Financial Services and Data Heterogeneity: Does Standardization Work?

Helani N. Galpaya

CISL WP#00-05
September, 2000

Sloan School of Management
Massachusetts Institute of Technology
Cambridge, MA 02142

FINANCIAL SERVICES AND DATA HETEROGENEITY:
DOES STANDARDIZATION WORK?

By
Helani N. Galpaya

Abstract

It is difficult to read a popular newspaper, magazine or web site without finding an article about the dramatic changes that are going on in the Financial Services (FS) industry. Changes in law and regulations and advances in technology are key forces driving the changes in the FS sector. In the race to succeed in the FS sector, FS institutions have to be more proactive, anticipate changing needs of its customers and come up with new offerings. In other words, the speed at which decisions have to be made at FS firms is growing. This demands the ability to analyze data from various internal (data in numerous data sources in the firm) as well as external (data from web sites, outside data feeds) sources.

But many firms face complex problems when trying to aggregate data that is obtained from inside or outside the organization. Often the data that is aggregated from multiple sources contradict each other, is formatted in non-standard ways that makes analysis impossible, or is meaningless to the use because its not what he or she expected to find. This is broadly identified as the problem of data heterogeneity.

Firms that realize the value of data and the ability to aggregate data accurately have taken numerous approaches to attempt to "clean" up their data. Most of these approaches center around the common theme of "standardization". That is, have standards that specify how each data element will be used, how it will be represented, how it will be interpreted, and what it really "means".

But we contend that standardization is often a sub-optimal solution to solving all the problems of data heterogeneity. While standardization solves some types of data heterogeneity problems, many others cannot be corrected by the adoption of standards. Even if they *could* (hypothetically) solve all the heterogeneity problems, the issues inherent in creating standards (complexity, high cost, organizational resistance) should force firms to *at least pause* to consider other alternatives.

This thesis starts out by discussing major technological and regulatory changes taking place in the FS sector and how this is changing the data needs of the FS firm. We will then introduce the concept of data heterogeneity and present a set of examples that illustrate what we mean with regard to (in the context of) FS firm. We will attempt to group the various examples into several key categories – types of data heterogeneity, and why these categories are introduced into a firms data sources. We will examine several examples of standardization attempts between FS firms and examine in length one example of a leading FS firm using a standards-based approach to find a solution to its data heterogeneity problem. We will analyze these approaches, point out their positive facets and suggest some potential shortcomings that may show how alternative approaches could help. One such alternative developed at MIT will be discussed at the end.

Advisor: Stuart E. Madnick
Title: John Norris Maguire Professor of Information Technology & Leaders for Manufacturing Professor of Management Science, MIT Sloan School of Management

Table of Contents

Table of Figures

Acknowledgements

This thesis was made possible with the support of many advisors, colleagues, friends and family.

I would first like to thank my advisor Professor Stuart Madnick for his guidance and support. The amount of focused attention he gave to this work was invaluable, even when he was only a few hours from catching a flight to leave on vacation!.

A special thank you is also due to Dr. Michael Siegel for reviewing my thesis. I appreciate the time he spent reading this work and giving me his comments.

Many chapters of this thesis would never have been possible if not for the generosity of everyone at Merrill Lynch. David Hirschfeld, John Bottega, John Mulholland and Joan Bader were particularly generous with their time and I am indebted to them.

Thanks also to Allen Moultan from my research group for helping me with technical difficulties and setting up the systems speedily so that I could finish my thesis on time.

I am indebted to my parents for their constant support not only during the writing of this thesis but always.

I also need to thank Jean Marie Jordy and Gail Hicky of the Technology and Policy program for making sure all the paper work was in place so that I had a realistic chance of graduating on time!

My academic advisor Dr. Richard Tabors deserves a big thank you as well for all the help he has given me during my time at MIT.

Finally I need to thank many friends in Boston and MIT who supported me and kept me entertained for the past two years and in particular during time I was writing this thesis. Special thanks to Arosha and Rujith for lending me their computers when I poured coffee on mine and burnt it. Finally Saurabh, my office mate and fellow slacker who kept me motivated through late night typing sessions and took care of my degree application - thanks!

