical and informational flows in patient services. A generic physical flow is shown in Figure 4.1. An emergency patient arrives directly at the admissions area of the emergency room for an unscheduled visit. An ER physician determines if

further care is needed or if the patient can return home. If the patient needs further attention, the process is similar to a scheduled patient arrival that is described next. The scheduled patient arrives for a planned procedure and might pass through five stages: scheduling and registration, preoperative care, the procedure followed by recovery in the ICU, and then a few days of recovery in the general floor.

To understand the informational view of the same patient, consider for example the operational status report and/or the patient care status report generated for hospital management once every half-hour. Based on these reports the management may dynamically schedule patient flow or change resource allocation to achieve better management of hospital resources and better patient care. The IPMAP representation for both these products are shown in figures 4.2, 4.3, and 4.4. Figure 4.2 shows the capture, processing, and storage of patient admission information. The registration office obtains personal information about the patient as well as information needed for emergency contacts and billing (Data Source DS<sub>1</sub>). Medical records for that patient may be obtained from other sources such as personal physician's office or other health care agencies (DS<sub>2</sub>). The patient is then examined and the initial patient conditions are also captured (DS<sub>3</sub>). The patient is then allocated a bed (in the ER/ICU/floor) that is also captured in the system (DS<sub>4</sub>). The latter two may be done using a palm top/PDA in a wireless network. All of this goes into the patient medical record storage (STO<sub>1</sub>). In the figure, raw data from sources is indicated by RD and processed data by CD, each with a suffix assigned in sequential order for identification. Data, during processing may move across multiple systems, some paper-based and some electronic. System Boundary (SB) blocks are used to explicitly capture this. Processing (P) blocks are used to represent data processing and quality blocks (QB) represent checks performed for validating the data. Business boundary blocks (BB) are used to represent the flow of data across organizational or business units and in cases where transfer from one business unit to another also implies transfer between two different systems a combined business and system boundary block (BSB) is used to represent it.

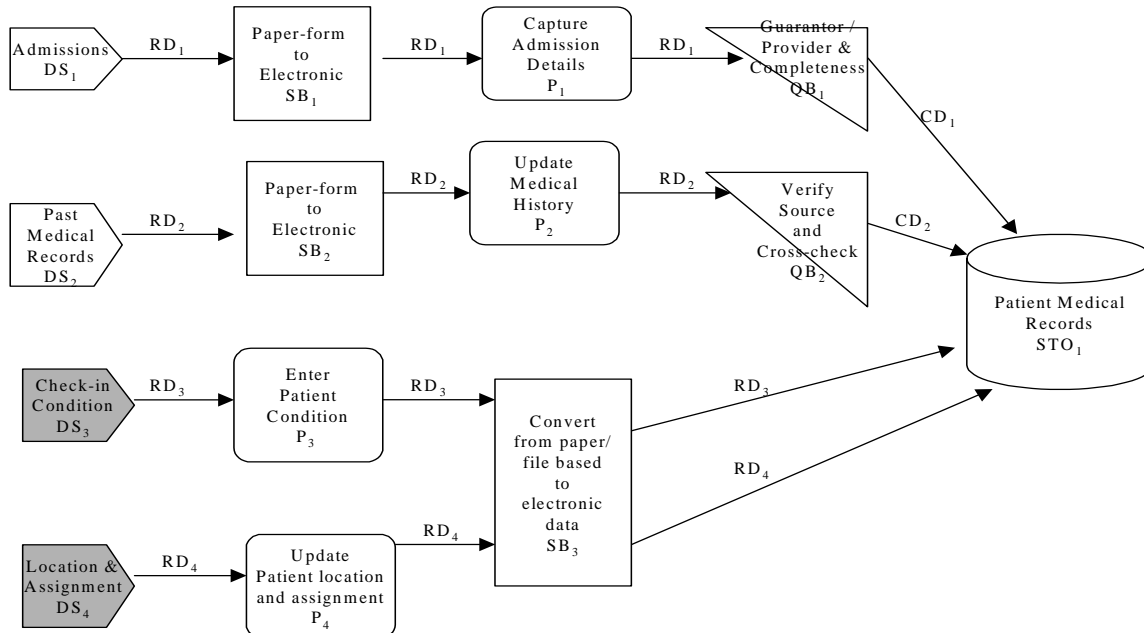


Figure 4.2: Capture, Processing, and Storage of Patient Admissions Information

Figure 4.3 describes the capture, processing and storage of patient treatment and care information. Lab/Radiology records and results (DS<sub>5</sub>) and information on surgical procedures performed (DS<sub>6</sub>) are captured into systems in corresponding departments and transferred into the patient treatment database. Further, recommendations from specialists (DS<sub>7</sub>), progress reports from attending interns (DS<sub>8</sub>), and vital signs continuously monitored by wireless devices (DS<sub>9</sub>) would also become part of the treatment database after necessary processing.

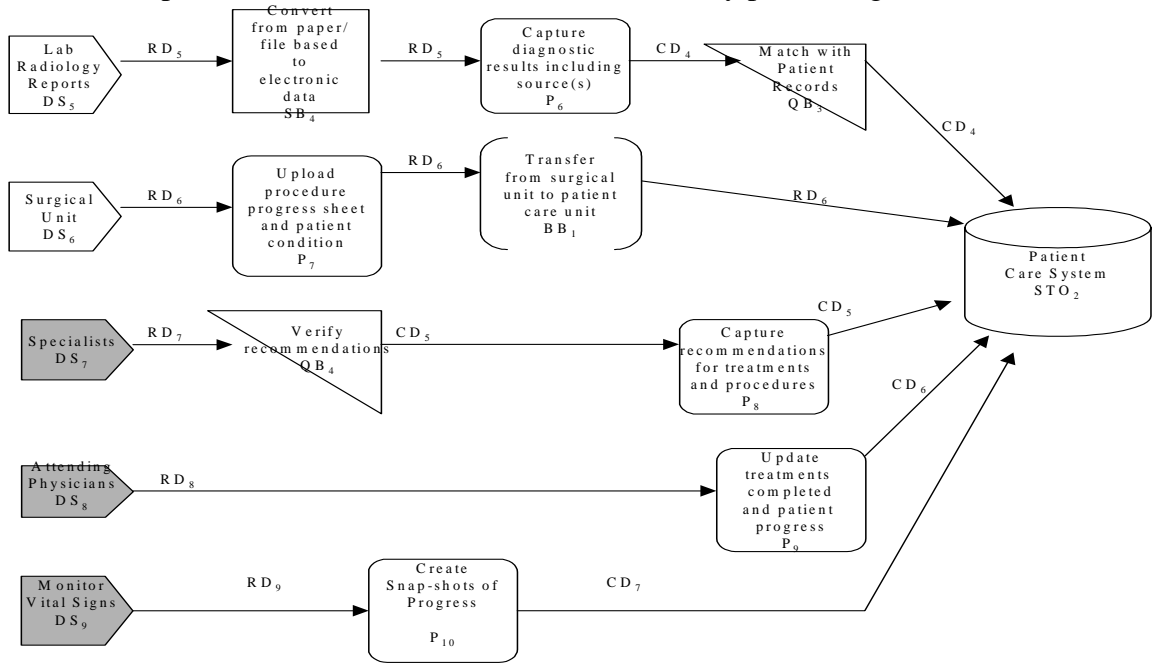


Figure 4.3: Capture, Processing, and Storage of Patient Treatment/Care Information

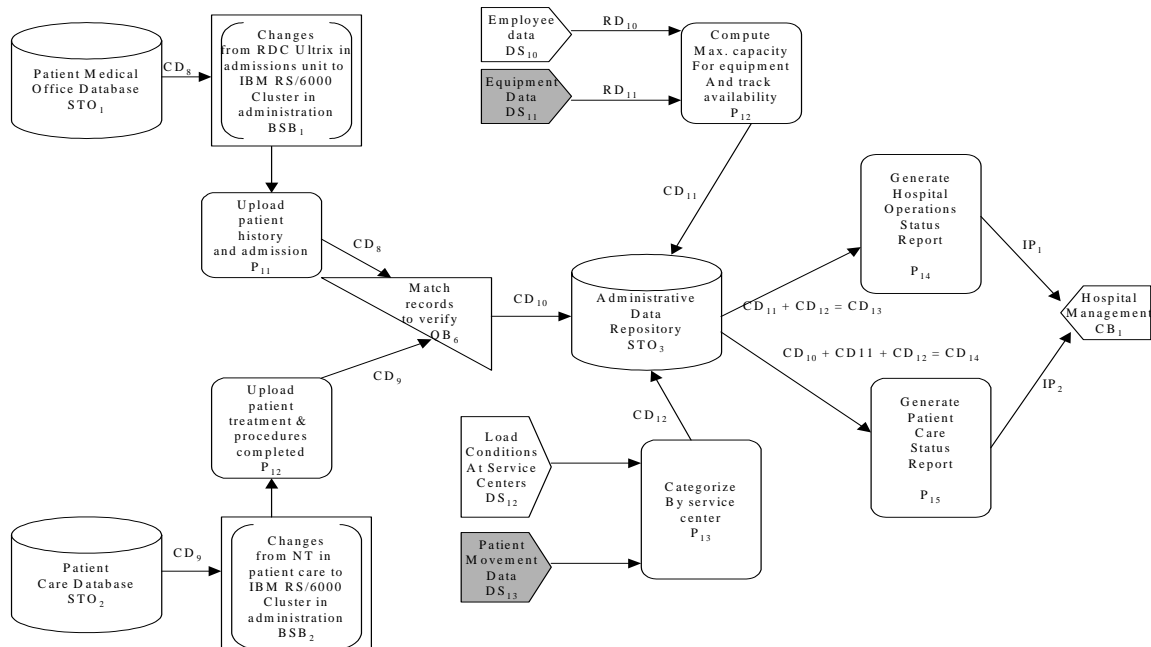


Figure 4.4: The creation of IP<sub>1</sub> and IP<sub>2</sub> for dynamic scheduling and patient flow monitoring



Combining this with data about employees (DS<sub>10</sub>) and equipments (DS<sub>11</sub>), data on load conditions at the various service centers (DS<sub>12</sub>) such as the MRI /CAT/X-Ray, as well as the data on patient and equipment movements (DS<sub>13</sub>) in the administrative data repository (STO<sub>3</sub>), the system creates the two information products. The first product (IP<sub>1</sub>) is the status report on the hospital's operational processes which can be used for dynamically allocating or re-allocating resources as well as identifying utilization of the different critical care centers. The second product (IP<sub>2</sub>), the status report on patient care will inform the administrators about the location /movement of the patient. Together with patient treatment and medical records, the administrators can determine how the patient should be cared for next keeping in mind the utilization of the different service centers. The data source blocks that are shaded in gray represent information that is captured by wireless or similar networks using infra red or radio frequency tags that may be attached to patients/equipments as described earlier.

<i>Name/Type</i>	<i>Department/Role</i>	<i>Location</i>	<i>Business Process</i>	<i>Base System</i>
Admissions /DS <sub>1</sub>	Admissions Office/ Patient	Admissions , OB/GYN, Emergency	Standard Form (#1101P)	Paper-based - Patient File
Past Medical Records / DS <sub>2</sub>	Admissions Office / Admissions clerk	Admissions Bldg., Records Room	Contact source and request with patient authorization.	Paper-based - patient file

**Table 1: Sample metadata for IP-MAP in figure 1**

The informational view is captured as meta-data. To complete the representation, we need to capture the information about each of the blocks and the data elements included in each flow in the model(s) above. This is akin to the data dictionary for a data flow diagram and we refer to this as the metadata associated with the model. The metadata repository resides within the data layer of the system architecture described next. For brevity only a sample of the metadata is shown in table 1. Besides metadata, each block in the IPMAP can include a set of quality dimensions such as completeness, accuracy, reliability, and timeliness. The information manager can assign (subjectively, with the help of users if necessary) weights to each of these dimensions. Using methods similar to the one described by Ballou et al [5], the overall quality of the information product(s) can be computed and visualized using appropriate metaphors.

## 5. Managing Information Quality in the Real Vision Laboratory

As a proof-of-concept, we are developing a VBE to support decision-making within the healthcare context. This environment will house the large quantities of data that will be collected by the sensory layer. The metadata about the quality of information and the databases will also be stored within the environment. Finally, the environment will also provide the visualization support for decision-making. Currently, we are in the process of developing the RealVision Laboratory (RVL) based on the architecture described. This lab will primarily use off-the-shelf (COTS) technology, and software and would serve as a platform to support and build multiple VBEs. Each environment will have a conceptual structure shown in Figure 5.1 below.

**Domain Resources Manager** is a subsystem that is responsible for managing the multiple resources such as expertise (knowledge), models, and data. Knowledge is captured here in the

form of documents, discussion threads, and other subtly structured data that cannot be managed with traditional data management techniques.

**Dialog Manager** is responsible for converting the answers to the queries as determined by the Engines into outputs for the user. The outputs can be of various forms as determined by the metaphors and can be rendered in different ways as determined by the visualization technology used.

**Business Context Engines and the Engine Manager:** In addition to Environments, the Process Domain typically contains a number of major software components called Business Context Engines. An Engine is defined as an analysis object that represents and implements a complex business capability requiring the integration of a variety of knowledge resources. Engines are shareable or reusable in disparate business specific applications or contexts. An Engine implements a particular, generic, business algorithm for the processing of a specific class of business Data. An Engine becomes useful by working on the Data that is relevant to the Environment(s) to which the Engine is "subservient". In this way, an Engine is able to provide diverse Environments (and consequently diverse business processes) with a standardized set of computational capabilities and/or evaluation functions. The Engines of this Architecture can be thought of as large blocks of reusable application code which instantiate complex business functions useful to one or more Environments. Engines typically integrates disparate data, synchronize processes to support get relevant information, normalize data to make fair comparisons, allocate resources dynamically, maintain business rules etc.

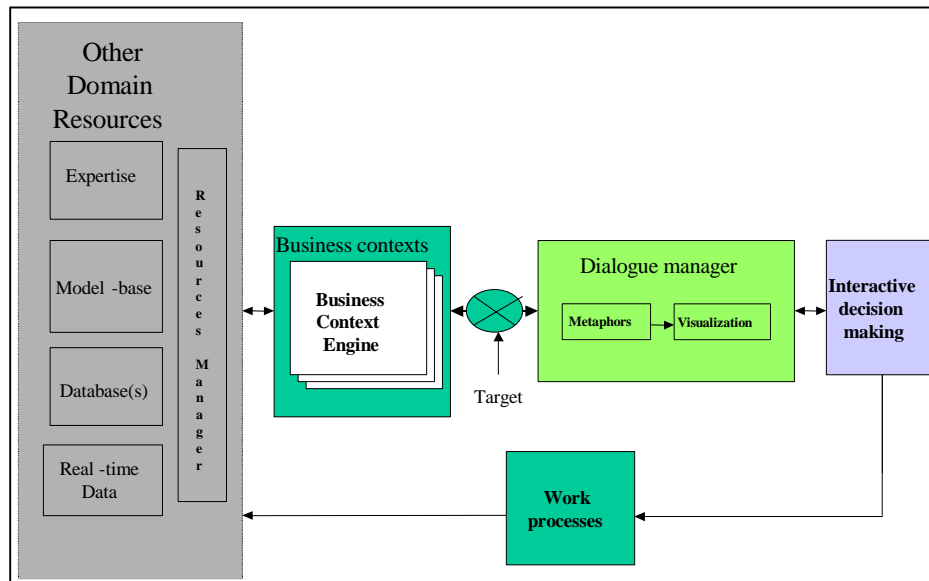


Figure 5.1: Conceptual Representation of a VBE

## Information Technology Architecture

As the enterprise continues to grow in size and complexity, several factors impede the ability of the enterprise to solve the problems that it faces. A point is rapidly reached at which there are too many factors that come into play in conducting the business of the enterprise. When dealing with such systems, designers have typically dealt with their complexity by breaking the problem into a set of smaller problems that are themselves less complex [8, 11, 23]. An architecture is a systems design that specifies the way the overall functionalities of the design

are to be decomposed into individual functional components and the way in which the individual functional components are to interact to provide the overall functionalities of the system design. *The decomposition of the enterprise into manageable parts, the definition of what those parts are, and the orchestration of the interplay among those parts are called the Enterprise Architecture.* The orchestration of the interplay is governed by a set of *design rules/Principles* or the organization's knowledge architecture [4, 19].

Data quality researchers have identified three stakeholders who participate the management of an information product: data collector, data custodian, and the data consumer [22]. This architecture integrates the interests of all three stakeholders and divides the world into three domains: decision network layer, data network layer and the sensory network layer. This collection of technology networks, data repository, and decision-making environments need to come together for us to benefit from tracking this real-time information. The architecture proposed next addresses this.

The sensory network layer has all the probes that collect information from the environment. For example, it could contain the RF network within a hospital room that collects information on patient movement or the GPS systems that help locate the ambulance and other assets that a hospital owns.

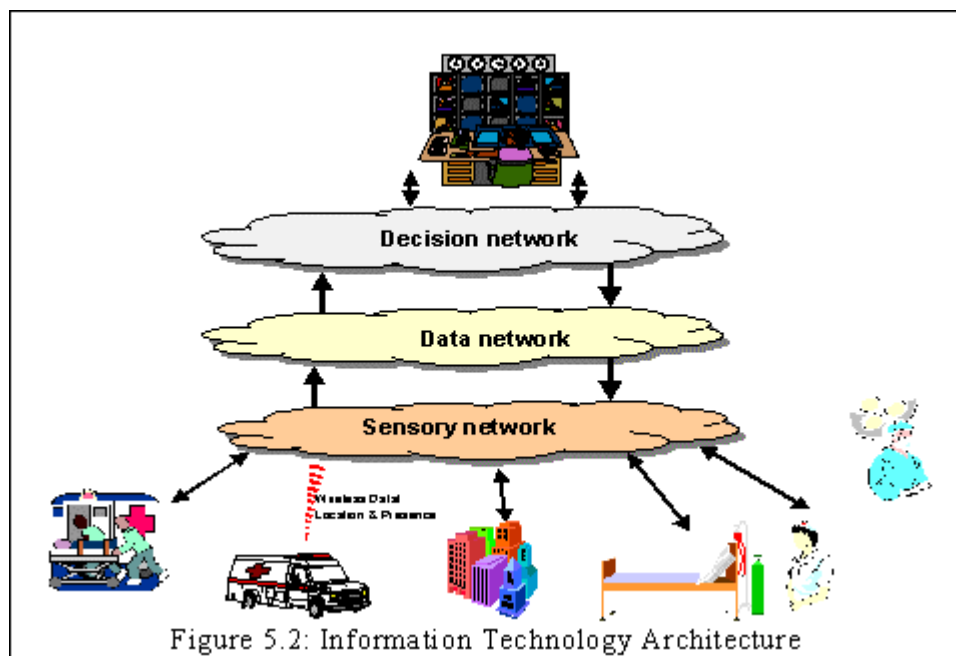


Figure 5.2: Information Technology Architecture

The data network layer has the schema information about the data that is stored in the enterprise. We use the term data broadly to include models as well. The meta information about models will be stored in this layer as described in [2]. Another example of information that may be stored in this layer is the metadata for the IP-MAP described in section 4.

The top layer is the decision network layer. This layer will contain all the processes, procedures, business tools, and rules to support decision-making within an organization. It includes applications needed for evaluating process design alternatives and their implications for cost and quality or models that can help determine staffing levels.

<i>The three C's in an information product management</i>	<i>The three network layers in the Real Vision Laboratory</i>
<ul style="list-style-type: none"> <li>• Data Collector (data creator, data collector, data entry)</li> </ul>	<ul style="list-style-type: none"> <li>• Sensory network (802.11b, RF devices, etc)</li> </ul>
<ul style="list-style-type: none"> <li>• Data Custodian (store, manipulate, and retrieve)</li> </ul>	<ul style="list-style-type: none"> <li>• Data Network (IP-MAP, database schema, etc.)</li> </ul>
<ul style="list-style-type: none"> <li>• Data Consumer (use information products for task at hand)</li> </ul>	<ul style="list-style-type: none"> <li>• Decision Network (metaphors and visualization)</li> </ul>

The above architecture and the VBE that is implemented using it will also support information quality management. Information that is aggregated, analyzed, and used in decision-making can be treated as a distinct information product and represented as an IPMAP. This offers several advantages to the information manager. The IPMAP associated with the different products are all captured in the data layer along with the metadata associated. Using visualization techniques, not only can we visualize the data in the information product but also the IPMAP for the product itself. By examining the IPMAP, the information manager can identify the sources of information, the organizational unit responsible for it, the individual(s) responsible for it, and more importantly, the organizational and system boundaries spanned by the manufacturing process, all of which are important when using real-time data. By subjectively assigning weights to quality dimensions (e.g. accuracy, timeliness, reliability, completeness etc.) at each of these blocks, the quality of the information processed at each block in the manufacture can be computed and visualized. A separate set of engines may be used to manage the IPMAP and its visualization. This engine will be part of the library of engines described earlier. Further more, by changing the weights of the quality dimensions on the IPMAP, information managers can visually examine the impacts of these changes on the final product. For example, hospital administrators may be visualizing the patient care status report and/or the hospital operations to determine the necessary re-routing or re-scheduling of resources with an eye on increasing efficiency and/or utilization. The information manager may simultaneously visualize the manufacture of these products (on a different screen(s)), and provide the administrators with the details on how good/reliable the information is and how it might impact decisions made using it including the best and worst cases.