1 Background and Motivation

It is difficult to read any popular newspaper or magazine without finding an article about the dramatic changes that are going on in the Financial Services (FS) industry. Changes in law and regulations and advances in technology are key forces driving the changes in the FS sector. Not only are similar types of banks (such as commercial banks) consolidating, merging or get taken over on a daily basis, but banking institutions that were previously separated by law (such as Investments banks and Commercial Banks) are coming together for the first time. In the race to achieve domination in the FS sector, FS institutions have to be more proactive, anticipate changing needs of its customers and come up with new offerings. What this means is that the speed at which decisions have to be made at FS firms is growing. A delay of one day may mean a loss of millions of dollars in the volatile market.

The data collected over the years is one of the most valuable assets at FS institutions. Data accumulated in multitude of data sources are used daily, weekly, monthly and annually to make short and long term strategic and tactical decisions at these firms. Often the data is collected by an organizational entity that is different from the organizational entity that uses it for decision-making purposes. The data is also used for purposes that were not envisioned when it was collected. The data gets used in a very different time frame than when it was collected. In other words, the context the data was gathered (collected) in is often different from the context that the data is used in. Due to the varying standards and norms that are used to represent data, data heterogeneity is introduced into the data elements. This makes the job of aggregating such data across different organizational units very difficult. When data aggregation takes longer than necessary, the nimbleness of the firms is compromised - because top-level management is waiting longer to receive a key comprehensive report, or because the report that is received contains inaccurate figures, without any one realizing it.

One obvious way to get rid of problems of data heterogeneity is to have standards - that is, specify how each data element will be used, how it will be represented and what it means. This is what would occur in a perfect world. But the amount of money, time and other resources needed to standardize millions of data elements often proves to be enormous. By the time many such standardization attempts are completed by a FS firm, the business (and hence data) needs of the firms have changed so much that the clock needs to be started again.
Therefore we will contend that there exists the need to find additional approaches to standardization in order to deal meaningfully with data heterogeneity.

The next chapter will present a brief history of technology and computing as it relates to the FS sector and present some of the key technological and regulatory changes that are changing the face of financial services. Given these changes ongoing in the FS sector, we will then examine how a FS firm's need for good quality data is changing.

Chapter 3 discusses the problems of trying to aggregate data from various data sources. We present examples of aggregation failure. We will introduce the concept of data heterogeneity, present examples of it and propose a way to categorize broadly the different types of heterogeneous data that have been observed. We then look at standardization as a broad approach to solving heterogeneity problems and discuss why this may or may not work. We also analyze the types of heterogeneous data problems that are most likely to be solved by standardization.

Chapter 4 is a discussion of some standardization attempts that have been taken on by the FS industry. We will also present some up-and-coming standards.

Chapter 5 is an observation of a standardization-based attempt to deal with data heterogeneity by a large FS firm. We will review and use existing theories on data standardization and data management to analyze how successful this particular firm's approach might be at achieving its stated goals.

Chapter 6 points out the importance of having alternatives to standardization. We will present briefly a description of one such alternative – the COIN (COntext INterchange) project developed at MIT.

Chapter 7 summarizes the work and provides future directions for further study, investigation and improvement.

2 Financial Services (FS) and the use of data

This chapter first discusses the evolution of computing technology and how it relates to the FS sector. Second it moves into a discussion of how FS firms have historically used data. Third is a discussion of how data usage is changing in today's economy, along with some factors that are driving this change. Finally the present and emerging data needs of FS firms are discussed.

2.1 The evolution of computing and technology

Computing, Information Technology, Information Systems: these all refer to a facet of technology that has had great impacts on firms through out history. On one hand, technology has played a vital part in shaping industry structures and the structure of individual firms. On the other hand, firms (and industries) have taken bold steps that have forced changes in technology.

As Michael Porter points out, a company can win against its competitors by two methods – obtaining a significant cost advantage over its competitors or by differentiating its product from that of its competitors. Technology allows a firm to do both [Porter85].

The literature provides many methods to analyze and to classify the evolution of technology. It is clear that the way individuals and corporations use, interact with, and think about technology has undergone many changes over the years. [Rockart86, pp 377-378] identifies four eras of computing as applicable to a firm:

1. **Accounting Era** – During this time, computer systems were installed primarily to automate accounting functions. Payroll systems, accounts payable, and general ledger systems were instituted on increasingly powerful computer systems throughout the world. Because computer hardware was expensive, systems and personnel to operate them were centralized to obtain higher productivity. A set of tools such as project management tools were developed. COBOL was determined as the key computer language. The data processing managers were "line" managers with all their people reporting directly to them.

2. **Operations Era** – In this era, the emphasis changed from systems serving the accountant to those designed to assist first-line operations personnel. Applications in this era included manufacturing control systems and online order-entry systems. As with the first era, most of the second wave of systems merely enabled companies to do what they had previously done with regard to paperwork processing in a faster and more accurate manner. However the on-line nature of these applications, and the need to access – in most cases – only *local* databases, tended to move the computer systems closer to their users. Therefore the world moved from an era in which "centralized" computing was paramount to one in which "distributed" computing, utilizing both central and local machines, became the norm. Moreover, in the second era, as the applications changed, the process of information systems management also changed. As computers were distributed throughout the company, so were information systems personnel. The Director of Information Systems thus had to become a matrix manager with a 'dotted-line' relationship to distributed information systems personnel and all the problems inherent in such an organization structure. The process of IS management thus become much more complex.

3. **Information Era** – Then came the third era that has been labeled *the information-communication application era.* As opposed to paperwork processing, these applications are concerned with both access to and the use of information and communication within an organization. The previous application era served the lower levels of the corporations. The third era applications serve

middle management, key staff personnel and even the executive suite. The applications, which include decision support systems, executive support systems, end-user computing of all types, electronic mail, and computer conferencing are unique and significant in several ways. These are:

- The procedures (for decision making and other uses) exist primarily in the minds of the managers and staff personnel who utilize them
- It is impossible to justify these applications in traditional cost/benefit ways
- Traditional means of systems development – through well-defined project management techniques – are no long applicable.
- The software languages being utilized are an entirely new breed – often not understood by the majority of information systems personnel
- Significantly higher quality information systems personnel are necessary to work with these applications that were required in the past.
- Demand for access to data and the ability to communicate with those in the corporation is coming from all quarters of the corporation, as opposed to being centered in one or more paperwork-processing clerical groups.

4. **Wired Society** – In this era, there exist multiple levels of connectivity: inter-corporate, intra-corporate, etc. There also is a tendency toward a senior staff-level IT "guru" sometimes known as a Chief Information Office (CIO) serving to bridge the gap between the IT people and the line. Significantly, in this era the line itself tends to take more of a lead in pushing IT developments and implementation of these developments. IT, as an organizational entity, thus reacts to technology-based ideas by the line rather than pro-actively pushing new developments the other way.

2.2 Technology and the FS firm

When we look at the above time line, we see that the FS industry has gone through all four phases identified above. In fact, the FS sector was one of the primary beneficiaries of the first era. The large volume of checks that get cleared and the large amount of accounting activity that is inherent in banks made the use of computers a vital factor in increasing productivity. In fact, if the more than 40 billion checks written annually in the United States were to be processed manually (instead of using automated readers and systems), the effort would require the service of over half of the US workforce.

Over the last thirty years, IT has become increasingly important for banking - cost effectively accommodating increasing transaction volumes, improving timeliness of settlement processing and enhancing the integrity of data computation and storage all depend on IT [SBC98].

During the second "operations" era, we saw financial services firms tying in many systems together to achieve higher levels of operational efficiency. Each business line or business unit in a bank was able to developed customized applications and systems that often tied into one back-office system. The IT personnel were distributed around the organization – some supporting financial products, some supporting divisions etc. – with primary accountability to the business manager, and only secondary (dotted line) reporting to a centralized IT manager or group.