## 6. Conclusions

In this research we have presented the concept of a Virtual Business Environment that supports dynamic decision-making and examined the implications for data quality in such environments. We have motivated the need for such environments using the operations and patient flows in a hospital. We have further described the critical need for high quality information and the need to track and measure quality in such environments where real-time data is collected and used. An important issue here is the need to seamlessly integrate real-time data and data collected by other traditional means. We have proposed an architecture that addresses this requirement.

Visualization is a technique that plays an important role in managing information quality. We have proposed the notion of information product maps (IPMAPs) as a modeling method for representing the creation, processing, and consumption of information products in these environments. Quality dimensions incorporated into the IPMAP permit the information manager to examine the quality of the product under different scenarios. This examination can be visually performed in a VBE for information quality management.

Currently, we are partnering with a technology vendor and a hospital to identify the requirements for a real-time environment to support decision-making within a hospital setting. We will apply the IP-MAP methodology to identify data quality requirement and implement processes to support it.

This study will help identify the specifications for the architecture described in section 5. For example, what types of sensors are needed in the sensory layer? What is the appropriate technology for the network (RF, 802.11b, barcode, etc.)? Metadata requirements for the data network layer. Finally, at the decision network layer, what tools and support technology are needed for rendering the complex business and real-time data to geographically dispersed users?

## References

- [1] P. Balasubramanian, R. Gottlieb, and R. Wang, "Virtual Business Environments: Concepts, Architecture, and Research Directions," Boston University, Boston, Working Paper March 2001.
- [2] P. Balasubramanian and M. L. Lenard, "Structuring Modeling Knowledge for Collaborative Environments," presented at 31st Hawaii International Conference on Systems Sciences, Hawaii, 1998.
- [3] P. Balasubramanian, K. Nochur, J. C. Henderson, and M. M. Kwan, "Managing process knowledge for decision support," *Decision Support Systems*, vol. 27, pp. 145-162, 1999.
- [4] C. Y. Baldwin and K. B. Clark, *Design Rules: The Power of Modularity*. Cambridge, MA: The MIT Press, 2000.
- [5] D. P. Ballou, R. Wang, H. Pazer, and G. K. Tayi, "Modeling Information Manufacturing Systems to Determine Information Product Quality," *Management Sciences*, vol. 44, pp. 462-484, 1998.
- [6] A. Bharadwaj, J. Choobineh, A. Lo, and B. Shetty, "Model Management Systems: A Survey," *Annals of Operations Research*, vol. 38, 1992.
- [7] A. T. Chun and B. Davidson, "Implementing the Information Quality Survey: A Case Study at Cedars-Sinai Health System," presented at Conference on Information Quality, Cambridge, MA: MIT TDQM Research Program, 1999.
- [8] T. DeMarco, *Structured Analysis and System Specification*. Englewood Cliffs, NJ: Yourdon Press, 1979.
- [9] L. English, "Plain English on Data Quality: Information Quality Management: The Next Frontier," *DM Review*, pp. 36-78, 2000.
- [10] J. Funk, Y. Lee, and R. Wang, "Institutionalizing Information Quality Practice: The S. C. Johnson Wax Case," presented at Conference on Information Quality, Cambridge, MA: MIT TDQM Research Program, 1998.
- [11] C. Gane and T. Sarson, *Structured Systems Analysis: Tools and Techniques*. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1979.
- [12] A. M. Geoffrion, "An Introduction to Structured Modeling," *Management Sciences*, vol. 33, pp. 547-588, 1987.
- [13] B. K. Kahn and D. M. Strong, "Product and Service Performance Model for Information Quality: An Update," presented at Conference on Information Quality, Cambridge, MA, 1998.
- [14] L. Landro, "Information technology could revolutionize the practice of medicine. But not anytime soon.," in *Wall Street Journal*, 2001.

- [15] E. Litvak, "The Program for the Management of Variability in Health Care Delivery," Boston University, Boston, Note 2001.
- [16] P. D. Mango and L. A. Shapiro, "Hospitals Get Serious About Operations," *McKinsey Quarterly*, pp. 74-85, 2001.
- [17] I. Nonaka and H. Takeuchi, *The Knowledge-Creating Company*. New York: Oxford University Press, Inc., 1995.
- [18] L. Prusak, "Knowledge In Organizations,". Newton: Butterworth-Heinemann, 1997, pp. 261.
- [19] R. Sanchez, "Modular architectures, knowledge assets and organizational learning: new management processes for product creation," *Int. J. Technology Management*, vol. 19, pp. 610--629, 2000.
- [20] G. Shankaranarayanan, R. Y. Wang, and M. Ziad, "Modeling the Manufacture of an Information Product with IP-MAP," presented at Conference on Information Quality, Massachusetts Institute of Technology: MIT TDQM Research Program, 2000.
- [21] R. H. Sprague and E. D. Carlson, *Building Effective Decision Support Systems*. Englewood Cliffs: Prentice-Hall, 1982.
- [22] R. Y. Wang, Y. L. Lee, and D. M. Strong, "Manage Your Information as a Product," *Sloan Management Review*, vol. 39, pp. 95-105, 1998.
- [23] E. Yourdon and L. Constantine, *Structured Design : Fundamentals of a Discipline of Computer Program and Systems Design*. Englewood Cliffs, NJ: Yourdon Press, 1986.

# **Non-Intrusive Assessment of Organisational Data Quality**

Binling Jin and Suzanne M. Embury

Department of Computer Science,  
University of Manchester,  
Oxford Road, Manchester, M13 9PL, U.K.  
{BJin|SEmbury}@cs.man.ac.uk

**Abstract:** Many organisations are becoming increasingly aware that the usefulness of their data is limited by its poor quality. Surprising (and sometimes alarming) proportions of data in databases are inaccurate, incomplete, inconsistent or out of date. One-off data cleaning methods can help the situation in the short term, but they are costly and do little to improve data quality in the long term. However, in order to plan and monitor the progress of long term data quality improvement programmes, it is necessary to be able to assess the quality of data across an organisation. Since resources for such programmes are generally limited, and since much of the data in question resides in mission critical systems, it is vital that these assessment activities do not intrude on normal day-to-day business processing.

In this paper, we present an approach to assess organisational data quality on a regular basis, which does not delay or disrupt revenue-generating data processing activities. We have adapted techniques from distributed query processing and distributed integrity checking to produce a system that takes account of the workload at each local site when distributing the defect checking work in the distributed information system. The approach assesses the data quality during the periods of low system activity, and ships data defects found to a global site, which time stamps them and records them for later analysis.

## **1 Introduction**

In recent years, many organisations have come to regard their accumulated data as a valuable asset that can be exploited in creative ways, to improve revenues and customer satisfaction. However, this increased use of data has also led to the discovery that much of it is of very poor quality [Redman1996]. Data is often found to be inaccurate, inconsistent, incomplete and out of date. One-off data cleaning efforts can help in the short term, but they are expensive and do nothing to improve data quality in the long term [English1999]. Ideally, organisations need to discover the causes of poor data quality and institute improvement programmes to remove them.

However, a fundamental element of any improvement program is the ability to assess the initial state of whatever we wish to improve. In order to plan an effective improvement programme, and make best use of the scarce resources available for it, an organisation must have a clear picture of the current quality of data in its systems. This allows the improvement effort to be directed at areas of greatest potential benefit, and provides a baseline against which the success of the programme can be gauged. In addition, it is necessary to reassess quality levels continuously at various points throughout the improvement programme, in order to track its effects and the degree of improvement achieved. What is required, therefore, is some means of regularly assessing the quality of an organisation's data at an acceptable cost, both in terms of staff

resources and processing time. Most importantly, since much of the data to be assessed will be stored in mission critical systems, it is necessary that normal (revenue generating) business processing is not disrupted or delayed by this assessment activity.

Although some aspects of data quality cannot be assessed by purely automatic means, levels of other aspects can be gauged by issuing queries over data, which search for data defects of a particular type. The number of defects located, relative to the amount of data searched by the queries, can give an indication of how far quality levels differ from the targets set by the organisation. For example, *consistency* and (in some cases) *accuracy* of organisational information can be assessed by issuing queries which compare data values stored in different information systems, while *completeness* can sometimes be determined by queries which identify records with certain null attributes, or where sets of records in different data sets cover different sets of real world data.

Given that certain forms of organisational data quality can be measured using cross-database queries, when are such queries to be executed? While it might be thought that they should be executed whenever an assessment of the data quality is required (i.e. at the beginning of the improvement programme, and at regular intervals throughout it) there are two disadvantages of this approach:

- Queries that assess data quality levels by comparing several data sets are typically very time consuming to evaluate. This means that execution of many such queries at one time requires significant data processing resources, which the owners of the databases involved may be unwilling to release (due to the potential delay to mission critical or revenue generation processing activities).
- Periodic assessment of data quality (e.g. every three months) provides a very imperfect and incomplete picture of the quality of data present in the system at the time of the assessment. For example, defects which entered the system since the last D.Q. assessment, but which were cleaned up before the next assessment activity, will not be detected. Also, the assessment provides only very coarse-grained details of when defects entered, or were removed from, the system (e.g. “some time in the last three months). Thus, much information vital to the correct interpretation of the root causes of data defects goes unrecorded.

An alternative strategy is to execute the queries continually (or, at least, whenever data is changed in a way that might produce a change in their results). This approach, which is essentially a more relaxed form of integrity checking [Nicolas1982], provides “perfect” information about data quality levels, at all times, but at an extremely severe cost to the organisation.

In this paper, we discuss a compromise approach in which system owners specify how much data processing resource they wish to give up to the data quality assessment activities, and at which times assessment can occur. For example, the owner of a system may be happy to allow defect detection queries to be run between the hours of midnight and 4.00am every day, when there is very little business processing activity for the system to support. We have adapted ideas from distributed query processing [Haas1997] and distributed integrity checking [Chawathe1996] to produce a system which schedules the distributed execution of data defect queries according to the workload patterns specified for each site, and which records details of any defects found, for



later analysis by the data quality improvement team. This approach has the advantage that assessment activities can be run often, thus providing a more complete picture of the changing levels of data quality, without intruding on important revenue-generating processing activities. However, it has the disadvantage that it may sometimes produce an inaccurate representation of the data quality levels, due to the delays which can arise in distributed query processing. We have therefore produced a cost model, which allows the system to allocate work between the distributed systems in a way that minimises the inaccuracies in the results.

The remainder of this paper is organised as follows. Section 2 discusses existing approaches to the assessment of data quality levels in information systems. In Section 3, we give an overview of our system for non-intrusive assessment of data quality, while in Section 4 we present its shortcomings in terms of accuracy and currency of the information generated, and describe how our cost model allows us to minimise these shortcomings. Finally, Section 5 concludes and makes some suggestions for future directions for this work.

## **2 Existing Approaches to Data Quality Assessment**

A number of approaches for assessing data quality (DQ) have been proposed. These approaches can be mainly divided into two classes: *assessment of subjective DQ* and *assessment of objective DQ*. The former measures individual stakeholders' subjective assessments of the DQ, typically using a questionnaire-based approach [Huang1999]. Analysis of the answers given to questions posed by a DQ questionnaire can help to identify specific problem areas, and to suggest actions which can be taken to monitor and improve the organisation's DQ. Subjective DQ measures are clearly valuable as a means of determining user satisfaction (or dissatisfaction) with their data. However, this method may not be suitable for measuring intrinsic data quality characteristics, as the results depend on the individuals surveyed and their intuition about what may be many years of experience working with a changing information system.

Assessment of objective DQ, on the other hand, focuses on measuring aspects of DQ that are concerned with "fact", rather than "opinion". For example, accuracy and currency of data are objective characteristics (a data item either is or is not 3 minutes out of date), while characteristics such as reputation and believability of data depend more on the eye of the beholder. One of the most common methods of objective DQ assessment is known as *database bashing* [Redman1996]. Essentially, this involves comparison of data in two or more databases that have some overlap in content. A typical application of the database bashing technique is to determine whether data quality levels are preserved through data feeds between systems. Database bashing can indicate where data fails to be consistent, accurate and complete, but the results must be interpreted with care, when for example inaccurate data is compared with inaccurate data or incomplete data with other incomplete data. It is also very expensive, and is too coarse-grained a method of measurement to really assist in the determination of root causes of data quality defects [Redman1996]. Database bashing is just one of a range of techniques often used in *data cleansing*; that is, the process of identifying and removing the defect from source data sets in order to prepare them for some new (unanticipated) use [Galhardas2001]. Other techniques used in data cleansing include the formulation of queries that search for hypothesised forms of defect, and manual inspection of samples of data or statistical summaries of data. While data cleansing can be very expensive in terms of staff time, it is often performed with only a minor impact on normal business processing activities, since it is common for data cleansing staff

to extract relevant sets of data from the original source systems (e.g. overnight), which can then be examined and corrected entirely off-line [Hernandez1998].

If the objective of the assessment is to gain an overall picture of data quality levels relatively cheaply, then examining every piece of data in the system and every process in which that piece of data is used is overkill. Two approaches have been suggested for this purpose: *statistical sampling* [English1999] and *data tracking* [Redman1996].

The statistical sampling approach involves the extraction of a representative subset of the data from the system, which is then examined (manually or with the help of a querying tool) for defects. Once the quality levels of this much smaller subset have been identified, they can be used to estimate the quality levels present in the system as a whole. In order to ensure that the analysis of the sample accurately reflects the state of the total data population being assessed, it is necessary to select a sufficiently large proportion of the data for the sample, while minimising the cost of the assessment process. Furthermore, it is usually recommended that anyone following this approach should spend some time determining which parts of the system are likely to be the best candidate for DQ improvement before the sample is selected [English1999, Bowen1998].

Data tracking differs from this statistical approach, in that it attempts to directly assess the effects of processes on data defect levels, rather than concentrating on the data only. When data tracking is used, the modifications made to selected data items as they are affected by some business process are tracked, in as much detail as is practicable. Thus, not only can defect rates be estimated by measuring the defects introduced in the sample (tracked) population, but we also gain considerable insight into exactly which parts of the process are responsible for creating them. However, data tracking cannot be used to gain a quick measure of data quality levels, as it must be carried out alongside the normal business processes, and cannot be simply invoked on demand [Redman1996].

One common feature of many of the methods proposed for assessment of objective DQ is that queries are formulated which search for defects in the data sets to be assessed. This is similar in practice (though not in intent) to techniques for integrity checking, in which queries are used to search for violations of the constraints. However, the aim of data quality assessment is simply to record details of the defects present (albeit often with the intention that these defects will later be corrected), whereas integrity checking has a much stricter aim of refusing to allow any modifications to data which would result in a violation of the constraint. Traditional integrity checking is unsuitable as a means of assessing data quality levels for improvement programmes. Checking complex constraints is a very expensive process, which inflicts the heaviest cost when transaction processing rates are highest. Moreover, they require strict adherence to all the constraints that they enforce, which is very limiting in the context of most messy, real world systems, where many constraints will have rare but legitimate exceptions [Caine2001].

However, the techniques which underlie integrity checking can be useful for data quality assessment, if adapted and weakened to suit this new application. In the remainder of this paper, we describe how we have adapted techniques from distributed integrity checking and distributed query processing, in order to produce a system which is more suitable for continuous, non-intrusive assessment of organisational data quality levels.

### 3 Non-Intrusive Assessment of Organisational Data Quality

We have developed two special software components that can be added to an existing network of information systems, to provide the capability for non-intrusive assessment of data quality within that system. The assessment method is based on the repeated execution of queries that can detect the presence of defects, at times which have been specified in advance to correspond to “quieter” periods for the individual systems. Any results found by execution of these queries are timestamped and recorded in a dedicated database, for later analysis. The two component types are:

- The Local Quality Assessor component (LQA), which schedules the repeated execution of the query fragments that have been allocated to a specific site within the network, according to the workload patterns specified by the owner of that site. This component also handles transfer of intermediate and final result sets to other sites, as necessary.
- The Global Quality Assessor component (GQA), which allocates the work involved in the defect detection queries to the local sites, according to the information provided by the cost model. The GQA also manages the repository in which details of any defects found within the network of systems are recorded.

In general, there will be one GQA per network of information systems, and one LQA for each site within that network that is expected to be involved in data quality assessment activities. The resulting architecture is illustrated in Figure 1.<sup>1</sup>

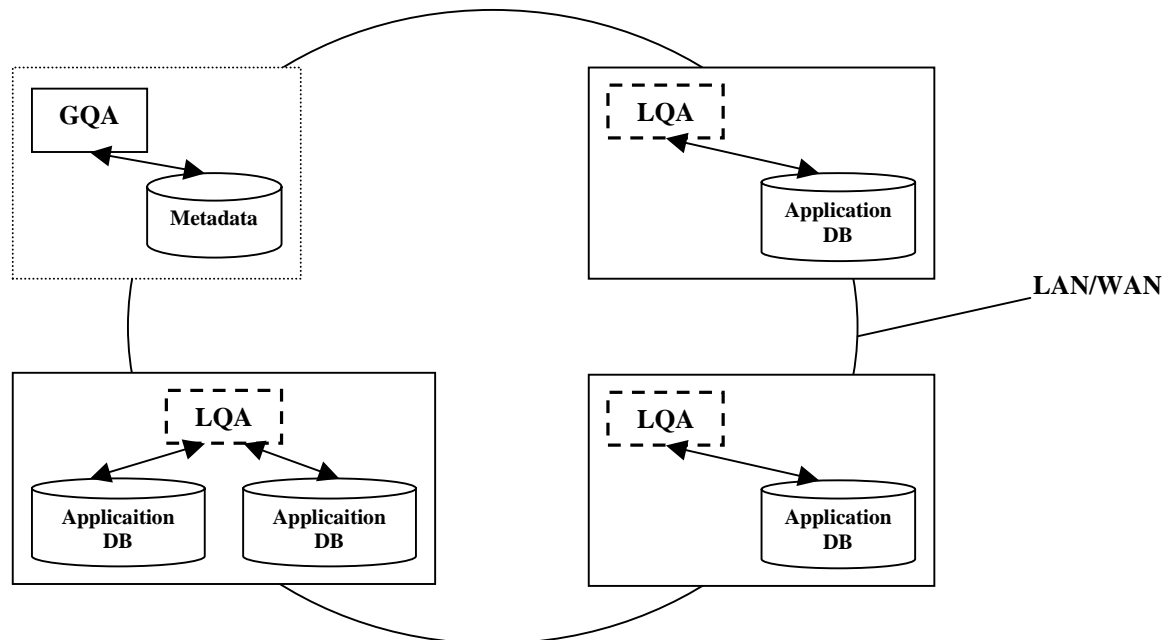


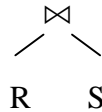
Figure 1. The System Architecture for Non-Intrusive DQA

<sup>1</sup> We also assume that all data sources within the network are accessible via the same SQL interface. Where individual sources do not provide this capability, it is assumed that appropriate wrappers will be provided, which will mimic a relational query interface for use by the LQAs.

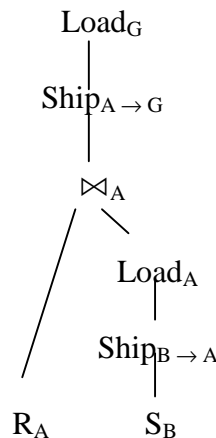
### 3.1 Allocation of Defect Detection Activity

Defect detection queries are specified as relational algebra expressions over a global schema that integrates the schemas of all the local systems. The relationship between the global schema and each local schema is defined by a series of mapping rules, and queries expressed against the global schema are transformed into equivalent queries over the local schema by “unfolding” the mapping rules within the query, as is common in distributed query processing [Ozsu1991]. This task is performed by the GQA, which also stores details of the data detection queries currently being monitored, the global and local schemas, the mapping rules and the data access characteristics of each local site. This latter set of information is used by the cost model to estimate how much time individual query fragments will take to execute at each local site, and includes such information as the block size of the local DBMS, relation sizes and selectivity statistics. Its use is discussed further in Section 4.