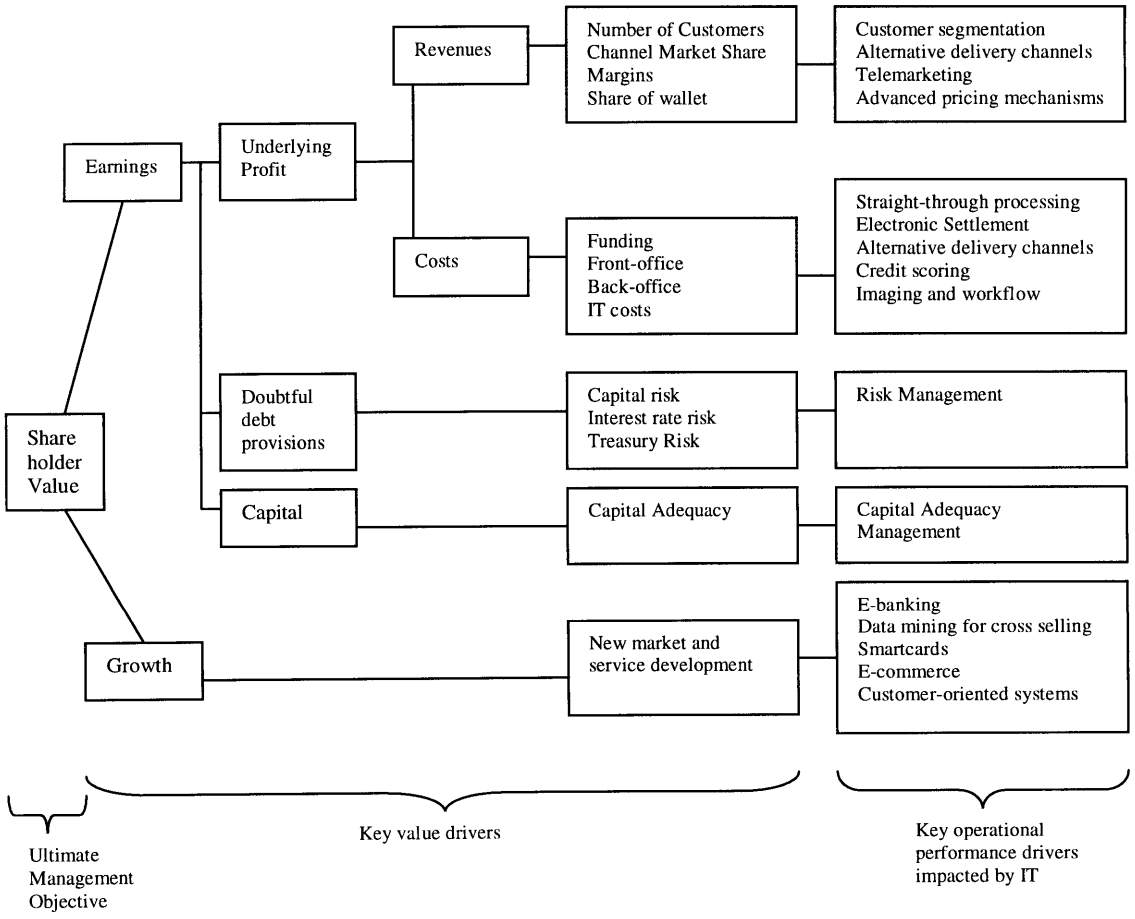
The third era started kicking in during the very early 80's and continued until the use of email, desktop computers, and user access to applications became common place. Due to these changes, the ability to access data from every user's computer was paramount. The users often were executives who used this data to make decisions that would effect the direction of the bank.

Today we are clearly in the era of Rockart's "wired society". Most systems within the bank are interconnected as well, reducing the need for human intervention in most financial communications. Banks have their large and small customers directly connect to their internal systems. More over, access to financial data via hand-held communications devices such as mobile phones is common.

Investments in computing during these times has increased firm level productivity [Brynjolfsson00] in a broad cross section of firms, even when there is not spectacular growth in the economy overall. The FS sector is no exception. Therefore the evolution of and investment technology has effected the bottom line of many FS firms.

Assuming that FS firms, like most other firms exist for the purpose of Increasing shareholder value, Figure 1: Impact of IT in the FS firm, below summarizes the key performance drivers in a bank that are impacted by IT and data.

Figure 1: Impact of IT in the FS firm[1]



[1] Information Technology in Retail Banking, AT Kearney/SBC Warburg Dillion Read, April 1988

2.3 Evolution in the use of data by FS firms

When we follow the evolution of computing, we can see the importance of "data" increasing as we get closer to the present day situation.

2.3.1 Historical uses of data

In the early days (the accounting era), since the attempt was on automating manual processes, the focus was mainly on the efficiency of the system. The processing speed was key. The data that was collected as a result of performing these check-clearances and other operations were of secondary importance. They would be used for netting and settlement purposes and for end of the month/year accounting purposes. The storage of data was often done only for record-keeping purposes (which was often required by financial regulations).

During the operations era, data was still collected and stored mostly for record keeping purposes. But some automated aggregation was being performed on that data by individual business units for the purposes of automating some activities.