After transforming the query so that it references only local tables and attributes, the GQA fragments the query into a number of sub-queries, and allocates each sub-query to a particular local site. This produces a set of sub-queries which each access data that is local to one site or that has been shipped to that site from another. In general, there will be several possible allocations that can be made for any given query. For example, consider the following very simple query, which joins two tables  $R$  and  $S$ , where  $R$  is stored at site  $A$  and  $S$  is stored at site  $B$ :



There are three elements to this query (the two local relations and the joined result relation), and theoretically either site ( $A$  or  $B$ ) could be given the task of executing the join operator, giving a total of two possible allocations. However, different allocations will result in different data transfer and load patterns. For example, if  $R$  is much larger than  $S$ , it will generally be more efficient to ship  $S$  to site  $A$  to compute the join, than to ship  $R$  to site  $B$ . The chosen allocation plan is indicated by inserting instructions to ship data, and to load it into local storage for further processing, into the original query expression. For example, we might produce the following allocation plan from the above query:



Notice that additional shipping and loading operations are added to the root of the allocation plan, in order to ensure that the final results of the query are shipped to the GQA, to be timestamped and logged in the database of detected defects. This occurs even if an empty result set is produced, as we need to record when defects have been corrected, as well as when they are introduced.

In traditional distributed query processing (and integrity checking) it is usual to try to choose an allocation that minimises the time spent transferring data between sites, as this is the major delaying factor in the evaluation. As we shall see in Section 4, the situation is not quite so simple for non-intrusive assessment of data quality, and a variety of other factors have to be taken into account. However, certain obviously stupid allocations can be ruled out of consideration straight away, and the GQA uses a couple of simple heuristics to avoid generation of such allocations:

- Always allocate production of a local relation to the site at which that local relation is stored.
- Always allocate a unary operator to the same site as its operand.

In addition to fragmenting each defect detection query, and allocating the sub-queries to the various LQA components for execution, the GQA also provides each LQA with some information about the order in which its query fragments must be executed. Largely, this order is based on the normal dependencies within the query, so that a sub-query which uses the result of some other sub-queries as its operand must be executed after that other sub-query. However, for complex queries, it is possible that two sub-queries may be independent of each other, but still be allocated to the same site. In these cases, the GQA tries to suggest the most efficient order for their execution, based on the cost model (to be described later).

### 3.2 Non-Intrusive Execution of Defect Detection Activity

Once a defect detection query has been allocated to the relevant local sites, the LQA components take over the task of repeatedly executing the sub-queries during times when this processing activity will not affect mission critical processing. The owner of each local information system specifies the expected workload pattern at that site in advance, in the form of a timetable showing predicted busy and idle periods. Figure 2 shows an example workload pattern. The times given in the workload are relative times, and the whole timetable describes a continuously repeating

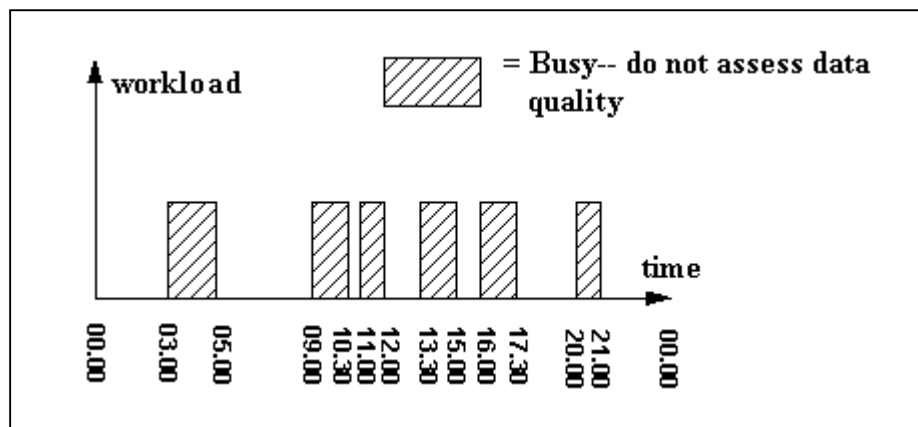
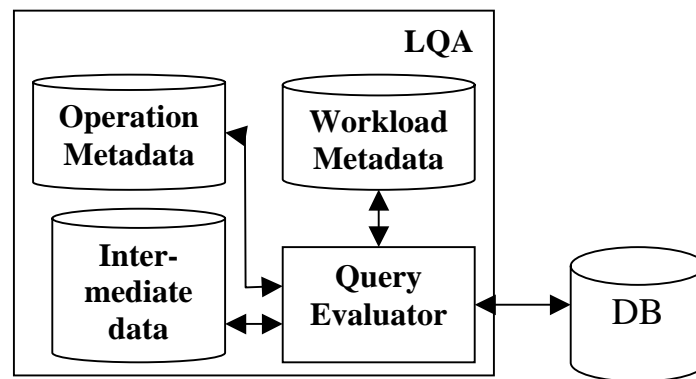


Figure 2. Example Timetable Showing Expected Local Workload Patterns

cycle. The workload pattern in this example would repeat every day, but timetables of arbitrary length may be specified (e.g. to capture the different workload patterns encountered in a typical week or month).

The internal structure of each LQA is shown in Figure 3. This architecture is similar to that used by the LOIS system [Caine2001], which checks integrity constraints in centralised systems in a non-intrusive manner. The LQA also contains a record of the sub-queries it is expected to evaluate (Operation Metadata), and any constraints on the order in which they should be evaluated, set by the GQA. Intermediate data sets, which have been shipped to this site from other LQAs, are also recorded, so that they can be loaded into the DB for further processing when required.



**Figure 3. Local Quality Assessor Architecture**

The query evaluator coordinates the execution of all the sub-queries that are the responsibility of each LQA. This component sleeps while the local information system is busy (as defined by the pre-specified workload) but wakes up when it enters a period of ‘idle’ time (or, alternatively, time when the system owner is prepared to devote some data processing resources to data quality assessment). It then repeatedly chooses a sub-query, data load operation or data ship operation from the Operation Metadata, and executes it, until the period of idle time ends. The exact criteria used by the LQA to select the next sub-query operation for execution is quite complex, and we can give only an overview of them here.

In designing the algorithm for selecting operations for execution, we were motivated by the following considerations:

- Since the time available for data quality assessment is necessarily limited, we wish to avoid wasting processing time when we know that we cannot produce any new information by the chosen action (e.g. by repeatedly executing a sub-query when its result cannot have changed).
- Most updates to data will occur during times when the system is “busy”, but all query evaluation activity for data quality assessment must wait until the system becomes idle again. There will therefore necessarily be a delay between defects entering (or being removed from) the system and their detection by the data quality assessment system. We

would like to minimise this delay as far as possible (in order to maximise the accuracy of the results).

In order to achieve the first of these aims, the LQA keeps track of when each sub-query result was last re-evaluated, when each local data set was last updated<sup>2</sup> and when each intermediate data set sent from another site is refreshed. This information allows us to determine the set of sub-queries that have not been evaluated since one or more of their operands was refreshed or updated. We say that these sub-queries are *eligible for execution*. All other sub-queries need not be considered, since no new information can be gained by evaluating them.

In order to achieve the second of our aims, we must find some way to choose one sub-query from those which are eligible for execution, so that the delay between changes to operands and evaluation of the sub-queries are minimised. We therefore choose to execute that sub-query which has the largest gap between refresh/update of one of its operands and the current time, subject to the ordering constraints imposed by the GQA.

By this means, the distributed defect detection query is executed (sometimes) concurrently by the LQAs, according to their local workload constraints. Each time execution of a query completes, the final result set (which may, of course, be empty) is sent to the GQA, where it is time-stamped and logged for future analysis. Thus, we can achieve “almost” continuous monitoring for defects, at a greatly reduced cost to the organisation than could be achieved with either full integrity checking or ‘one-off’ data cleaning attempts. In the next section, we will discuss whether the non-intrusive approach described here can provide information that is as useful as that provided by these other alternative approaches, and how its shortcomings can be addressed.

#### **4 Addressing the Shortcomings of Non-Intrusive Data Quality Assessment**

We have already discussed how in order to improve data quality within an organisation it is necessary to measure the initial quality levels, then to analyse the resulting measurements in order to decide what action to take to prevent more defective data from being introduced. However, if the data quality is not measured accurately enough, then the results may not reflect the real status of the organisation’s data [English99]. Just as poor quality operational data can result in mistaken decisions being taken by managers, so inaccurate or incomplete measurements of data quality can mean that the actions instituted by data quality staff fail to achieve the desired improvements.

In order to achieve “perfect” knowledge of the data defects that are present in a system, and of when they entered and were removed from the system, it is necessary to monitor every single modification made to data and to compute its effect on the current set of data defects.<sup>3</sup> This is effectively what integrity checking does. Whenever an update occurs which may violate a constraint, a query is issued in order to determine whether that constraint is indeed violated [Embury1995]. We could adapt the same behaviour for data quality assessment, in order to

---

<sup>2</sup> We assume that the local information systems have a facility for informing the LQA when local data is updated. If the local system has a trigger mechanism, then this is easily achieved. If not, it may be possible to discover that data sets have been updated by examination of the transaction log or other metadata.

<sup>3</sup> Of course, our view of “perfection” here is relative to the set of defect detection queries we have formulated. If we formulate a query incorrectly, we will get inaccurate details of the defects that are present. Moreover if we fail to add a query for a particular type of defect that we are interested in, then we will obtain a highly incomplete picture of the prevalence of that kind of defect.

obtain fully accurate and complete knowledge of data defects. Unfortunately, however, the impact on normal business processing would be prohibitive. At the other end of the spectrum, we have one-off data cleansing type activities, in which the effect on normal business processing is extremely limited, but the information obtained is highly incomplete. For example, if data quality levels are only assessed every three months, then many defects may have entered and been removed from the system in the interval between assessments. These defects may be the cause of much lost revenue, but they are not considered by the improvement programme staff, as they are effectively invisible.

The non-intrusive method of defect detection described in the previous section represents a compromise between these two extremes. The cost in terms of data processing resources is much higher than for off-line data cleaning but is considerably less painful than the cost of integrity checking. In addition, because defect checking activities will be performed on (hopefully) a daily basis, the system is much more likely to detect the presence of short-lived but costly defects. The resulting knowledge of defect levels and types is therefore much more complete than under the data cleaning approach.

Unfortunately, we have introduced a new problem by delaying and distributing the defect detection queries: namely, inaccuracy of defect data. The problem is caused by the phenomenon of “phantom states” in distributed query processing [Grefen1997], in which defects may be incorrectly detected or ignored. These are caused by the fact that different parts of the data are checked for defects at different periods in time, and that intermediate results are cached throughout the system (for the purposes of comparison of data from different sites). The delays inherent in distributed query processing mean that these caches quickly become out of date.

For example, consider a defect detection query that finds all examples of “orders” where some product on the order is priced differently to the stated price in the product catalogue. The table *Orders(ProductID, ProductPrice)* is stored at site 1 and the catalogue table *Products(ProductID, ProductPrice)* is stored at site 2. The evaluation of this query can be expressed as three sub-queries:

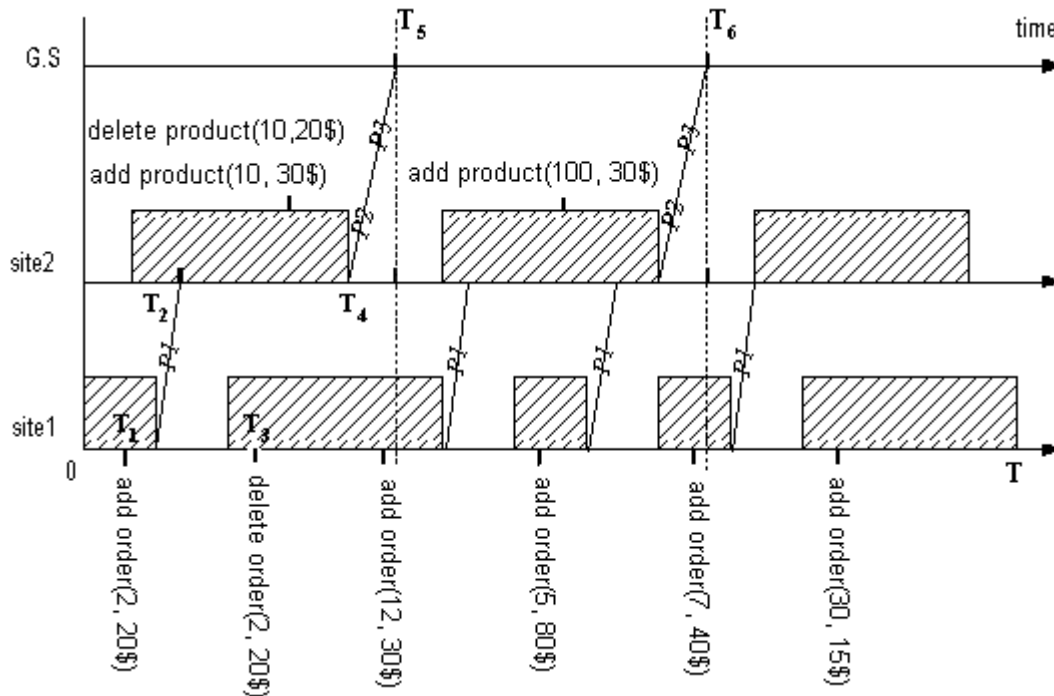
- P<sub>1</sub>: ships the *Orders(ProductID, ProductPrice)* table from site1 from site2.
- P<sub>2</sub>: executes the query:
 

```
select * from Orders, Products where (Orders.ProductID = Products.ProductID)
AND (Orders.ProductPrice != Products.ProductPrice),
```

 at site2.
- P<sub>3</sub>: ships the result of P<sub>2</sub> from site2 to the global site G.S.

The workloads at the two sites in the example system are shown in Figure 4. Assume that a new record is added to the *Orders* table at time T1. This record is in fact defective, but it is deleted from the system at time T3. However, because of the delays inherent in the non-intrusive execution of the query, the defect still appears to be present in the database at time T5. Moreover, we do not discover its disappearance until time T6.





**Figure 4. Example of Non-intrusive Quality Assessment Allocation 1**

An alternative allocation of the work for this same defect detection query is given by the following three sub-queries:

- P<sub>4</sub>: ships the *Products(ProductID, ProductPrice)* table from site2 to site1.
- P<sub>5</sub>: executes the query:
 

```
select * from Orders, Products where (Orders.ProductID = Products.ProductID)
AND (Orders.ProductPrice != Products.ProductPrice), at site1.
```
- P<sub>6</sub>: ships the result of P<sub>5</sub> from site1 to G.S.

The pattern of checking for this allocation is shown in Figure 5. A comparison of these two figures indicates the different patterns of defect detection that can be achieved simply by deriving a different allocation of work. For example, we can see that the defect enters the system at time t<sub>1</sub>, and is detected at time t<sub>2</sub>, thus suggesting that this allocation may result in more accurate results than the previous one. Moreover, we can see that this second allocation results in more frequent reports of defects than the first. We see that defects are reported on four occasions in Figure 5 (indicated by the dotted vertical line), but just two occasions in Figure 4. We could therefore prefer this second allocation to the first.

This second allocation may be more accurate than the first, but it does still have the potential for inaccuracy. Unfortunately, we cannot completely remove this inaccuracy without increasing the amount of data processing resources required for defect checking to unacceptable levels. However, we can attempt to make use of the resources available to us (as described by the local site workloads) in such a way that maximises the accuracy and completeness of the resulting data defect records. We have already seen how choice of allocation of sub-queries to the different sites in the system can affect the amount of time that must be spent transferring data from one site

to another. Similarly, the fact that each site may be operating to a different workload pattern means that different allocations will result in more accurate or more complete defect detection patterns than others. We can therefore maximise the benefits to be gained from our defect detection activities, if we can find some means to choose the allocation that will result in the most accurate and complete defect reports possible.

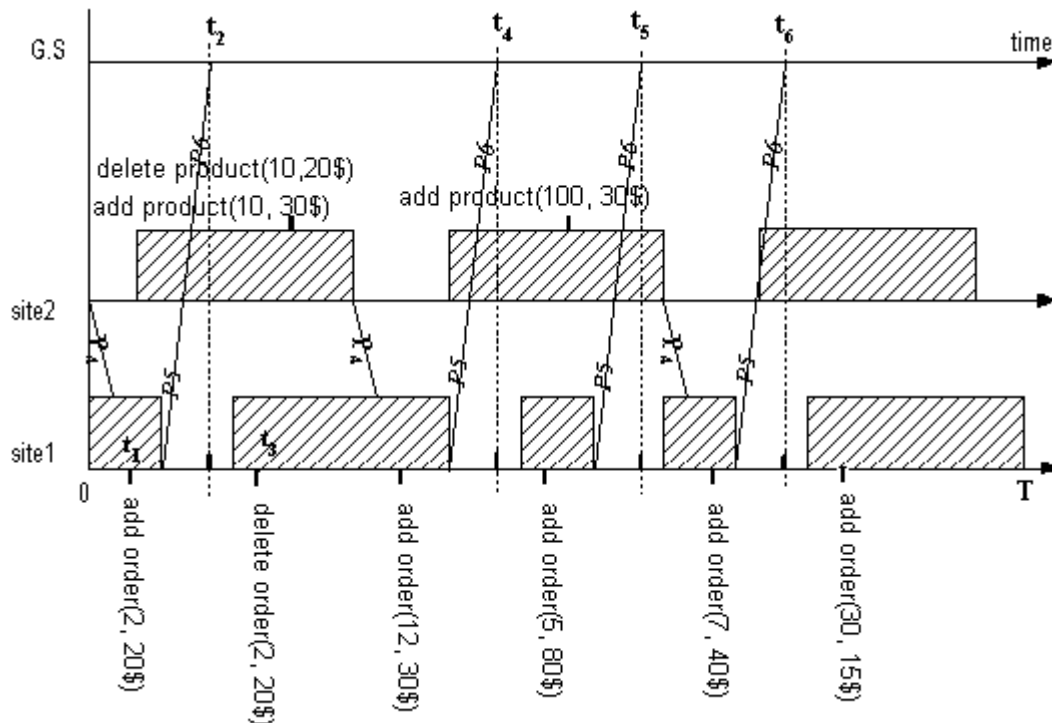


Figure 5. Example of the Non-intrusive Quality Assessment Allocation 2

In order to achieve this, we have developed a cost model which can be used to analyse the characteristics of different allocation plans, so that the most appropriate one can be selected. The cost model is based around a number of parameters. These will now be described before discussing how they can be used to determine the characteristics of a particular allocation.

The cost model simulates the behaviours of each individual allocation plan over a time period according to the workloads specified in advance. In order to reflect the workloads completely, the duration of this simulation period needs to be adequate. In a particular simulation period, if the defect checking query is executed more frequently, then the DQA is likely to be more accurate and to provide a more complete record of the defect data. We define the number of the times that the result of defect checking arrives at the global site during the simulation period as the ‘Frequency’ of DQA in this period. For example, in figure 5 mentioned previously, the ‘Frequency’ equals six during the period 0-T. In addition, for one allocation plan *AP1*, if the time interval between the evaluation of a sub-query and the corresponding update to one of its source relations is shorter than the time interval for another allocation plan *AP2*, then the assessment result of *AP1* is likely to be more accurate and complete than that of *AP2*. We call this characteristic the ‘Delay’. To compute the delay inherent in the complete evaluation of some defect detection query, we sum the time intervals between the update to each of the source relations and the recording of the final query results at the global site. Clearly, due to the

complex interactions of multiple workloads, some evaluations of the defect query will result in a larger 'delay' than others, even with the same allocation plan. In order to gain an impression of the overall level of delay caused by a given allocation plan, the average delay per complete defect check is computed for the entire simulation period. This is done using the following formula:

$$\text{Inaccuracy} = \text{sum}(\text{Delay}) / \text{Frequency}$$

Thus, at compile time, the GQA uses the above cost model to select the most appropriate distribution of defect detection work, and hopefully to make sure that the defect detection queries assess DQ as accurately and as completely as possible given the available resources.