During the third (information) era, we finally see the value of data beginning to rise. Senior managers begin to realize that the information lying dormant in the bank's databases can be used for decision-making purposes. Line managers begin to realize that the information (if analyzed properly) will enable them to make better decisions that effect the firms internal as well as external heath. Therefore the emphasis on data increases. Now it is important to have access to current and historical data, to slice/dice it in different ways so as to be able to make decisions. Higher the level of data usage by executives in the organization, higher the level of interconnectivity required between previously disconnected system. For example, previously, a manager evaluating customer credit risk would have referred to one database while the bank's marketing personnel referred to another database. But now a senior manager who wishes to use credit ratings as a viable method of identifying new customers will need to access data from both systems. Therefore investments in technology and data are driven by the need to focus on increasing revenues instead cutting costs (as was the case previously [SBC98]).

2.3.2 Emerging data uses

Today, we are seeing a completely radical view emerging with regard to data in the Financial Services (FS) sector. Gone are the days when data was part of the back-office. Today the FS industry is evolving into a knowledge based industry. The amount of knowledge a firm can gather about its customers has been increasing exponentially due to the Internet - it is easy to reach the customer and it is easy to identify each customer individually, using login profiles, cookies and other methods), trends in the customer's finances. By the same token, customers find it easy to access their financial service provider online, and are demanding access to data immediately. It is reported that over 50 million US households are online currently and about 15 million of these use the web to access some form of financial service provider. By the year 2004, these figures are predicted to grow to 70 million and 32 million respectively [Callinan00]. Therefore FS firms are scrambling to put their financial information (data) online – sometimes not willingly, but often out of necessity because they stand to loose their existing customer base otherwise. Banks have lost out (with a few exceptions) in the sophisticated game of credit cards. Credit card companies use their data to target increasingly narrowly defined segments of the market. Pricing, rewards, and affinity and co-branding programs are now the key success factors for new offerings. But banks are yet to develop the skill of exploiting the customer information that is at their disposal [Bowers95].

But by all indications, FS firms should be in a good position to exploit the new dependence on data. They already have a trusted relationship with their customers, process many forms of financial products for their customers electronically and have access to a large (and often ignored) amount of valuable data about each customer. Selected surveys say that consumers are more likely to buy financial products from financial institutions than from other types of companies (such as technology companies) [MSDW99]. Therefore they are in an ideal position to offer new products and services to their customers because they know each customer's profile, creditworthiness, buying patterns etc. But because the dependence on, and importance of, data has grown almost overnight, most FS firms have been caught off guard. Those who didn't anticipate the emergence of a "knowledge based" economy are scrambling to get access to the vast stores of their own data. But often, before any meaningful analysis can be done, data that is lying in disparate systems have to be collected, cross-referenced and aggregated. This has turned into a nightmare for many firms. The individual systems that were developed (often along operational or business unit lines, without much communication between them) contain large amounts of redundant, useless or corrupted data. Therefore very labor intensive "clean-up" efforts need to be undertaken before the data can be used to make business critical decisions.

2.3.3 Emergence of various data types

Not only has the volume of useful data increased with the emergence of the Internet, but new *types* of data are also emerging. No longer is it just customer name, address and phone number that is collected, but things like "click streams" are turning out to be invaluable information (a "click stream" provides information about a customer's behavior once he or she visits a web site. It is the best source of information about the customers buying patterns and Internet usage habits).
Today the FS firm has access to (or at least the ability to collect) almost all of the following types of trackable data [Kasrel98].

1. Through human contact - mail sent, interaction with agents/brokers/financial consultants
2. Through implicit deductions - Click-stream, demographic profile, page views
3. Through legacy systems - Credit history, Databases, ATM records, Loans and mortgages, Investments
4. Through explicit actions - Survey feedbacks, transaction histories

For example, by collecting the click-stream data for each customer that visits www.citibank.com, Citibank can figure out which products the customer is interested in, which other sites the customer visits and what data queries he or she performs. This information can enable Citibank to target the customer to specific advertising the next time he visits the site. Similarly, using aggregation/screen-scraping/trusted agent technologies, Citibank can present the customer with information about interest rates on home mortgages. If Citibank is savvy, it will not only present the competitor's rates, but also calculate a special rate in real-time and present it to the customer as a special "deal". This is the kind of activity that will increase customer loyalty, increase profits and increase market share. To do all this and more, the ability to analyze and process data from disparate sources in real time is vital.

2.4 The regulatory environment and the use of data

Technology and computing is not the only factor that is changing the use and value of data in the FS sector. Regulatory and legal factors and changes play an equally (and even more) important part in the way data is used.