## **5 Conclusions**

We have described an approach to the assessment of data quality across a network of linked information systems. This approach balances the value of the information obtained against the cost to the organisation of achieving this information bearing in mind limitation processing resources. In particular, we allow the owners of each individual system in the network to specify the expected workload on their system, and thus to define the times at which they are willing for data processing resources to be devoted to data quality assessment. Thus, we can achieve regular checking for defects with minimal disruption to crucial revenue-generating data processing activities. While this approach unavoidably introduces elements of inaccuracy and incompleteness into our resulting knowledge of defect levels, we have shown how a more sophisticated cost model can help to minimise these elements, so that maximum benefit can be gained from the small amount of data processing resource given over to DQ assessment.

The next step in our work is to attempt to validate and refine our cost model, through experimentation. We are also working on the development of a range of visualisation tools which can be used to analyse the detected defect data, in order to assist in the discovery of the root causes of defect classes [Baillot2001]. The design of these tools is made non-trivial by the need to cope with the inherent inaccuracies within the defect detection process, and we plan to make use of techniques from temporal logic to extract weaker, but more reliable knowledge from our defect logs. We hope that these tools will be of particular value in allowing defect causal analysis techniques [Card1993] to be adapted for use with data defects, as well as in underpinning the design and monitoring of data quality improvement programmes.

## **References**

- [Baillot2001] A.Baillot. Visualising Data Quality Problems. *MSc thesis* (in preparation), the Department of Computer Science, the University of Manchester, UK, 2001.
- [Bowen1998] P.L. Bowen. *Continuously Improving Data Quality in Persistent Databases*. <http://www.dataquality.com/998bowen.htm>.
- [Caine2001] N.J. Caine and S.M. Embury. *LOIS: the "Lights Out" Integrity Subsystem*, in Proceedings of 18th British National Conference on Databases (BNCOD'01), B.J. Read (ed.), Chilton, UK, July, Springer LNCS V. 2097, pp. 57-74, 2001.

- [Card1993] D.N. Card. Defect-Casual Analysis Drives Down Error Rates. *IEEE Software*, 10(4), pp. 89-100, July 1993.
- [Chawathe1996] S.S. Chawathe, H. Garcia-Molina, and J.Widom. *A Toolkit for Constraint Management in Heterogeneous Information Systems*, in Proceedings of the 12<sup>th</sup> International Conference on Data Engineering, S.Y.W. Su (ed.), New Orleans, Louisiana, IEEE Computer Society, pp. 56-65, 1996.
- [English1999] L.P. English. *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits*. Wiley Computer Publishing, New York, USA, 1999.
- [Embury1995] S.M. Embury and P.M.D. Gray. *Compiling a Declarative High-Level Language for Semantic Integrity Constraints*, in Proceedings of the 6th IFIP TC-2 Working Conference on Data Semantics (DS-6), R. Meersman and L. Mark (eds.), Stone Mountain, Georgia, USA, Chapman and Hall, pp. 188-226, 1995.
- [Galhardas2001] H. Galhardas, D. Florescu, D. Shasha, E. Simon and C. Saita. *Declarative Data Cleaning: Language, Model and Algorithms*, to appear in Proceedings of the 27<sup>th</sup> International Conference on Very Large Data Bases, Roma, Italy, Morgan Kaufmann, 2001.
- [Grefen1997] P.W.J. Grefen and J. Widom. *Protocols for Integrity Constraint Checking in Federated Databases*, *Distributed and Parallel Databases* 5(4), pp. 327-355, January 1997.
- [Haas1997] L.M. Haas, D. Kossmann, E.L. Wimmers, and J. Yang. *Optimizing Queries across Diverse Data Sources*, in Proceeding of the 23<sup>rd</sup> VLDB Conference, M. Jarke, M.J. Carey, K.R. Dittrich, F.H. Lochoysky, P. Loucopoulos, M.A. Jeusfeld (eds.), Athens, Greece, Morgan Kaufmann, pp. 276-285, 1997.
- [Hernandez1998] M. Hernandez and S. Stolfo. *Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem*, *Data Mining and Knowledge Discovery* 2(1), pp. 9-37, January 1998.
- [Huang1999] K.T. Huang, W.L. Yang, and R.Y. Wang. *Quality information and knowledge*. Prentice Hall PTR, New Jersey, USA, 1999.
- [Nicolas1982] Jean Marie Nicolas. *Logic for improving integrity checking in relational databases*, *Acta Informatica*, Volume 18, pp. 227-253, 1982.
- [Oszu1991] M.T. Oszu and P. Valduriez. *Principles of Distributed Database Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1991.
- [Redman1996] T.C. Redman. *Data Quality for the Information Age*. Artech House, Boston, USA, 1996.

# Using Control Matrices to Evaluate Information Production Maps

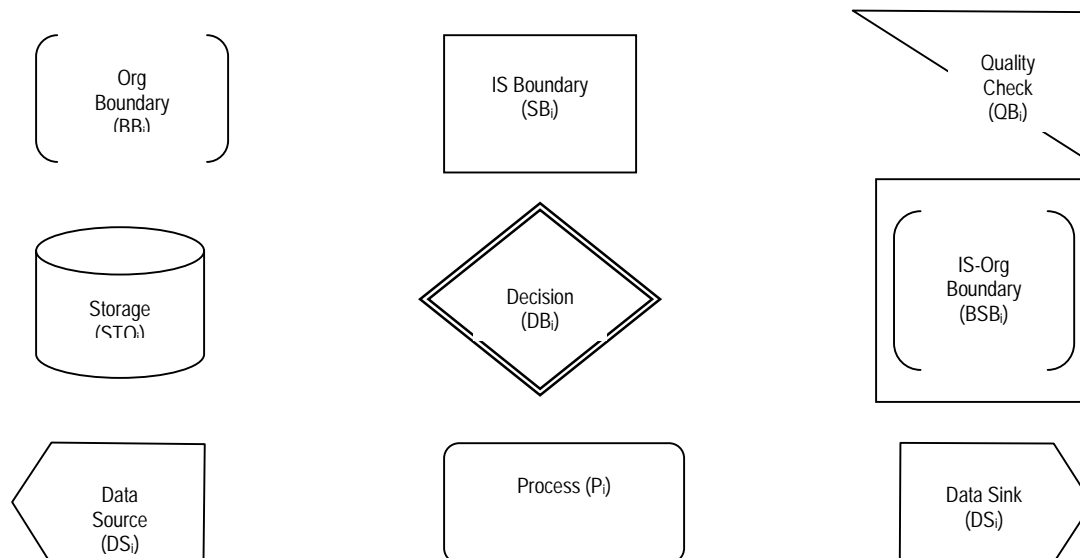
Elizabeth M. Pierce  
 Eberly College of Business  
 Indiana University of Pennsylvania  
 Indiana, PA 15705 USA  
 empierce@grove.iup.edu

**Abstract:** An information production map (IP-MAP) is a graphical model designed to help people to comprehend, evaluate, and describe how a data product such as an invoice, customer order, or prescription is assembled. In this paper, an information product control matrix is introduced in conjunction with the IP-Map to help data administrators assess the overall quality of their information products.

## 1. Information Production Maps

Information Production Maps (IP-Maps) are based on the idea that information outputs produced by information systems are analogous to products produced by manufacturing systems. Just as the quality of products produced in a manufacturing systems can be analyzed and tracked, the same can be done for information products. IP-maps are a means by which one can model the creation of the information product from its original raw data sources through the various processing, quality checking, and storage stages to its final form. Ballou et al. (1998) used this approach to model the information production process using symbolic blocks similar to those used in data flow diagrams. To distinguish this model, which emphasizes terminology and parameters based on manufacturing concepts as well as introducing the data quality construct, Ballou et al. (1998) referred to these diagrams as information manufacturing models. Shankaranarayanan et al. (2000) further enhanced the work done by Ballou et al. (1998) by adding three new blocks to model decision points, IS boundary points, and organization boundary points. Shankaranarayanan et al. (2000) also added metadata describing the department/role, location, business process, data composition, and base system to each of the blocks so that the IP-Map would contain relevant descriptive information that would allow the data administrator to more easily retrieve details about each of the IP-Map constructs.

**Figure 1: List of IP-Map Constructs**



## **2. Information Product Control Matrices**

An information product such as a bill or prescription is prone to a variety of errors such as missing fields, incorrect information, or improper formatting. These errors may come from either the data sources or be introduced through the data process itself. Correction controls in the form of manual inspection or clean up programs may prevent, detect, or correct these errors, but not necessarily with 100% results. Although Information Production Maps give a good overall picture of the process by which data is transformed into the final information product, these diagrams do not highlight data quality parameters that describe the sources of error or the reliability rate of the various process components in a way that easily facilitates the assessment of the overall data quality of the information product.

To aid the data administrator in evaluating the information contained in the IP-Map and the resulting ramifications on the quality of the information product, this paper will borrow on ideas that IT auditors have used since the 1970's. Control matrices are a concise way to link data errors and irregularities to the quality controls that should detect and correct these data problems during the information manufacturing process. The columns of the matrix show sources of data errors or irregularities that occurred during the information manufacturing process. From a data administrator's perspective, the columns of the matrix list the data quality problems that can afflict the information product and identify the location in the IP-MAP model where the data administrator believes these data quality problems originate.

The rows of the matrix are the IP-Map constructs such as the quality checks or corrective processes that were exercised during the information manufacturing process to prevent, detect, or correct these data errors or irregularities. The elements of the matrix are some rating of the effectiveness of the IP-Map construct at reducing the level of data errors or irregularities. These ratings could take several forms:

- *! or X* - A quality check exists to prevent certain error(s) from appearing in the information product. In this case, the data administrator has examined the information production process and has identified that there is a corrective or detective process in place that should prevent that type of error from appearing in the final information product. Notice the ! or X provides the lowest level of assessment information since this only indicates a quality check is present, and does not address how well the quality check performs its function.
- *Category* - A quality check exists in the IP-MAP to prevent certain error(s) from appearing in the final information product and the data administrator is able to describe its effectiveness at error prevention, detection and/or correction as Low, Moderate, or High. This categorical assessment provides more information than a simple ! or X since it captures the data administrator's belief as to how reliably the quality check performs its function.
- *Number* - A quality check exists in the IP-MAP to prevent certain error(s) from appearing in the final information product and the data administrator is able to describe its performance as 95% effective at removing the data irregularity. To obtain a precise numerical assessment of the quality check's effectiveness, the data administrator would need to devise a test in order to evaluate the effectiveness of the IP-MAP control. For example, the data administrator may create several test data sets seeded with known errors. The quality check is then applied to the test data set. If the quality check is able to correct on average 95% of the known errors then the quality check can be considered to be 95% effective at preventing those types of data irregularities from appearing in the information product.
- *Formula* - A quality check exists in the IP-MAP to prevent certain errors(s) from appearing in the final information product; however, its reliability rate depends on a relationship between itself and some other variables. Under this scenario, the data administrator has obtained through process experimentation an understanding of how the reliability of a quality check may fluctuate and is able to describe that fluctuation through a mathematical function. For example, clerks who take orders over the phone may be less reliable in checking the data quality of the orders as the day wears on and this behavior can be modeled using a function incorporating the time of day.

Note that not every IP-MAP construct will be included in the Information Product Control Matrix. Only those quality checks or corrective processes that impact the data quality of the information product are included. While the IP-MAP is designed to model the overall process and is an important first step in understanding how the various components of the information production fit together, the IP-Control Matrix is designed to focus only on those

parts of the IP-MAP that prevent, detect, or correct data irregularities in the final information product. It is also important to note that not every quality check will detect every type of error. It is quite possible that multiple quality checks are employed during the information manufacturing process, each one designed to catch different types of data problems.

Once the information production control matrix is complete, the data administrator examines each data error columns of the matrix in order to weigh up the effect of the various data quality controls and to determine whether the quality of the information product is at an acceptable level. Acceptable levels of quality will depend on the organization's commitment to data quality as well as to the cost and benefits of maintaining the information product at a given quality level.

In addition, cost and frequency information can be added to the Information Product Control matrix to estimate the dollar impact of an unreliable information product. To help illustrate the assessment process, Table 1 shows a sample information product control matrix. For simplicity, it is assumed that each transaction corresponds to a sales order (information product).

**Table 1: Sample Information Product Control Matrix**

	Information Product					
	Source of Data Errors or Irregularities That Occurred					
	Duplicate data in Component Data produced during Process 1	Data became obsolete during Storage	Typos from Data Source 1	Missing data from Data Source 1	Bad data from Data Source 1	Wrong format used in Process 2 to produce Component Data
Estimated Frequency of Error	2% of transactions	3% of transactions per month	5% of transactions	10% of transactions	6% of transactions	4% of transactions
Estimated Cost of Data Error per Information Product	\$1	\$1	\$2	\$5	\$3	\$4
Reliability Ratings of IP-Map constructs						
Quality Check 1	98% of transactions					
Corrective Process 1		90% of transactions per month				
Quality Check 3			85% of transactions	95% of transactions	88% of transactions	
Quality Check 4						97% of transactions
Overall Quality = Error Rate x (1 - Reliability Rate of IP-Map Construct)	.04% of transactions lead to IP's that are duplicates	.3% of data in storage lead to IP's that contain obsolete data	.75% of transactions lead to IP's that have typos	.5% of transactions lead to IP's that have missing data	.72% of transactions lead to IP's that have wrong data.	.12% of transactions lead to IP's that have the wrong format

For each column (i.e. data irregularity), the data administrator assesses how many errors would still show up in the information product after the quality checks have been performed. His assessment depends on his understanding of the information production process as described in the IP-MAP. To get an overall assessment of the quality of the information product, the data administrator must then combine the individual data irregularities rates into an overall rating. In the simplest case where the error rates are independent and in the same units, the data administrator can apply basic probability rules to determine that the probability that a given information product is free of any defects:

$\prod(1 - \text{Error Rate}_i) = (1 - .04\%)(1 - .3\%)(1 - .75\%)(1 - .5\%)(1 - .72\%)(1 - .12\%) = 97.6\%$  of IP's are correct.

Expected Cost Per 1,000 IP's =  $\sum \text{Cost}_i * 1000 * \text{Error Rate}_i = \$69.80$  per 1,000 IP's.

In the case where numerical assessments of reliability were not obtained, the data administrator would need to subjectively combine the categorical or ! / X ratings to get an overall feel of the quality level of the information products. In addition, if the error rates are expressed in terms of functions, a spreadsheet or simulation program may be needed to gauge the overall reliability of the information product. In particular, where probability density functions are involved, the overall quality of the information product may be better expressed as a graph rather than a single number such as an average.

### **3. An Illustration of the Steps Used to Construct IP-Map Control Matrix**

This section describes the steps a data administrator would follow in order to assess the reliability of an information product using an IP-Map and Information Product Control Matrix. For this simple example, a fictitious scenario is set up based on the author's experience in studying the operations of alumni affairs. For the purposes of this fictional case, the information product is a mailing label. A school called Big State University uses these mailing labels to send out publications to its alumni. Incorrect or out-of-date mailing labels are a problem for the university. Undeliverable mail costs the school money in terms of unrecoverable printing and mailing costs, as well as potential lost donation revenue from missing or disgruntled alumni.

After the end of each academic year, data (including current address information) about graduating seniors are taken from the Big State University's active student database and transferred to the Alumni database. Alumni are encouraged to send name/address corrections and changes to Alumni Affairs so that their address information can be kept up to date. Alumni may choose to phone in the information, send email via the Big State University's alumni web site, or send updates via regular mail. The secretary at Alumni Affairs records this information into the Alumni database on a weekly basis.

Unfortunately, only about 1 in 10 alumni remember to inform Big State University of their changes name and address changes. To track down moving alumni, every quarter Alumni Affairs sends a list of its mailing labels to a Change of Address Service, which compares the address of alumni against its master list and identifies those addresses that have changed. This service has demonstrated it is able to detect and correct 98% of the changed addresses and can also identify misspelled addresses and names. In other words, 2% of the changed addresses remain undetected by the Change of Address Service.

Besides the problem of incorrect or out-of-date addresses, there is also the problem of identifying deceased alumni. While relatives may contact Big State University about 20% of the time to stop the mailings to deceased alumni, in many cases, there is no notification. To help it identify deceased alumni, once a year, Big State University sends its active mailing list to an Obituary Service, which compares the list to its master list and can identify about 80% of the alumni who are now dead.

When it is time for Big State University to send out an alumni publication, Alumni Affairs runs a program to create a list of mailing labels, which are then pasted onto the outgoing publication by the University Mail Service. To minimize the number of out-of-date mailing labels, Big State University has embarked on a data quality campaign to better manage the production of its alumni mailing labels.

#### **3.1 Create the IP Map**

Big State University's Alumni Affairs Division begins its data quality campaign by constructing an information production map to describe its production of mailing labels. The IP-MAP is described by both a graphical depiction of the mailing label production process as well as a table of metadata that completes the representation by capturing information about each of the blocks and the data elements included in each flow in the IP-Map model.



Figure 2: IP-Map for Mailing Labels

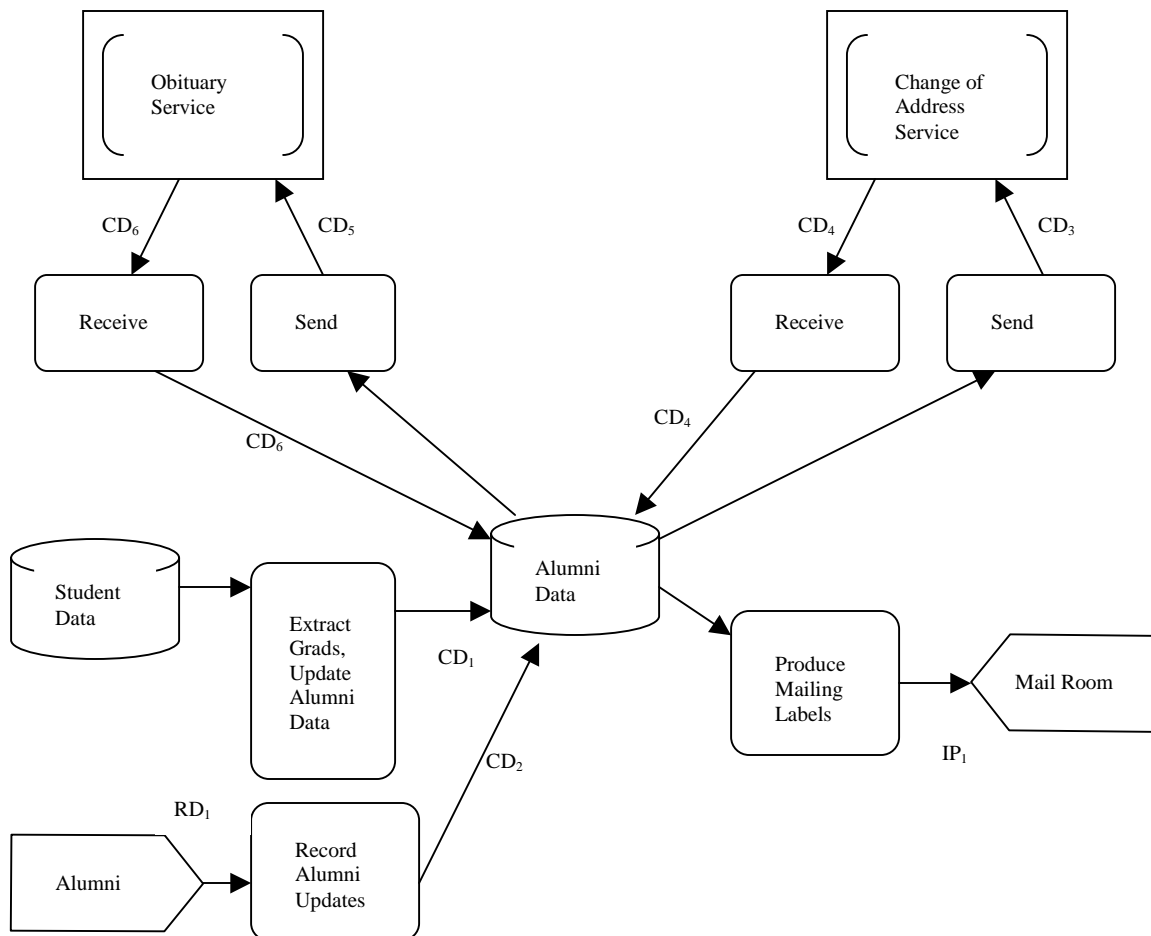


Table 2: Metadata for IP-Map in Figure 1

Name/Type	Dept/Role	Location	Business Process	Composed Of	Base System
Student (STO <sub>1</sub> )	BSU IS Academic Division	222 Carpenter Hall	IS Academic maintains data on active students	Raw data produced by another system.	Oracle Database
Extract Grads, Update Alumni Data (P <sub>1</sub> )	BSU IS Academic Division	222 Carpenter Hall	IS Academic extracts list of graduates each academic year.	Extract from Student Database (STO <sub>1</sub> )	Cobol Program
Alumni (DS <sub>1</sub> )	Alumni	Alumni	Alumni send in their mailing label updates.	Name, Address, and Status Changes	Mail, phone, or electronic correspondence.
Record Alumni Updates and Corrections (P <sub>2</sub> ) (Corrective Process)	BSU Alumni Affairs	202 Breezedale Hall	Secretary records updates.	RD <sub>1</sub> (Raw Data)	Entered via electronic form into Oracle Database.
Alumni Data (STO <sub>2</sub> )	BSU Alumni Affairs	204 Breezedale Hall	Alumni Affairs maintains data on alumni.	Alumni Mailing Information, Other Alumni Data	Oracle Database
Send to Address Service (P <sub>3</sub> )	BSU Alumni Affairs	204 Breezedale, Hall	Alumni Affairs prepares an extract file to send every quarter.	Extract from Alumni Data (STO <sub>2</sub> )	COBOL Extract program, transmit tape.

Address Service (BSB <sub>1</sub> )	NCOA, Inc, BSU Alumni Affairs	NCOA HQ, MD 204 Breezedale Hall	Address Service performs a matching operation.	CD <sub>3</sub> (Component Data)	Vendor System
Receive Updates from Address Service (P <sub>4</sub> ) (Corrective Process)	BSU Alumni Affairs	204 Breezedale Hall	Alumni Affairs receives address updates from NCOA. Applies updates to Alumni database shortly after receiving.	CD <sub>4</sub> (Component Data)	Tape received. COBOL program used to load updates into database.
Send to Obituary Service (P <sub>5</sub> )	BSU Alumni Affairs	204 Breezedale Hall	Alumni Affairs prepares an extract file to send every year.	Extract from Alumni Data (STO <sub>2</sub> )	COBOL Extract program, transmit tape.
Obituary Service (BSB <sub>2</sub> )	Obituary, Inc., BSU Alumni Affairs	Obituary Service HQ, VA 204 Breezedale Hall	Obituary Service performs a matching operation.	CD <sub>5</sub> (Component Data)	Vendor System
Receive Updates from Obituary Service (P <sub>6</sub> ) (Corrective Process)	BSU Alumni Affairs	204 Breezedale Hall	Alumni Affairs receives obituary updates. Applies updates to Alumni database shortly after receiving.	CD <sub>6</sub> (Component Data)	Tape received. COBOL program used to load updates into database.
Produce Mailing Labels (P <sub>7</sub> )	BSU Alumni Affairs	204 Breezedale Hall	Alumni Affairs runs a program to produce a set of mailing labels.	Alumni Data (STO <sub>2</sub> )	COBOL program used to extract data, produce mailing labels.
Mail Room (IP <sub>1</sub> ) (Information Product)	BSU Post Office	111 Sullivan Hall	Mailing labels are used to address alumni publications.	Name, Address, City, State, Zip	Set of Paper Labels

### 3.2 Identify the IP Components and their Potential Problems

Every information product can be viewed as a collection of parts in the same way that a manufactured product such as a car is itself a collection of different parts. These parts are often in turn a collection of smaller parts. For example, a car has many parts (wheels, engine, chassis, seats, etc.). Some of these parts (ex. wheels) can be divided into smaller pieces (ex. tire, hub cap, bolts, etc.). An information product like a mailing list can be viewed as a collection of individual labels. Each label can in turn be broken down into components (name, address, city, state, zip). Some of these components (ex. name) can be further subdivided into smaller pieces (i.e., first name, middle name, last name). Just as different manufactured parts have different problems and failure rates, the different components that make up the information product can have different types of data irregularities and error rates.

Before Big State University's Alumni Affairs Division can estimate the reliability of their mailing labels, they must determine the data components that make up their information product as well as the potential problems that may plague certain components (see Table 3). While some data irregularities affect the information product as a whole such as a duplicate or obsolete mailing labels, some data problems such as an alumni move to a different borough of the city may only affect certain parts of the mailing label. It will be up to the Alumni Affairs Division to decide at what level of detail to track the quality of the mailing label (i.e. mailing label as a whole, main components, or sub-components).

For this example, the Alumni Affairs group uses the standard three-line format used for domestic mail in the U.S. For the sake of simplicity, Alumni Affairs does not send publications overseas or to military address that require a four-line format. The problem that Alumni Affairs is most interested in tracking is out-of-date addresses, names, and deaths. Mistakes in spelling occur relatively infrequently and are of less consequence to the Alumni Affairs group. Note: Since this case is fictional, the error rates cited in Table 3 are meant only to illustrate the Control Matrix technique. In a real-life scenario, the data administrator would need to base the error rates on historical observations, U.S. Census or U.S. Postal data, or experimental study.

**Table 3: Components that make up a mailing label**

Mailing Label		
Components	Sub-Components	Potential Problems
Recipient Line	Name treated as a whole	<p>Typos - Mistakes in the spelling of the name occur infrequently and typically do not interfere with the publication's delivery.</p> <p>Changed (most often occurs in Alumni Data (STO<sub>2</sub>) when female alumni marry or divorce). Alumni Affairs estimates that about 6% of the alumni per year request a name change. Alumni Affairs also estimates that about 60% of the name changes also involve an address change as well.</p>
Delivery Address Line	Address treated as a whole	<p>Alumni Affairs estimates that a small percentage of alumni have mistakes in their address. Typically typos include incorrect or missing direction suffix, wrong street name or number, wrong or missing route number, and wrong or missing apartment number. The Address Change Service will generally catch and correct these mistakes.</p> <p>Changed (most often occurs in Alumni Data (STO<sub>2</sub>) when alumni move). Alumni Affairs estimates that 16% of their alumni change addresses each year.</p>
Post Office, State, Zip Code + 4 Line	i. Post Office (City)	<p>Alumni Affairs estimates that a small percentage of alumni have mistakes in the City, State or Zip fields. The Address Change Service will generally catch and correct these mistakes.</p> <p>Changed (most often occurs in Alumni Data (STO<sub>2</sub>) when alumni move to a new city). Alumni Affairs estimate that 1/2 of the moves involve a change of city.</p>
	ii. State	<p>Changed (most often occurs in Alumni Data (STO<sub>2</sub>) when alumni move out of state). Alumni Affairs estimates that 1/4 of the moves involve a change of state.</p>
	iii. Zip Code + 4	<p>Changed (most often occurs in Alumni Data (STO<sub>2</sub>) when alumni moves to a new zip code region). Alumni Affairs estimates that 3/4 of the moves involve a change in zip code.</p>
Mailing Label as a Whole		<p>Obsolete (occurs in Alumni Data (STO<sub>2</sub>) when alumni dies). Alumni Affairs estimates that 8% of its alumni die each year.</p> <p>Duplicate Labels - This issue is not considered a serious problem by Alumni Affairs.</p>

### 3.3 Construct Information Product Control Matrix

Big State's Alumni Affairs Group is now ready to set up their Information Product Control Matrix. Their focus will be on out-of-date labels for now, but they could expand the matrix to include address mistakes as well. Alumni Affairs has also included frequency and cost information in their matrix along with an assessment of the reliability of current data quality checks in their process.

Information Product Control Matrix						
Label	Recipient Line	Delivery Address Line	City	State	Zip	Post Office, State, Zip+4 Line
Alumni	Name	Address	City	State	Zip	
Dead	Changes	Changes	Changes	Changes	Changes	
.67% of alumni per month	.5% of alumni per month	1.33% of alumni per month	.67% of alumni per month	.33% of alumni per month	1 % of alumni per month	
Avg. Dollar Error Cost per individual mailing	\$3	\$1*	\$1*	\$1*	\$1*	
IP-Map Constructs that control for that data error.						
Record Alumni Updates and Corrections (P <sub>2</sub> ).	20% of dead alumni detected per month	10% of address changes detected per month	10% of city changes detected per month	10% of state changes detected per month	10% of zip changes detected per month	
Receive Updates from Address Service (P <sub>4</sub> )	98% of name changes detected per quarter	98% of address changes detected per quarter	98% of city changes detected per quarter	98% of state changes detected per quarter	98% of zip changes detected per quarter	
Receive Updates from Obituary Service (P <sub>6</sub> )	80% of dead alumni detected per year					
<p>Legend &amp; Notes: Percentages in the body of chart reflect the ability of a quality check or corrective process to detect/correct a particular type of error in the alumni records that it processes.</p> <p>*2/3 of alumni use mail forwarding, 1/3 do not. The \$1 represents the average expected loss of the publication and mailing fees.</p>						

### 3.4 Assess the reliability of the information product

From the information product control matrix, the data administrator can use several methods to assess the overall reliability of the information product depending on the type of data provided. The IT auditing literature cites examples using deterministic models, software reliability models, engineering reliability models, bayesian models, and simulation models to evaluate the use of controls on data integrity. For the data provided in this example, one can use a spreadsheet to analyze the data quality levels over the short term. Given that the both the error rates and corrective processes are time dependent, the results for measuring the reliability of the mailing labels are best displayed graphically over time rather than as a single number. From these results, the data administrator can make a determination based on the needs of the business as to whether or not the data quality levels are acceptable.

#### 3.4.1 Reliability of the Information Product

Using the data supplied by the IP Control Matrix along with a spreadsheet, one can forecast the number of labels with out-of-date information using the formula:

$$\text{Num\_Bad\_Labels}_j = \sum_{i=1}^n (\text{Num\_Bad\_Labels}_{ij-1} + \text{New\_Bad\_Labels}_{ij} - \text{Corrected\_Labels}_{ij})$$

where

$i = 1, 2, 3, \dots, n$  reasons for out-of-date labels,  $j = 1, 2, 3, \dots, m$  periods.

$\text{Num\_Bad\_Labels}_j$  = the overall number of bad labels in period  $j$ .

$\text{Num\_Bad\_Labels}_{ij-1}$  = the number of labels containing a particular type of error  $i$  in period  $j-1$ .

$\text{New\_Bad\_Labels}_{ij}$  = the number of additional labels that are afflicted with a particular type of error in period  $j$ .

$\text{Corrected\_Labels}_{ij}$  = the number of labels containing a particular error  $i$  that were detected and corrected in period  $j$ .

	Time Period											
	1	2	3	4	5	6	7	8	9	10	11	12
<b>Name Changes</b>												
% of alumni who change name each month	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
% of name changes self-reported & corrected each month	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
% of name changes caught by Address Service & corrected per month	0	0	1.35	0	0	1.35	0	0	1.35	0	0	1.35
Cumulative % of labels with undetected name changes each month	0.93 *****	1.38	0.48	0.93	1.38	0.48	0.93	1.38	0.48	0.93	1.38	0.48

\*\*\*\* Period one assumes a starting name error rate among alumni mailing labels of .48%.

	Time Period											
	1	2	3	4	5	6	7	8	9	10	11	12
<b>Address Changes</b>												
% of alumni who change address each month	1.33	1.33	1.33	1.33	1.33	1.33	1.33	1.33	1.33	1.33	1.33	1.33
% of address changes self-reported & corrected each month	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13
% of address changes caught by Address Service & corrected per month	0.00	0.00	3.60	0.00	0.00	3.60	0.00	0.00	3.60	0.00	0.00	3.60
Cumulative % of labels with undetected address changes each month	2.47 ****	3.67	1.27	2.47	3.67	1.27	2.47	3.67	1.27	2.47	3.67	1.27

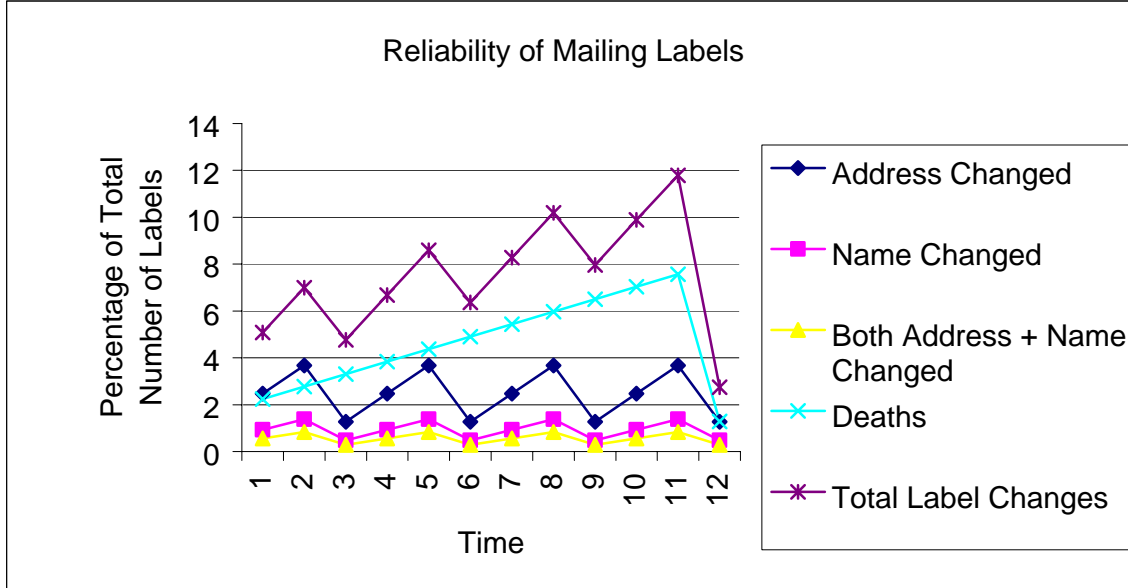
\*\*\*\* Period one assumes a starting address error rate among alumni mailing labels of 1.27%.

	Time Period											
	1	2	3	4	5	6	7	8	9	10	11	12
<b>Alumni Deaths</b>												
% of alumni who die each month	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67
% of alumni deaths reported & corrected each month	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13
% of alumni deaths caught by Obituary Service & corrected per month	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	6.82
Cumulative % of labels addressed to dead alumni each month	2.24 ****	2.77	3.31	3.84	4.37	4.91	5.44	5.97	6.51	7.04	7.57	1.29

\*\*\*\* Period one assumes a starting death error rate among alumni mailing labels of 1.71%.

	Time Period											
	1	2	3	4	5	6	7	8	9	10	11	12
<b>Totals</b>												
Cumulative % of labels with undetected name changes each month	0.93 *****	1.38	0.48	0.93	1.38	0.48	0.93	1.38	0.48	0.93	1.38	0.48
Cumulative % of labels with undetected address changes each month	2.47 ****	3.67	1.27	2.47	3.67	1.27	2.47	3.67	1.27	2.47	3.67	1.27
Adjustment for 60% Dual Name & Address Changes	-0.56	-0.83	-0.29	-0.56	-0.83	-0.29	-0.56	-0.83	-0.29	-0.56	-0.83	-0.29
Cumulative % of labels addressed to dead alumni each month	2.24 ****	2.77	3.31	3.84	4.37	4.91	5.44	5.97	6.51	7.04	7.57	1.29
<b>Total % of labels with undetected errors each month</b>	5.08	6.99	4.77	6.68	8.59	6.36	8.28	10.2	7.96	9.88	11.8	2.75

The graph below shows the mean percentage of mailing labels that have data quality issues for the next 12 periods.



### 3.4.2 Average dollar impact of erroneous information product

Using the cost information from the IP Control Matrix and the formula:

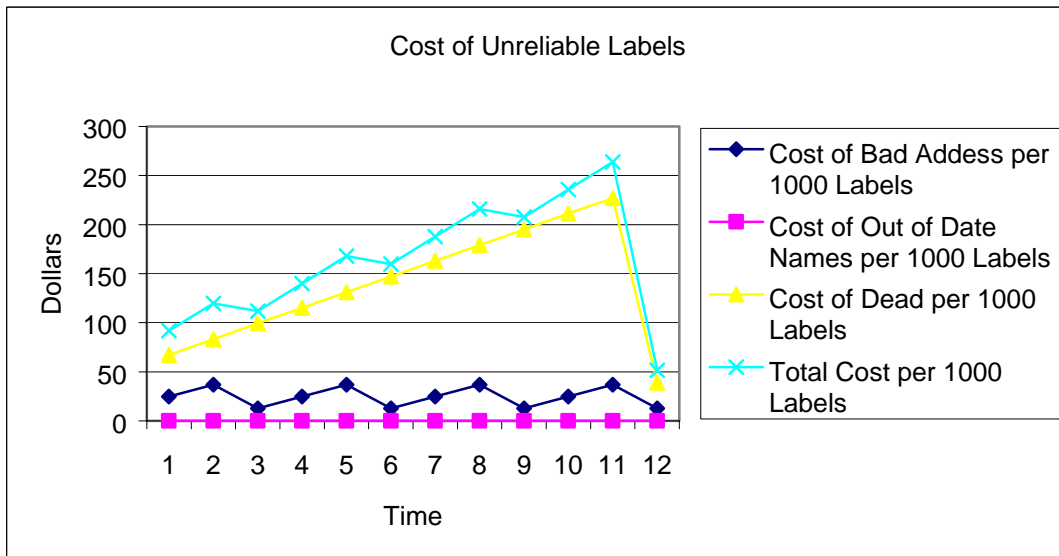
$$Cost_j = \sum_{i=1}^n Cost\_Error_i * Num\_Bad\_Labels_{ij}$$

where

Cost<sub>j</sub> = Overall cost of bad labels mailed out in period j.

Cost\_Error<sub>i</sub> = Cost of a label containing a particular type of error i.

one can display a graph of the average cost of out-of-date mailing labels for the next 12 periods.



### **3.5 Discussion of Results**

Once the data administrator has established the reliability level of the labels and the associated costs of bad labels over time, decisions can be made as to whether the quality is acceptable or if improvements should be made. In this fictional case, the Alumni group at Big State University will have to weigh the costs of additional quality checks or corrective processes against the perceived benefits of increasing the percentage of correct labels. One factor that was not explicitly considered in this case is the timing of the alumni publications. It may be that there is no problem with letting the quality of the mailing list deteriorate provided a clean up is performed prior to the commencement of a mass mailing.

If further data quality improvements are desired, the IP-Map is a useful tool for helping the data administrator to identify potential areas of improvement. Revisions in the information production control matrix and the subsequent what-if analysis using spreadsheets can help project the amount and value of the quality improvements as part of the cost-benefit analysis.

### **4. Summary and Future Research Directions**

The information product control matrix is a tool that can be used to evaluate the reliability of an information product. Essentially, it is an application of control matrices, a tool that IT auditors have long used to help them to make an evaluation decision on how well a system safeguards assets and protects data integrity. The information product control matrix is designed to be used in conjunction with IP-Map in the same way that a traditional control matrix can be used in conjunction with a data flow diagram that shows the controls exercised over the data flows through a system.

An information product control matrix differ from the data manufacturing analysis matrix that Ballou (1998) used to track the data units through the various stages of the data manufacturing process insofar that it focuses just on the corrective processes and data quality checks that most influence the rate at which errors occur in the data elements most pertinent to the final information product. In some respects, the information product control matrix is a specialized version of the data manufacturing analysis matrix, which looks at all the data units and the processes they undergo in their entirety.

Further research and development must be done to refine the information product control matrix to ensure that this tool can adapt to all the complexities that an IP-Map can model. It should also be noted that this technique is only as good as one's understanding and knowledge of the information manufacturing process. The more detailed the measurements, the better the estimates of the reliability of the information product. In addition, the development of software that automates the process of creating, maintaining, and analyzing the IP-Map and the IP Control Matrix is greatly needed to encourage practitioners to adopt these data quality tools in the field.

### **References**

Ballou, D. P., R. Y. Wang, H. Pazer and G.K. Tayi, "Modeling Information Manufacturing Systems to Determine Information Product Quality," *Management Science*, 44(4), 1998, pp. 462-484.

Shankaranarayanan, G., R. Y. Wang, M. Ziad, "IP-MAP: Representing the Manufacture of an Information Product," *Proceedings of the 2000 Conference on Information Quality*, Cabridge, MA.

Weber, Ron, "Information Systems Control and Audit," Prentice Hall, Upper Saddle River, NJ, 1999.



## **Data Quality Issues in Service Provisioning & Billing**

Tamraparni Dasu & Theodore Johnson  
AT&T Labs – Research  
180 Park Avenue, Florham Park, NJ 07932

### **Introduction**

Large corporations maintain a complex array of databases and data warehouses to record data when a customer orders a particular product or services and has recurring interaction with the corporation. The data are used for many core functionalities of the company, such as billing a customer, maintaining the customer's history of interaction with the company (e.g. complaints), billing history and history of purchases and usage. The warehoused data are analyzed and mined for customer segmentation, customer retention and acquisition activities (CRM programs), target marketing, advertising, developing new product and service offerings and many other activities (e.g. revenue assurance) that can save and generate significant revenues for the company. Thus, considerable money is spent on collecting, storing and analyzing the data.