2.4.1 *Early regulatory structure*[2]

The United States' FS industry has been heavily regulated through history. Starting from the depression era of the 1930's, the amount of regulation increased. The Glass-Steagall Act (Sections 16, 20, 21 and 32 of the 1933 Banking Act) was one of the most prominent among these regulations. This imposed the separation of commercial banking from most forms of investment banking. It permitted affiliates of commercial banks to engage in investment banking as long as they are "not principally engaged" in this activity. The creation of the FDIC (Federal Deposit Insurance Corporation) was another key development that took place in 1933.

In 1956 the Bank Holding Company Act was passed, requiring companies controlling banks to register with the Federal Reserve and establishing standards for non-banking activities that are "closely related to banking". The Douglas Amendment to this (Bank Holding Company) Act effectively prevented bank holding companies from acquiring banks in more than one state.

The above-sited legislation (along with some others) are primarily responsible for the banking and finance structure that evolved in the United States up until the early 1980's. Due to the clear separation of insurance, commercial banking and investment banking activities from collaborating from each other, FS firms were left to pursue clients and sell them products within the legally permitted portfolio of products. No opportunities for cross-selling radically different investment, savings and insurance products were available. Furthermore, geographical restrictions (across interstate boundaries) existed, preventing FS firms from targeting whole groups of consumers.

2.4.2 *First wave of reforms*

Then the 1980's brought along a wave of reforms and deregulation that began to change the FS sector to a certain degree. The Depository Institutions Deregulation and Monetary Control Act eliminated all federal limits on the payment of interest on deposits and permitted interest bearing checking accounts. The 1989 Financial Institutions Reform, Recovery and Enforcement Act generally brought savings associations up to the commercial bank standards for supervisory purposes and increased regulatory enforcement and authority. The 1991 FDIC Improvement Act established risk-based deposit insurance and the 1994 Riegle-Neal Interstate Banking And Branching Efficiency Act provided for full nationwide banking with nationwide branching allowed.

These waves of reforms had the effect of breaking down the geographical barriers as well as reducing the obstacles that existed with regard to cross selling different types of (insured and uninsured) financial products. Suddenly banks were able to change and match their product offerings to their customers' demographics in more ways than before.

2.4.3 *Recent regulatory changes*

The most radical regulatory change so far has been the 1998 overhauling of the Glass-Steagall Act (known as the Financial Reform Act). This has for the first time lifted the operational limitations placed on insurance, commercial and investment banking firms - they are allowed to merge and create partnerships (with certain restrictions and subject to approval of course). The merger between Citibank and Travelers Insurance Group to form Citigroup was perhaps the first large-scale consolidation that took place in the aftermath of the repealing of Glass-Steagall. The merger of Bank Boston (a consumer bank) with Robertson Stevens (an investment management firm) is another example of the type of consolidation that has taken place. There are only two examples - many other

[2] [Wilson95]

mergers, takeovers and joint ventures are taking place between diverse FS firms due to this regulatory change.

These recent regulatory changes have had 3 effects on FS firms and the way they use and view their data and systems.

First, the numerous mergers that were mentioned above is making numerous systems integration necessary. Now that the law allows various types of mergers, one of the main justifications for going ahead with a merger is the ability to achieve operational efficiencies. The previously independent banks need to merge all their banking processes in order to see the cost reductions that are gained through economies of scale. Systems integration is needed in order to do this. But physical connectivity between systems is only the first step – the real change in data usage occurs when disparate systems need to be logically connected.

The second effect of the recent regulatory changes an increase in competition. Segments of markets that were previously left to only one type of FS firm are now being attacked by other types of FS firms. For example, companies that were previously investment banks are now beginning to offer checking and savings accounts, thus encroaching into the arena of traditional consumer banks. Moreover, non-FS firms are entering the field - Microsoft and other technology based firms are beginning to offer financial management products. Therefore FS firms have to offer better products and services in order to retain their customer base and attract new ones. In this process, this, information technology serves as the key tool. With the proper information technology tools and the right kind of data, a consumer bank can analyze their customer profiles, figure out which customers have college aged children and offer health, life and college-dorm-renters insurance to them. Such identification of new customer segments is invaluable for FS firms trying to defend their positions and is made possible only by the analyzing, aggregation and consolidation of various types of data. New opportunities such as this example sited here have increased the importance of data to all FS firms. In fact, it has been said that the Glass-Steagall repeal would cause major changes in the IT sector in FS firms and unleash a hunt on Wall Street for experienced information technology executives [Marino96].