However, the data collected are often riddled with problems and glitches. Bad data can lead to bad decisions, loss of revenues, loss of credibility and damage to customer goodwill. More recently, there is a pressure on corporations to provide a customer with electronic access to the customer's account for self-monitoring. Ideally, a customer would want to add, delete or update their order in real time, online. This requires a company to have an accurate and timely view of every customer's services, products and configurations. It is the pressure to provide such e-services to e-enable the customers that has forced many companies to examine the quality of their customer data and push for automatic flow of data seamlessly across various databases and warehouses.

In this paper, we discuss data quality issues as they arise in the process of recording a customer's data, starting from the point of sale, through provisioning of the service, to the mailing of a bill to a customer for products and services. We cannot discuss specifics due to proprietary reasons, however we illustrate with a hypothetical example. There are many interconnected stages involved such as recording the customer order, provisioning it, tracking customer care issues, computing recurring and non-recurring charges, and finally billing the customer. Typically, each stage is maintained by a different organization, in a different database, with poor communication between different databases. All these factors contribute to data quality issues. Using an illustrative example, we will describe potential problems at each stage and propose possible solutions. Many factors, such as legacy systems and decisions driven by the need for quick "time to market" remedies, make it difficult to solve data quality issues that arise in the service provisioning and billing process, especially in a large corporate setting.

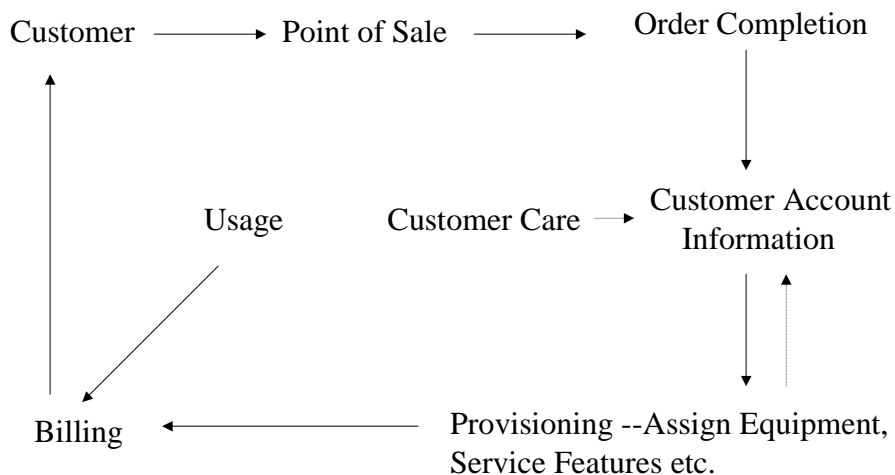
The data quality problems and processes discussed in this paper motivated the development of a database browsing and profiling tool that we discuss in a separate paper [5]. Unfortunately, the paper is too long to include as an appendix to this paper, so we have tried to quote briefly from it or refer to it wherever possible.

Note: We do not consider the process prior to the placing of an order (telemarketing, sales) or after the bill has been rendered (bill collection etc.).

## Description of the Process & Motivation

Consider the case where a customer calls to order a service with certain physical requirements (a T1 line) and service features (data transfer speed). Let us assume that a single provider can meet all the requirements without involving a third party, so that we do not have to consider the data generated by the interaction between the parent company and third party vendors.

### Outline of Sales-Provisioning-Billing Process



**Figure 1**

Figure 1 represents the outline of a sales-provisioning-billing process. We show only the main steps, whereas in reality each stage involves a cluster of activities recorded in numerous databases. The process is set in motion when a customer orders a product or service. The sales representative who handles the call will typically record a minimal amount of information since his/her priority is to “sell, sell, sell”. Furthermore, there are **no standards or requirements** for data entry at this stage. Often, different GUIs are used in different sales branches, each with its own eccentric limitations, so that the amount, manner and type of data collected are dictated by such limitations. A follow-up call or mailing is made to the customer to gather more complete information and create a customer profile in the customer account database. Here again, there are several opportunities for incorrect information being entered – the customer might be unreachable, the person who placed the order might be different from the contact person who provides subsequent feedback and there might not be a strong communication between them, manual misrecording, etc. The customer’s order is then passed on for provisioning of physical, logical and service features (physical cable connection, IP address assignment, modem speed etc.). See Figure 1. Once the provisioning is completed, usage is monitored for recurring charges and the customer is billed regularly. Other databases such as customer care also begin to be populated.

All the databases are inter-dependent and need to communicate smoothly for efficient provisioning, maintenance, care and billing of a customer. As mentioned earlier, large customers would like to access their accounts to manage, update and request changes to their services

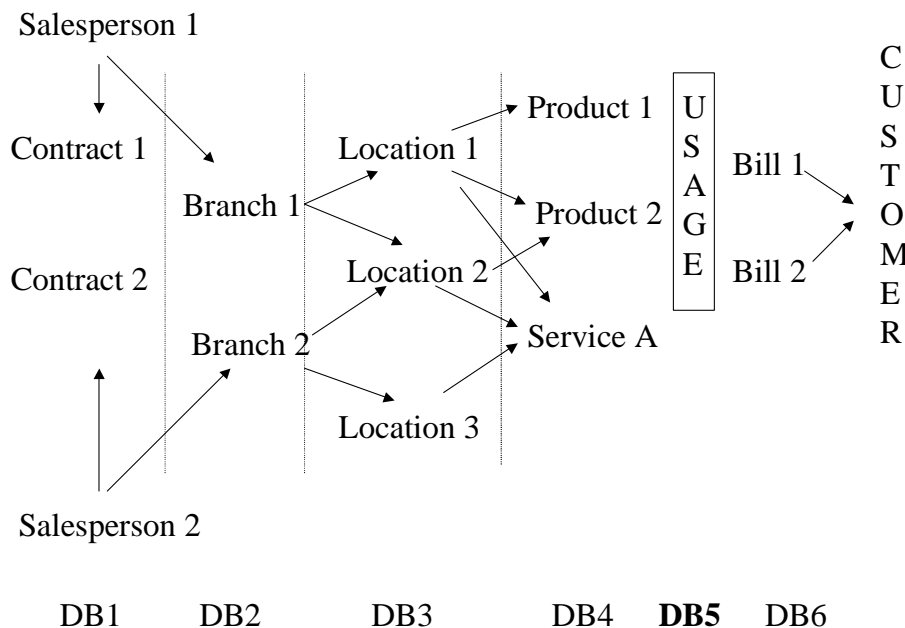
online. In order to enable such services, it is important to maintain a complete, accurate and timely view of the customer's account.

## Data Quality Demons

In the process that we described above, there are many opportunities for data problems to arise.

- Collecting customer data at point of sale. As mentioned earlier, there are opportunities for **manual errors** (180 Park Avenue typed as 108 Park Avenue), **limitations of the interface to the system** (will allow only 10 characters, will allow only 15 fields of information etc.) and **lack of incentive** (priority is to sell, not gather accurate information for subsequent analysis.) Optional data fields are seldom populated resulting in spotty availability of information that could be of value subsequently. It is difficult to address this problem since it is not desirable to tie down valuable selling time with data entry tasks.
- There are **no standards** that are followed across organizations. The Los Angeles branch might require the Industry Code to be populated for the sales person to close the sale, whereas the New York branch might not. Such a situation leads to large amounts of **missing data**.

### Different Views of a Customer



**Figure 2**

- Each organization has a different view of what a customer means. See Figure 2. For example, it could be a contract, a billing entity, a physical resource (a port) or the consumer of a particular product or service. The particular meaning drives the design and relationships within a database. It is **hard to correlate** these views since they often have hierarchies that result in many-to-many relationships that can be confusing. For example, a customer might have many branches scattered geographically, with different billing

hierarchies, volume discount plans etc. In Figure 2, there can be arbitrary dependencies between the various views of a customer maintained in different databases. For instance:

- It might be hard to establish that Bill 1 and Bill 2 belong to the same customer, making the **rendering of a single bill** (a very popular customer demand) difficult. Especially, if the two bills originate in parts of the company that have been recently acquired from outside.
- Suppose the sales force is compensated on the basis of revenues generated from the products and services sold. In the above example, while the mapping between Sales Person (database DB1) and Product (DB4), and usage (DB5) and billing (DB6) might be accurate, the relationships between Product (DB4) and Usage (DB5) are difficult to maintain due to the complexity of the services provided. This makes it difficult to **determine the sales compensation**.
- Since each database is maintained by different organizations, with a high likelihood, there does not exist a common identifier for a customer across databases. Therefore, it is not easy to create a complete history of the customer from sales order to billing. Often, one has to rely on “soft keys” such as names and addresses to merge the databases to get complete information. Name and address matching can be expensive and sometimes inaccurate. For example, one database might list Ted Johnson while another might list Theodore Johnson or T. Johnson. Similarly, one database might list Park Avenue, while the other might list Park Ave. There are commercial vendors that sell software to address such issues.
- The feedback across organizations and databases is not strong so that the databases get out of sync rapidly. For example, the customer database might know that a customer has moved but might not notify the provisioning database so that they can de-allocate the resources assigned to the customer. As a consequence, valuable network resources might remain tied up, creating **spurious capacity saturation problems**.
- When different companies merge or when one acquires another, they might have a large common customer base (e.g. when phone companies merge with cable companies) but nothing other than a name and address (in potentially different formats, conventions and database management systems) as a common identifier between the two companies’ databases. As mentioned earlier, merging databases based on name and address strings is difficult and error-prone.
- Furthermore, a company might inherit databases from an acquired company, but not with inadequate documentation and metadata. In such a case, it might be difficult to determine the key attribute (unique identifier) in any given table of the database. When there are hundreds of tables with thousands of variables, searching manually for a key-like attribute is impossible. (See our related paper [5] about designing an automatic database browser).
- Related to the previous point, it is difficult to find join paths between database tables to extract complete information about a customer. For example, finding the IP address assignment and the usage for the month of August for customer A might require joining several intermediate tables. While many database management software systems automatically generate schemas and diagrams to indicate the join paths, they are often based on field names and not necessarily on an interpretation of the content. Two tables might have completely different keys both of which are called “address”, which might then show up as a potential join path.

## Potential Solutions

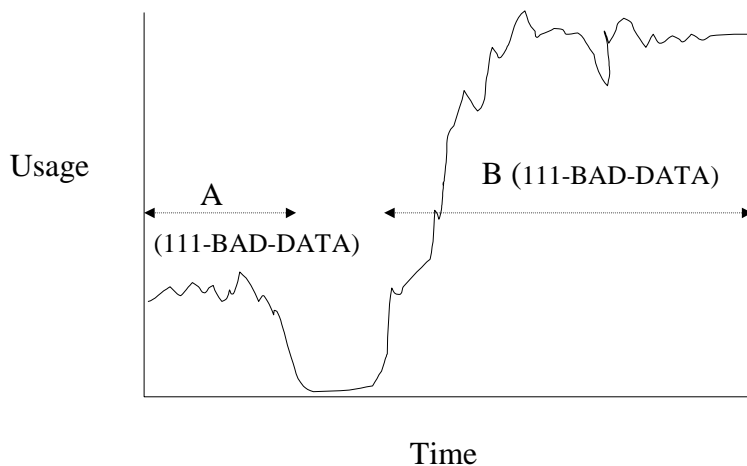
The problem of making information flow from sale to billing (and beyond) is a difficult one, especially in old, legacy systems. There are commercial ventures that enable migration into modern, easy to manage databases. However, the fundamental problem of insufficient documentation and lack of communication between related but independently run databases remains. Often, in a hurry to take a service to market, old technology is re-used in the context of new applications, creating predictable as well as unforeseen problems.

While we cannot address organizational and political issues in this paper, we will outline a few technical solutions that can be employed to facilitate a smoother flow of data.

- Create the infrastructure for knowledge and expertise sharing, as elucidated by Kuan-Tsae Huang, Yang W. Lee, Richard Y. Wang in their chapter on Network Knowledge Infrastructure. See [4] for details.
- As Nigel Turner mentioned in an IQ2000 paper [6], centralize and document metadata and domain expertise. While this is an organizational/process suggestion, the interfaces used by sales personnel and database managers can be setup to require the entry of certain information. This can be enforced through appropriate incentives (compensation tie-ins, other rewards, recognition) as well as easy and **efficient design of interfaces**. The source of many data quality problems can be traced back to a lack of metadata, documentation, interpretation and domain expertise.
- In the absence of good documentation, use heuristics to **automatically generate keys** and potential join keys. For example, **any field that has unique or almost unique values is a potential key**. Similarly, if a combination of two keys results in a unique or near-unique set of identifiers, then we have identified a two-way key. The key finding algorithm computes the counts of unique values of pairs, triples, etc. of the fields of the table. Therefore as a side effect the key finding algorithm finds *approximate dependencies*, in which one set of fields determines another. The key finding algorithms will not find all such dependencies (because it avoids doing all counts), and does not check the quality of the dependence. These dependencies are marked accordingly when they are entered into the profile repository. See our related paper [5] for details about data profiling. The paper is too lengthy to include in the appendix of this paper.
- Once we have rapidly identified keys within each table, we can start identifying "join paths", namely ways of joining different tables to extract comprehensive information about any given entity, e.g. a customer. Note that this task is difficult to perform manually, given that a database has hundreds of tables with dozens of attributes in each table. Without proper documentation, finding join keys can be very hard. Using "similarity measures", we can reduce the set of potential "join keys" to a manageable number. For example, using **string-matching** algorithms (see [3]) we can determine that a 9-digit-key in one table is similar to the two-way key ZIP+4 in another table. In practice, the algorithm has been highly successful in matching databases based on names and addresses. Similarity measures can be used to establish the confidence in the result of merging two tables.
- As a next step, we can **validate the joins** that we have performed using the keys that we have found. We can do the validation by using redundant information that resides in the

two tables. As a simplistic example consider: if we have joined the two tables using name+address, we can use other attributes for **asymmetric validation**. Therefore, if we see that there is revenue in one table, but the service indicator for that particular service is not populated, then there is a reason to believe that the two parts of the merged record do not belong to the same customer. If the information is consistent, then we need to do further checks to validate the match. The asymmetry is due to fact that, if the validation is unsuccessful we have good information that the record has not been properly merged, but if the validation is successful we are not certain of success of the merge. Let us consider a slightly more sophisticated example. Suppose one table contains usage and another table contains billed amount. We can verify an approximate join based on name+address, using **regression validation**. That is, we can use the matches that we are confident about (**labeled examples** to use machine learning terminology) to build a regression type model that estimates billed amount based on usage, and check whether the approximate joins are consistent with this model by comparing the actual billed amounts with the estimated billed amounts derived from the regression model (**supervised learning**). Another form of validation is based on **mutual information**. We use the labeled examples to determine attributes that have high mutual information. Loosely, mutual information is a measure of the extent of association between two attributes. If the mutual information is high, under certain circumstances, one attribute can be used to guess the value of the other attribute. We can then use the guessed value with the actual value to validate the join. See [1] for details on mutual information.

### Potentially Incorrectly Merged Records



**Figure 3**

- We can extend the above concept to **validate time series** joins. For example, in the telecommunications industry, telephone numbers are re-assigned after alarmingly short waiting periods. So when usage tables are joined based on a telephone number (111-

BAD-DATA), it is quite possible to get spurious matches between customer A's history and a new customer B's history. See Figure 3. It is easy to screen out obvious mismatches and check them using more rigorous techniques. (See [2]). The rapid screening of time series to isolate a subset for further investigation is again an asymmetric validation technique. We can use linear models to investigate deviation from expected values.

- While it is not possible to mandate the use of a common identifier across organizations in a company, it might be possible to put in place a mechanism where the mapping between a common identifier and an individual organization is kept current. There is no burden on a particular organization to use a centrally dictated identifier that might not meet the specific needs of the organization. We have found this strategy to be successful in practice.

## **Conclusion**

In this paper, we have focused on the data flow from sales order to billing of a customer. We have documented potential sources of problems and proposed technology based solutions. Other solutions based on political and organizational criteria are not considered in this paper. We refer to [5] for greater detail on the implementation of some of the technical solutions and the resulting tool.

## **Bibliography**

- [1] Thomas Cover & Joy Thomas : Elements of Information Theory, (1991), Wiley Series in Telecommunications. John Wiley & Sons.
- [2] Tamraparni Dasu, Theodore Johnson & Eleftherios Koutsofios : "Hunting Glitches in Massive Time Series Data", IQ 2000, MIT, Boston, MA.
- [3] Luis Gravano, Panagiotis G. Ipeirotis, H. V. Jagadish, Nick Koudas, S. Muthukrishnan and Divesh Srivastava : "Approximate string joins in a database (almost) for free". In Proceedings of the International Conference on Very Large Databases (VLDB), 2001.
- [4] Kuan-Tsae Huang, Yang W. Lee, Richard Y. Wang : Quality Information and Knowledge Management, (1998), Prentice Hall.
- [5] Theodore Johnson & Tamraparni Dasu : "Database Browser" accepted to IQ 2001.
- [6] Nigel Turner & John Hodges : "Deploying Information Quality Tools in a Federated Business" IQ 2000, MIT, Boston, MA.

# INTRODUCING DATA QUALITY IN A COOPERATIVE CONTEXT

Paola Bertolazzi

Istituto di Analisi dei Sistemi ed Informatica  
Consiglio Nazionale delle Ricerche (IASI-CNR)  
Viale Manzoni 30, 00185 Roma, Italy  
bertola@iasi.rm.cnr.it

Monica Scannapieco

Dipartimento di Informatica e Sistemistica  
Università di Roma “La Sapienza”  
Via Salaria 113, 00198 Roma, Italy  
monscan@dis.uniroma1.it

**Abstract:** Cooperative Information Systems are defined as numerous, diverse systems, distributed over large, complex computer and communication networks which work together to request and share information, constraints, and goals. The problem of data quality becomes crucial when huge amounts of data are exchanged and distributed in such an intensive way as in these contexts. In this paper we make some proposals on how to introduce data quality into a cooperative environment. We define some specific quality dimensions, we describe a conceptual data quality model to be used by each cooperative organization when exporting its own data, and we suggest some methodologies for the global management of data quality.

## 1 Introduction

With the explosion of e-Business, Cooperative Information Systems (CIS's) ([18], [3]) are becoming more and more important in all the various relationships among businesses, governments, consumers and citizens (B2B, B2C, C2C, B2G, G2B, etc.). By use of a CIS, autonomous organizations, sharing common objectives, can join forces to overcome the technological and organizational barriers deriving from their different and independent evolutions.

In order to make cooperation possible, each organization has to make available its own data to all other potential collaborators. One possible way is that organizations agree on a common set of data they wish to exchange and make them available as *conceptual schemas* that are understood and can be queried by all cooperating organizations ([14], [12]). Technological problems deriving from the heterogeneity of the underlying systems can be overcome by using component-based technologies (such as OMG Common Object Request Broker Architecture [20], SUN Enterprise JavaBeans Architecture [16] and Microsoft Enterprise .NET Architecture [23]) to realize access to data exported by these schemas.

In a CIS it is imperative to deal with the issue of data quality, both to control the negative consequences of poor cooperative data quality, and to take advantage of cooperating characteristics to improve data quality. In fact, exchanges of poor quality data can cause a huge spread of data deficiencies among all the cooperating organizations. However, CIS's are



characterized by replication of their data across different organizations. This replication can be used as an opportunity for improving data quality, by comparison of the same data at each organization.

The aim of this paper is to give some methodological suggestions to introduce data quality in a cooperative context. In our vision, organizations export not only conceptual models of their data, but also conceptual models of the *quality* of such data, therefore giving rise to many opportunities.

This paper is organized as follows. Section 2, after a brief review of the state of the art concerning data quality, introduces and defines the data quality dimensions we consider most relevant in a cooperative environment. Section 3 proposes a conceptual data quality model that can be exported by cooperative organizations. Section 4 considers a possible tailoring of the TDQM cycle [29] to a cooperative context and in Section 5 we illustrate an application scenario, the e-Government Italian initiative, which provides motivations for our work and the test bed in which we will test our approach. Section 6 concludes the paper with possible future work areas.

## 2 Data Quality

### 2.1 Related Work

The notion of *data quality* has been widely investigated in literature; among the many definitions we cite those of data quality as “fitness for use” [28] and as “the distance between the data views presented by an information system and the same data in the real world” ([21], [26]). The former definition emphasizes the subjective nature of data quality, whereas the latter is an “operational” definition, although defining data quality on the basis of comparisons with the real world is a very difficult task.

We here consider the concept of data quality as defined by a set of *dimensions*, usually considered in data quality literature as quality properties or characteristics of data (e.g. accuracy, completeness, consistency).

Many definitions of data quality dimensions have been proposed, including the identification of four categories (regarding intrinsic, contextual, representation and accessibility data aspects) for data quality dimensions [28], and the taxonomy proposed in [22], in which more than twenty data quality dimensions are classified into three categories, namely conceptual view, values and format. A survey of data quality dimensions is given in [27].

We will inherit some dimensions already proposed in literature, and we will introduce some new quality dimensions, specifically relevant in cooperative environments.

Data quality issues have been addressed in several research areas, e.g. quality management in information systems, data cleaning, data warehousing, integration of heterogeneous databases and web information sources. Based on the analogy between data and manufacturing products an extension of Total Quality Management (TQM) to data is proposed in [29]: Total Data Quality Management (TDQM). Four phases are recognized as necessary for the managing of the Information Product (IP): definition, measurement, analysis and improvement. In this last the Information Manufacturing Analysis Matrix [1] can be used. We here consider the TDQM approach and its extension to CIS.

To our knowledge, many aspects concerning data quality in CIS have not yet been addressed; however, when dealing with data quality issues in cooperative environments, some results already achieved for traditional and web information systems can be “borrowed”. In CIS’s, the main data quality problems are:

- Assessment of the quality of data exported by each organization.
- Methods and techniques for exchanging quality information.
- Improvement of quality.
- Heterogeneity, due to the presence of different organizations, in general with different data semantics.

Results achieved in the data cleaning area ([6], [9], [7]), and the data warehouse area ([25], [9]) can be adopted for the Assessment phase. Heterogeneity has been widely addressed in literature, especially with respect to schema integration issues ([2], [8], [24], [11], [4]).

Quality improvement and methods and techniques for exchanging quality information have been only partially addressed in literature (e.g., [15]). This paper particularly addresses the exchange of quality information by proposing a conceptual model for such exchanges, and makes some suggestions on quality improvement based on the availability of quality information.

## 2.2 Data Quality Dimensions

Two categories for data quality dimensions can be distinguished. *Intrinsic data quality dimensions* characterize properties inherent to data, i.e., which depend on the very nature of data. *Process specific data quality dimensions* describe properties that depend on the cooperative process in which data are exchanged.

### 2.2.1 Data Intrinsic Dimensions

Only the most important dimensions [26] and those we consider most relevant in a cooperative environment are discussed here. These are:

- accuracy,
- completeness,
- currency,
- internal consistency.

Standard literature definitions for these are assumed (e.g. [22]).

### 2.2.2 Process Specific Dimensions

The need for context-dependent data quality dimensions has been recognized [28]. In CIS, the cooperative process provides the context and data quality dimensions are related to data evolution in time and within the process. We have therefore chosen and adapted some of the dimensions proposed in [28] (timeliness and source reliability), and in addition propose a new dimension dependent on the specificity of our context (importance).

Process specific dimensions are tied to specific data exchanges within the process, rather than to the whole process. Hence, in the following definitions, we consider a *data exchange* as a triple `<source organization i, destination organization j, exchange id>`, representing the cooperating organizations involved in the data exchange (i.e. source and destination organizations) and the specific exchange<sup>1</sup>. Moreover, in the following, we will refer to *schema element* meaning, for instance, an entity in a Entity-Relationship schema or a class in an object oriented schema expressed in Unified Modeling Language (UML) [19].

---

<sup>1</sup> The exchange `id` has the role of identifying a specific data exchange between two organizations, as they may be involved in more than one exchange of the same data within the same cooperative process.

### *Timeliness*

⇒ *The availability of data on time, that is within the time constraints specified by the destination organization.*

For instance, we can associate a low timeliness value for the schedule of the lessons in a University if such a schedule becomes available on line after the lessons have already started. To compute this dimension, each organization has to indicate the *due time*, i.e., the latest time before which data have to be received. According to our definition, the timeliness of a value cannot be determined until it is received by the destination organization.

### *Importance*

⇒ *The significance of data for the destination organization.*

Consider organization B, which cannot start an internal process until organization A transfers values of the schema element X; in this case, the importance of X for B is high.

Importance is a complex dimension whose definition can be based on specific indicators measuring:

- the number of instances of a schema element managed by the destination organization with respect to a temporal unit;
- the number of processes internal to the destination organization in which the data are used;
- the ratio between the number of core business processes using the data and the overall number of internal processes using the data.

### *Source Reliability*

⇒ *The credibility of a source organization with respect to provided data. It refers to the pair <source, data>.*

The dependence on <source, data> can be clarified through an example: the source reliability of the Italian Department of Finance concerning Address of citizens is lower than that of the City Councils; whereas for SocialSecurityNumber its source reliability is the highest of all the Italian administrations.

## **3 Data and Data Quality Models**

### **3.1 Data Model**

All organizations involved in a CIS need to export their data according to some specific schemas; these are referred to as *cooperative data schemas*.

These are class schemas defined in accordance with the ODMG Object Model [5]. Specifically they describe types of exchanged data items, wherein types can be:

- classes, whose instances have their own identities;
- literals, when instances have no identities, and are identified by values.

New classes can be defined as collections of objects (as instances are objects) or as structured literals, as a record of literals.

### **3.2 Data Quality Model**

This describes the conceptual data quality model that each cooperating organization must define in order to export the quality of its own data.

A *cooperative data quality schema* is a UML class diagram associated to a cooperative data schema, describing the data quality of each element of the data schema. It can be divided into two types, intrinsic and process specific, described in the following sections.

### 3.2.1 Intrinsic Data Quality Schemas

Intrinsic data quality dimensions can be modeled by considering specific classes and structured literals called here *dimension classes* and *dimension structured literals*.

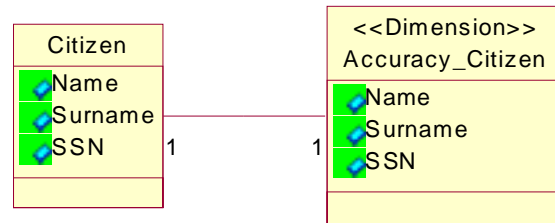


Figure 1. Example of an intrinsic data quality schema.

Each data quality dimension (e.g., completeness or currency) is modeled by a specific class or structured literal. These represent the abstraction of the values of a specific data quality dimension for each of the attributes of the data class or of the data structured literal to which they refer, and to which they have a one-to-one association.

A dimension class (or a dimension structured literal) is represented by a UML class labeled with the stereotype <<Dimension>> (<<Dimension\_SL>>), and the name of the class should be <DimensionName\_ClassName> (<DimensionName\_SLName>).

An *intrinsic data quality schema* is a UML class diagram, the elements of which are: dimension classes, dimension structured literals, the data classes and data structured literals to which they are associated and the one-to-one associations among them.

Consider the class `Citizen`. This may be associated to a dimension class, labeled with the stereotype <<Dimension>>, the name of which is `Accuracy_Citizen`; its attributes correspond to the accuracy of the attributes `Name`, `Surname`, `SSN`, etc. (see Figure 1).

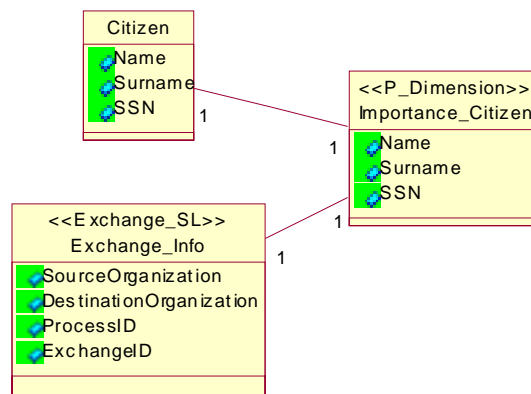


Figure 2. Example of a process specific data quality schema.

### 3.2.2 Process Specific Data Quality Schemas

Tailoring UML in a way similar to that adopted for intrinsic data quality dimension, we introduce *process dimension classes* and *process dimension structured literals*, which represent process specific data quality dimensions, just as dimension classes and dimension structured literals represent intrinsic data quality dimensions.

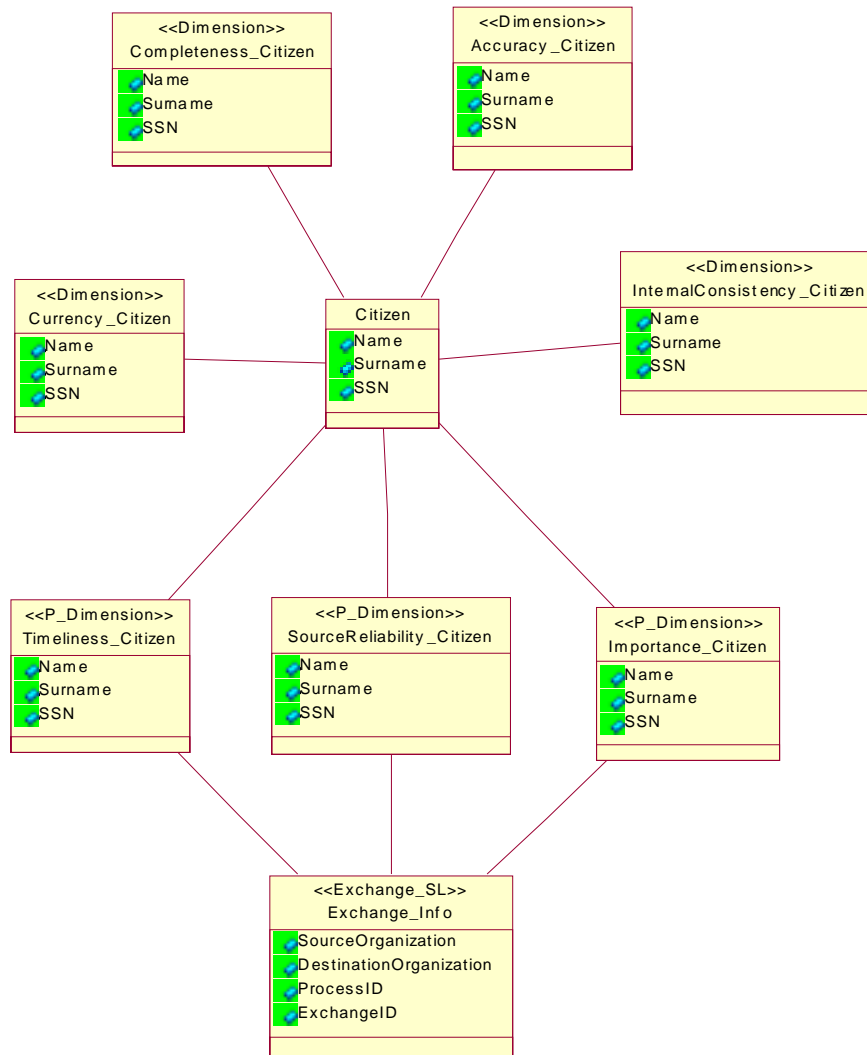


Figure 3. A cooperative data quality schema referring to the Citizen class. All the associations are 1-ary.

Process dimension classes and literals are represented by the UML stereotypes `<<P_Dimension>>` and `<<P_Dimension_SL>>`. The name of the class should be `<P_DimensionName_ClassName>` (`<P_DimensionName_SLName>`).

Also necessary is an *exchange structured literal* to characterize process dimension classes (and structured literals). As described in Section 2.2, process specific data quality dimensions are

tied to a specific exchange within a cooperative process. This kind of dependence is represented by exchange structured literals. They include the following mandatory attributes:

- source organization,
- destination organization,
- process identifier,
- exchange identifier.

Exchange structured literals are modeled as UML classes stereotyped by <<Exchange\_SL>>.

A *process specific data quality schema* is a UML class diagram, the elements of which are: process dimension classes and structured literals, the classes and structured literals to which they are associated, exchange structured literals and the associations among them. Figure 2 gives an example.

The considerations discussed in this section are summarized in Figure 3, in which a cooperative data quality schema describes the quality of both intrinsic and process specific dimensions for the `Citizen` class: the intrinsic data quality dimensions (accuracy, completeness, currency, internal consistency) are labeled with the stereotype <<Dimension>>, whereas the process specific data quality dimensions (timeliness, importance, source reliability) are labeled <<P\_Dimension>>, and are associated to the structured literal `Exchange_Info`, labeled <<Exchange\_SL>>.

## 4 TDQM<sub>CIS</sub>: a Cycle for Quality Treatment in Cooperative Environments

The Total Data Quality Management (TDQM) cycle has been proposed with the aim of providing users with high data quality by considering data as a manufactured product [29]. In this section we show the first steps towards a tailoring of the TDQM cycle to cooperative environments. The TDQM cycle consists of the following phases:

- definition - the identification of data quality dimensions and of the related requirements;
- measurement - producing quality metrics. These provide feedback to data quality management and allow the comparison of the effective quality with pre-defined quality requirements;
- analysis - identifying the roots of quality problems and then studying their relationships;
- improvement - information quality improvement techniques.

These four phases have been redesigned in the context of CIS's, giving rise to the *cooperative TDQM cycle (TDQM<sub>CIS</sub>)*, applicable in the practical cases deriving from the Italian e-Government initiative described in Section 5.

There are five phases to the (*TDQM<sub>CIS</sub>*): Definition, Measurement, Exchange, Analysis and Improvement. They are illustrated in Figure 4. Like the TDQM cycle, TDQM<sub>CIS</sub> is a continuous cycle, in the sense that it must be applied in an iterative way.

### 4.1 TDQM<sub>CIS</sub> Definition

In the TDQM cycle the Information Product (IP) is defined at two levels: its functionalities for information consumers and its basic components, represented by the Entity-Relationship schema.

In the TDQM<sub>CIS</sub> cycle quality data is associated to an IP and specified in terms of intrinsic and process specific dimensions. Both IP's and their associated quality data need to be exported by each cooperative organization through cooperative data and quality schemas, as described in Section 3.

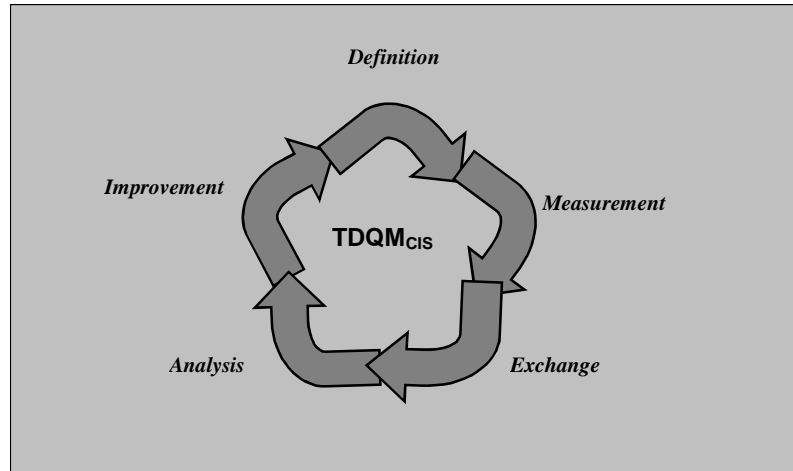


Figure 4. The phases of the TDQM<sub>CIS</sub> cycle.

What organizations have to export is driven by the cooperative requirements of the processes they are involved in. The definition phase therefore also needs to:

- model cooperative processes;
- specify cooperative requirements in terms of what data must be exported and what quality information is needed in each cooperative process.

Note that our focus is a business-to-business context, in which the consumers of exported data are members of the same CIS as the exporter.

## 4.2 TDQM<sub>CIS</sub> Measurement

Two measurement types are made:

- **Static:** source reliability and all intrinsic data quality dimensions are measured statically, i.e. each cooperating organization assesses the quality of its data once using traditional methods (for example the statistical methods proposed in [17]). Data quality values must be computed with respect to the conceptual specification of the defined cooperative data quality schemas. There should also be a general agreement on the metric scales used for data quality dimension measurements.
- **Dynamic:** only timeliness is measured dynamically, i.e. during execution of the cooperative process. To calculate timeliness each organization must indicate the due time, as described in Section 2.2.2.

The importance dimension is not measured at all: values must be specified by each organization, on the basis of how important exchanged data are for the cooperative process. Moreover, importance is used to evaluate data quality measurements of the other dimensions, as it will be clarified in the description of the analysis phase.

## 4.3 TDQM<sub>CIS</sub> Exchange

This phase is additional to the standard TDQM cycle. It is related to the quality of data exchanged among cooperating organizations and includes the exact definition of a transferred unit (TU). Quality data to be transferred include intrinsic dimension values and source reliability values. Importance and timeliness are calculated by the destination organization. With respect to the specified data and quality conceptual models, we distinguish the following types of TU:

- **Type a:** a single attribute value X with its associated quality data, consisting of the values of all the data quality dimensions calculated in the static measurement phase (see Figure 5, which must be completed with the values of the dimensions for X).

Attribute Value	Accuracy Value	Completeness Value	Currency Value	Internal consistency Value	Source reliability Value
X	..	..	..	..	..

Figure 5. Transferred Unit of type a .

- **Type b:** a composite (i.e. multi-attribute) unit with its associated quality. Note that our conceptual model makes a distinction between classes and literals, but the composite unit effectively transferred includes a class instance with all the associated literal instances. Quality data include the values of all the transferred data quality dimensions for each of the attribute values of the composite unit. In Figure 6, the type b TU related to a composite unit including three attribute values (X,Y,Z) is shown.

Attribute Value	Accuracy Value	Completeness Value	Currency Value	Internal consistency Value	Source reliability Value
X	..	..	..	..	..
Y	..	..	..	..	..
Z	..	..	..	..	..

Figure 6. Transferred Unit of type b.

#### 4.4 TDQM<sub>CIS</sub> Analysis

This phase differs from its correspondent in the TDQM cycle, as an analysis step is introduced during the execution of the cooperative process. In particular we distinguish:

- the analysis phase of organization A which sends the TU and
- the analysis phase of organization B, which receives the TU.

A's analysis is similar to the classical analysis phase in the TDQM cycle: the internal processes are analyzed and the causes of poor quality are determined. B's analysis phase is discussed in detail below.

##### 4.4.1 Destination Organization Analysis

This is performed during the execution of a cooperative process. Organization B receives from A a TU including the values of the intrinsic data quality dimensions and the source reliability.

B has three tasks:

- Calculate timeliness as the difference between the due time and the arrival time.
- Interpret the TU's intrinsic quality values. We evaluate dimension values, such as accuracy or completeness, on the basis of importance and source reliability. All intrinsic



data quality values can be weighted with their associated importance and source reliability values, using a weighting function chosen by organization B. For instance, a “low” source reliability for an attribute X of the TU should be weighted with the result of a “high accuracy” for X’s value. The evaluation of timeliness is affected only by importance - source reliability is not relevant. If importance is “high” but the data are not delivered in time, they will be probably discarded by the receiving organization. The interpretation and evaluation phases of TU data quality may also include the calculation of data quality values for the entire TU, starting from the values of dimensions related to each of attributes included in the TU. Though this problem is not in the scope of this paper, we can say that with DIM being a specific dimension, TU a transferred unit, and xi an attribute of TU, the quality value of the dimension DIM for TU is a function F of the value of DIM for each attribute xi of TU, that is:

$$Q_{TU} (DIM) = F_{xi \in TU} (DIM)$$

For each dimension a particular function F can be chosen. We also observe that on the basis of this analysis B can choose to accept or reject the TU.

As an example let us consider the Citizen class with the attribute Name, Surname, SSN, shown in Figure 1. If we consider an “average” function for accuracy and a “boolean-and” function for completeness, we have:

$$Q_{Citizen}(Accuracy) = AVERAGE_{Name, Surname, SSN} (Accuracy)$$

$$Q_{Citizen}(Completeness) = BOOLEANAND_{Name, Surname, SSN} (Completeness)$$

- Send a TU to another organization. B's analysis phase introduces an activity typical of cooperating systems where an organization is at the same time both a consumer and a producer of an IP. In this case B receives a TU X from A, performs a task based on X and then sends a new TU Y to C (see Figure 7).

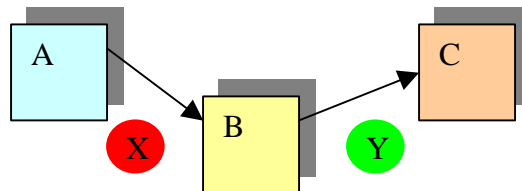


Figure 7. Cooperative exchanges among three different organizations.

Y can be seen as “derived” from X in some way. If the more general case in which both X and Y are type b TU's is considered, the following cases can be distinguished:

1. *B sends X to C without modifying it (Y = X).* If B had specified a due time for B then a value of the timeliness is calculated. All other quality dimension values remain unchanged and are sent to C.
2. *B changes some attribute values and sends Y to C.* Here we consider only one attribute value change; other cases can be easily reduced to this. In this case, let X . ai be the changed attribute. For each intrinsic quality dimension, except consistency, the values previously calculated by B in the measurement phase are replaced.

Consistency must be recalculated as we consider an internal type of consistency<sup>2</sup>. The value for source reliability must be changed from that of A to that of B.

3. *B uses X to produce Y and sends Y to C.* Y is a different TU, so B must calculate the values of all the transferred data quality dimensions. In relation to the possible ways of calculating these attributes, we can distinguish the following cases:
  - If an attribute of Y is obtained by arithmetic operations starting from attributes of X, possible ways of combining the values of the quality for the different dimensions are proposed in [1].
  - If the value of an attribute Y.ai is extracted from a database of B on the basis of the attribute value X.ai then:
    - The accuracy of Y.ai depends on the accuracy of X.ai, with respect to semantic aspects<sup>3</sup>.
    - All other data quality dimension values are known from the measurement phase.

#### 4.5 TDQM<sub>CIS</sub> Improvement

A cooperative environment offers many opportunities for actions that can improve the quality of data shared by cooperating organizations and exchanged in cooperative processes.

The data quality measurement phase enables organizations to understand and address the quality weaknesses of their data on the basis of a comparison with the quality of the same data owned by other organizations. As already observed, the quality of the data held by an organization must be “filtered” according to the source reliability dimension. Source reliability values of cooperating organizations may be set by a *source reliability manager*, which might be one of the CIS members or an external organization. Its main role should be to certify the source reliability of each data exchange within a cooperative process and supply such information to the destination organization on request.

Some other improvements can be made on the basis of the analysis phase performed by receiving organizations. Evaluating and interpreting the quality of delivered TU’s gives the opportunity of sending accurate feedback to the source organizations, which can then implement corrective actions to improve their quality.

Another important opportunity for improvement derives from the dynamic evaluation of timeliness during cooperative process executions. It may be possible to trace the timeliness values for each of the organizations involved in a specific process execution, thus identifying the most critical exchanges with respect to the timeliness of the whole process.

### 5 An e-Government Application Scenario

The approach presented in this paper will be validated in the Italian e-Government initiative [13]. In Italy, in 1993, the Italian Parliament created the Authority for IT in Public Administration (Autorità per l’Informatica nella Pubblica Amministrazione, AIPA) with the aim of promoting

---

<sup>2</sup> Consistency implies that two or more values do not conflict with one other. By referring to internal consistency we mean that all values compared to evaluate consistency are internal to a specified schema element.

<sup>3</sup> Semantic accuracy can be seen as the proximity of a value  $v$  to a value  $v'$  with respect to the semantic domain of  $v'$ ; we can consider the real world as an example of an semantic domain. For example if  $X.ai$  is the key to access to  $Y.ai$ , it may cause access to an instance different from the semantically correct instance: if  $X.ai=Josh$  rather than correctly  $X.ai=John$ ,  $Josh$  can be a valid key in the database of B so compromising the semantic accuracy of  $Y.ai$ .

technological progress, by defining criteria for planning, implementation, management and maintenance of information systems of the Italian Public Administration<sup>4</sup>. Among the various initiatives undertaken by AIPA since its constitution, the Unitary Network project is the most important and challenging.

This project has the goal of implementing a “secure Intranet” capable of interconnecting public administrations. One of the more ambitious objectives of the Unitary Network will be obtained by promoting cooperation at the application level. By defining a common application architecture, the Cooperative Architecture, it will be possible to consider the set of widespread, independent public administration systems as a Unitary Information System of Italian Public Administration (as a whole) in which each member can participate by providing services (e-Services) to other members ([14], [13]). The Unitary Network and the related Cooperative Architecture are an example of CIS. Similar initiatives are currently undertaken in the United Kingdom, where the e-Government Interoperability Framework (e-GIF) sets out the government’s technical policies and standards for achieving interoperability and information systems coherence across the UK public sector. The emphasis of these approaches is on data exchanges, and is therefore focused on document formats (as structural class definitions). The approach presented here aims at introducing a methodology so that organizations can also exchange the quality of their data, and obtain feedback on how data quality can be improved.

## **6 Concluding Remarks and Future Work**

This paper has proposed a possible way to deal with the issue of data quality in a cooperative environment. The importance of introducing specific data quality dimensions was dealt with first. A conceptual modeling language, to represent the quality of the data exported by cooperating organizations was then obtained by the tailoring of the Unified Modeling Language. Finally we discussed how TDQM cycle might be adapted for a cooperative context.

The future directions of our work will principally address a more specific definition of the tailoring of the TDQM cycle, and a validation of our ideas in the context of the Italian e-Government initiatives.

## **7 Acknowledgements**

The authors would like to thank Carlo Batini (AIPA), Barbara Pernici (Politecnico di Milano) and Massimo Mecella (Università di Roma “La Sapienza”) for important discussions about the work presented in this paper.

## **8 References**

- [1] Ballou D. P., Wang R. Y., Pazer H., Tayi G. K., Modeling Information Manufacturing Systems to Determine Information Product Quality, *Management Science*, vol. 44, no. 4, 1998.
- [2] Batini C., Lenzerini M., Navathe S.B.: A Comparative Analysis of Methodologies for Database Schema Integration, *ACM Computing Survey*, vol. 15, no. 4, 1984.
- [3] Brodie M.L.: The Cooperative Computing Initiative. A Contribution to the Middleware and Software Technologies. GTE Laboratories Technical Publication, 1998, available on-line (link checked July, 1st 2001): <http://info.gte.com/pubs/PITAC3.pdf>.

---

<sup>4</sup> See AIPA’s web site, <http://www.aipa.it> for details.

- [4] Calvanese D., De Giacomo G., Lenzerini M., Nardi D., Rosati R.: Information Integration: Conceptual Modeling and Reasoning Support. In Proceedings of the 6th International Conference on Cooperative Information Systems (CoopIS'98), New York City, NY, USA, 1998.
- [5] Cattell R.G.G., Barry D. et alii: The Object Database Standard: ODMG 2.0. Morgan Kaufmann Publishers, 1997.
- [6] Elmagarmid A., Horowitz B., Karabatis G., Umar A.: Issues in Multisystem Integration for Achieving Data Reconciliation and Aspects of Solutions. Bellcore Research Technical Report, 1996.
- [7] Galhardas H., Florescu D., Shasha D., Simon E.: An Extensible Framework for Data Cleaning, in Proceedings of the 16th International Conference on Data Engineering (ICDE 2000), San Diego, California, CA, 2000.
- [8] Gertz M.: Managing Data Quality and Integrity in Federated Databases. In: Second Annual IFIP TC-11 WG 11.5 Working Conference on Integrity and Internal Control in Information Systems, Airlie Center, Warrenton, Virginia, 1998.
- [9] Hernandez M.A., Stolfo S.J. : Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem. Journal of Data Mining and Knowledge Discovery, vol. 1, no. 2, 1998.
- [10] Jeusfeld M.A., Quix C., Jarke M.: Design and Analysis of Quality Information for Data warehouses. In Proceedings of the 17th International Conference on Conceptual Modeling (ER'98) , Singapore, 1998.
- [11] Madnick S. : Metadata Jones and the Tower of Babel: The Challenge of Large –Scale Semantic Heterogeneity. Proceeding of the 3rd IEEE Meta-Data Conference (Meta-Data '99), Bethesda, MA, USA, 1999.
- [12] Mecella M., Batini C.: Cooperation of Heterogeneous Legacy Information Systems: a Methodological Framework. Proceedings of the 4th International Enterprise Distributed Object Computing Conference (EDOC 2000), Makuhari, Japan, 2000.
- [13] Mecella M., Batini C.: Enabling Italian e-Government Through a Cooperative Architecture. In Elmagarmid and McIver 2001.
- [14] Mecella M., Pernici B.: Designing Wrapper Components for e-Services in Integrating Heterogeneous Systems. To appear in VLDB Journal, Special Issue on e-Services, 2001.
- [15] Mihaila G., Raschid L., Vidal M.: Querying Quality of Data Metadata. In Proceedings of the 6th International Conference on Extending Database Technology (EDBT'98), Valencia, Spain, 1998.
- [16] Monson-Haefel R.: Enterprise JavaBeans (2nd Edition). O'Reilly 2000.
- [17] Morey R. C.: 'Estimating and Improving the Quality of Information in the MIS, Communication of the ACM, vol. 25, no. 5, 1982.
- [18] Mylopoulos J., Papazoglou M. (eds.): Cooperative Information Systems. IEEE Expert Intelligent Systems & Their Applications, vol. 12, no. 5, September/October 1997.
- [19] Object Management Group (OMG): OMG Unified Modeling Language Specification. Version 1.3. Object Management Group, Document formal/2000-03-01, Framingham, MA, 2000.
- [20] Object Management Group: The Common Object Request Broker Architecture and Specifications. Revision 2.3. Object Management Group, Document formal/98-12-01, Framingham, MA, 1998.
- [21] Orr K.: Data Quality and Systems Theory. In Communications of the ACM, vol. 4, no. 2, 1998.
- [22] Redman T.C.: Data Quality for the Information Age. Artech House, 1996.
- [23] Trepper C.: E-Commerce Strategies. Microsoft Press, 2000.
- [24] Ulmann J.D.: Information Integration Using Logical Views, in Proceedings of the International Conference on Database Theory (ICDT '97), Greece, 1997.
- [25] Vassiliadis P., Bouzeghoub M., Quix C.: Towards Quality-Oriented Data Warehouse Usage and Evolution. In Proceedings of the International Conference on Advance Information Systems Engineering (CAiSE'99), Heidelberg, Germany, 1999.
- [26] Wand Y., Wang R.Y.: Anchoring Data Quality Dimensions in Ontological Foundations. Communication of the ACM, vol. 39, no. 11, 1996.

- [27] Wang R.Y., Storey V.C., Firth C.P.: A Framework for Analysis of Data Quality Research. IEEE Transaction on Knowledge and Data Engineering, Vol. 7, No. 4, 1995.
- [28] Wang R.Y., Strong D.M.: Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems, vol. 12, no. 4, 1996.
- [29] Wang R.Y.: A Product Perspective on Total Data Quality Management. In Communication of the ACM, vol. 41, no. 2, 1998.

## **Information Quality and Large Scale Project Budget Tracking**

Viktor Dvurechenskikh,  
Chairman, Office of the  
Comptroller of Moscow  
36, New Arbat, Moscow  
121019, Russia  
[nalimova@ksp.mos.ru](mailto:nalimova@ksp.mos.ru)

Vladimir Baranov  
Expert, Office of the  
Comptroller of Moscow  
36, New Arbat, Moscow  
121019, Russia  
[nalimova@ksp.mos.ru](mailto:nalimova@ksp.mos.ru)

George Huntington  
Consultant  
Simsbury, CT, USA  
[gchjr@mail.ru](mailto:gchjr@mail.ru)

**Abstract:** The acceptance of the “E-Russia” project by the government of the Russian Federation presents a significant challenge to existing management structures. Particularly apparent is the need to be able to optimize the financial aspects of a project of this magnitude. Large project budget monitoring and control, including effective warnings regarding impending trends and deviations, is an imposing task. In normal circumstances this is done retrospectively and with minimal predictive capability. By implementing appropriate analytical techniques along with determinations of non-linear instability in existing trends, dramatic improvement in project control can be attained. Additionally, appropriate models exist that enable the availability of public funding sources to be more accurately analyzed and projected. The risk factors in both the expense and funding components are more readily visible and early corrective measures may be applied to mitigate these risks. These techniques offer the potential to more readily assess the relevance and impact of the data normally collected and yields pertinent information that exposes a much more accurate view of the project from a management and financial perspective.

### **EXTENDED ABSTRACT**

The government of the Russian Federation has proposed and accepted the principle objectives of the project “E-Russia.” This project intends to significantly expand the scope and reach of computer communications throughout the federation and foster the development of three key objectives:

- Enhance educational opportunities
- Promote business development
- Provide individuals with access to information

The acceptance of this project has been subject to considerable discussion and controversy. Some of the pertinent criticism has centered on the need to develop improved auditing and management of projects of this size and scope. While information quality now approaches ISACA (Information Systems Audit and Control Association) standards in many areas, the accumulated experience and judgment within government and industry to effectively manage and cope with the implications of a project of this size may be somewhat lacking. This paper is part of a series of steps intended to foster increased awareness of strategic management of information and its quality pertinent to large scale IT projects.

For “E-Russia” as with all large scale projects there are several elements of methodology that are essential to the organization and management of the project. First, the ‘Maturity Model’

establishes the criteria for determination of the successful completion of the key phases of the project. Next, the identification of 'Critical Success Factors' will establish foundational risk avoidance and essential actions required for full realization of the objectives. High quality information is one of the foundational Critical Success Factors. The final component is the identification of key goals and the financial monitoring and control commensurate with efficient attainment of the project objectives.

The monitoring and control of all budgets – from the small family through large State institutions – can be described via similar mathematical models. This is analogous to the ability to describe motion from the microscopic to galactic sizes, via the fundamental equations of mechanics. In actual experience the exactness realized in celestial mechanics is neither practical nor attainable in predictive financial analysis. This paper illustrates fundamental aspects of information requirements, mathematical models and attainable predictive scenarios that may be expected to assist in tracking and controlling the budgets of large-scale projects.

In Russia, project and budget tracking is multifaceted. Much of the effort required is political. Additionally there is considerable technical and procedural work. Finally, there is the analysis of past financial performance and prediction of the future performance. Historically, much of the analysis and predictive work that has been done in the past, in both Russia and the west, has utilized simple linear models (straight line) to ascertain past and future performance.

In principal, the task of accumulating project costs appears to be a simple and straightforward process. Payment data are verified and classified, ancillary charges appended and the data on source documents are transformed from paper to electronic form. It will be necessary to put control barriers in place to ensure an absolute minimum of data entry errors. Typically the resultant information is presented in a tabular form with appropriate totals and the process is regarded as complete. It is only after significant difficulties are experienced that a credible analysis is performed on the departures from expectations and rarely is any predictive effort attempted.

Drawing on the work that has been done in macroeconomics [P.Samuelson and W. Nordhouse, 1995], there are techniques and indicators that are useful for management of projects of this type. A significant aspect of success in this area is the optimization and strategic allocation of resources between the capital costs and the research and development expenses associated with the project: these activities demand data of the highest quality. Formalization and accuracy in this allocation has been demonstrated to be a key component of successful completion of large-scale development projects. In the instances where the roles of significant components is either ignored or merely approximated can result in negative consequences and even the failure of the project as a whole.

A thorough understanding of the econometric work of Samuelson, et al, [P.Samuelson and W. Nordhouse, 1995] and the utilization of standard mathematical techniques, enable a much more accurate evaluation of the present and future status of budgetary performance. This will allow the non-linear aspects of a complex project to be better seen and timely adjustments made before damaging effects get out of control.

By utilizing the econometric model, termed the “Multiplier-Accelerator Model” developed by Nobel laureate Paul Samuelson, [R Shone ,1997], the dynamic nature of project costing can be evaluated. The judicious use of this model, combined with appropriate design parameters of the project itself, will enable careful monitoring of cost performance.

The development of this model involves accurate definition as well as determination of the following components: capital expenditures related to implementation of the project, research and development expenses including continuing expenses to refine auditing and measurement expenses and finally, administrative and management expenses.

The mathematical model proposed in the paper fits the characteristics of the established econometric work of Samuelson, et al. The utility of this model allows several dynamic conditions of project performance to be identified.

First, as development and implementation is progressing in a satisfactory manner the dynamics of project milestones and the total cost of the project is not fluctuating in an unstable manner and subsequent accomplishments can confidently be projected to remain within acceptable ranges. Mathematically this indicates that the model is in a stable state without oscillating costs of requisite tasks.

Second, the model will allow the identification of situations where each step towards completion requires increasing amounts of time, money, information, and resources. This often observed condition is a predictor of runaway costs that can frequently result in either unsatisfactory attainment of project function or project failure due to unsustainable cost of attainment. Early identification of this condition will often enable proper project redefinition and control to eliminate the risks associated with event.

The final output of the model is the identification of the condition where relatively simple circumstances produce wildly fluctuating costs per unit of task completion. This particular instance is often difficult to comprehend, since the general attitude to relatively simple systems is the production of simple, not complex behavior. Yet recent advances in mathematics and econometrics has enables the identification and understanding of this intrinsic ‘chaotic’ behavior. Identification of the limits of the time domain of this condition will allow judicious decisions regarding corrective actions necessary to mediate the risks of this situation.

With proper input and analysis of the functional and budgetary aspects of large-scale projects, early identification of trouble indicators, such as poor quality data over appropriate time frames, project management and costs can be dramatically improved and enhanced.

## **References**

Paul Samuelson and William Nordhaus, *Macro-Economics* (paperback), 15th Edition, (New York: McGraw Hill, 1995).

R Shone *Economic Dynamics* CUP 1997, pp. 94-97).





**IQ-2002**

## ***The 7<sup>th</sup> International Conference on Information Quality***

### **Conference Co-Chairs**

**Yang W. Lee**, Northeastern University  
y.lee@neu.edu, ylee@mit.edu  
(617) 373-5052

**James D. Funk**, S.C. Johnson  
JDFunk@computer.org  
(262) 260-3034

### **Program Co-Chairs**

**Craig Fisher**, Marist College  
Craig.Fisher@Marist.edu

**Bruce Davidson**, Cedars-Sinai Health System  
bruce.davidson@cshs.org

### **Publication Chair**

**Donald P. Ballou**, SUNY Albany  
d.ballou@albany.edu

### **Publicity Chair**

**Leo L. Pipino**, UMASS Lowell  
Leo\_Pipino@uml.edu

### **Finance Chair**

**Richard Y. Wang**, Boston University  
Rwang@bu.edu

### **KEY DEADLINES (FIRM)**

- June 28:** Submission Deadline for papers and panels  
**Aug. 15:** Notification of Acceptance  
**Sep. 16:** Soft Copy and Registration fee due for at least one author  
**Sep. 29:** Registration Deadline for all participants  
**Nov. 8:** IQ-2002 Conference Starts

The 7<sup>th</sup> International Conference on Information Quality (IQ-2002) will be held at MIT from November 8 (Friday evening) to November 10 (Sunday noon), 2002. The purpose of the conference is to promote the exchange of knowledge about IQ research and practice. The conference registration fee for IQ 2002 is US\$295.

### **CALL FOR PAPERS**

The conference program will include tracks of practice-oriented papers, rigorously reviewed research papers, and panel sessions. *IQ-2002 strongly encourages practitioners to submit papers that report experiences, lessons, and perspectives.*

**Research papers** should be initially submitted single-spaced, font 12, Times Roman, and no more than 15 pages in Microsoft Word format (.doc). The abstract must be less than 250 words.

**Practice-oriented papers** can either follow the research paper format or simply include an executive summary of 250 words or less, accompanied by a Power Point presentation.

Each submission must be identified as either a research paper or practice oriented paper. For research papers, please also indicate whether submission is a complete paper or research-in-progress. Submit your paper as an attachment to ylee@mit.edu, Craig.Fisher@Marist.edu, bruce.davidson@cshs.org, and JDFunk@computer.org.

For further information such as program committee, reviewer guidelines, notes on final, accepted papers, directions to the conference site, and conference registration, please refer to <http://web.mit.edu/TDQM>.

### **SUGGESTED TOPICS BUT NOT LIMITED TO:**

IQ Concepts, Tools, Metrics, Measures, Models, and Methodologies  
IQ Improvement Case Studies  
Information Product Implementation, Delivery, and Management  
IQ Education and Curriculum Development  
IQ in the Internet, Web, and e-Business  
Data Warehouses and Data Mining  
IQ Policies and Standards  
Experience Reports on IQ Practices  
Effects of Data Quality  
Explicit and hidden costs of data quality  
Cost/Benefit Analysis of IQ Improvement

The key deadlines are FIRM. Late submissions will not be reviewed or accepted for presentations and publications. Late registration will only be accepted with a late fee, space permitting.