But the third, and perhaps most important effect of the latest regulatory reforms is the change in regulations that FS firms are going to face. Consumer banks (and other types of deposit-taking financial institutions) have been traditionally heavily regulated in order to protect consumers. Meanwhile, investment banks, though certainly regulated, have managed to avoid the type of constant scrutiny that consumer banks are used to. But a side effect of repealing Glass-Steagall is that all banks will now come under more regulations (similar in some ways to the previous consumer bank regulations). Glass Steagall's repeal means that three or more regulatory agencies will supervise a FS firm, where previously there was one. For example the Securities and Exchange Commission will look at securities activities, the banking regulators will look at banking activities and the Federal Reserve Board will have enhanced powers to look at all banking activities [Hamilton99]. A summary of the changes in regulation that are brought on is provided by the FFIEC (Federal Financial institutions Examination Council, whose regulatory agencies include the Federal Reserve Board, Federal Deposit Insurance Corporation, Office of the Comptroller of the Currency etc.)[3] This type of detailed and diverse regulatory requirements are a new experience to most FS firms - particularly the big investment banks, but in general to all banks that were not previously commercial banks. Most are simply unable to provide [to regulators] some of the information legally required of them - simply because it is impossible to obtain this information from their organization's data sources. This is

[3] http://www.ffiec.gov/s2-frb.pdf provides a good document that summarizes all the regulations that banks and bank holding companies have to meet. See Section *D. Summary Status Reports*, page 11-38)

causing many firms to take a hard look at their data collection, storage, aggregation and analysis methods. In the next chapter, we will explore some of the problems these firms face when they try to aggregate data from various databases.

One example was seen at a FS firm we observed. A proposed SEC (Securities and Exchange Commission) Reporting Requirement includes something called the Fischer Template. This is an attempt by the SEC to obtain (and provide to investors) better information concerning the nature and quality of the underlying assets, the structural ABS (asset-based security) offerings, and associated risk factors that may have a direct impact on investor return [PSA96]. The template requires that "the measure of risk should cover structural exposures across entire institution and include all relevant exposures, compiled on a net basis across assets, liabilities and off-balance sheet exposures. This information should exclude actively managed exposures covered in Section 1.B[4]" etc. The firm we talked to said that the first 14 pages of the Fischer Template would be left incomplete if they were to fill it in today.

This and similar proposed legislature has caused much concerns for FS firms - enough to cause industry alliance groups to petition the Treasury Dept., SEC and other regulatory bodies. The quote below[5] shows the concerns a coalition of FS firms that are members of the PSA (Public Securities Association) had regarding a proposed "Large Positions" rule:

> "PSA is greatly concerned, however, that certain features of the proposed large position reporting system would inadvertently undermine many of the benefits of an on-demand reporting system. First, in PSA's view, the inclusion in the definition of "gross financing position" of Treasury securities received in pledge will require firms to develop and implement costly systems and related procedures [emphasis added by author] to monitor collateral received on a daily basis in all types of transactions.
> Second [material deleted] .....
> Finally, in PSA's view, the one and one-half day response time to submit the large position report would, in effect, require firms to develop and implement complex systems with the ability to monitor daily Treasury positions in order to gather the information for the report in a timely fashion."

2.5 Need for data for internal purposes

We see from the above discussion how the importance of data within a FS firm is rising due to the natural evolution of technology and the marketplace as well as the regulatory changes. These can be thought of as "external" drivers of data management. But data is not only used to fill lengthy forms for SEC filings and to create new investment products. Ever since FS firms started using data (and IT) for post-operational monitoring instead of merely using data to support daily operational needs, data has been used as a valuable internal decision-making tool [Rockart91, pp10]. Both long term (strategic) and short-term (tactical) decisions are made on the base of extensive data analysis. According to a Forrester report [Schadler97], when asked what business benefit was expected from their data management and data warehousing efforts, 42% of CIO's selected "Better data for making decisions" and "Ability to make decisions faster" as the top benefits. There are numerous examples that highlighted the need for different types of data.

---

[4] Quoted from presentation titled "Enterprise Information and EDSI Standards" by John Helm at the interviewed firm.

[5] From: PSA, To: Director, Bureau of Public Debt, Dept. of Treasury, Date: Mar 18, 1996, Re: Proposed Large Position Rule RIN 1505-AA53, available at http://www.psa.com/reg_sec/LARGCOM6.shtml

In the long-term strategic arena, a FS firm may decide to use its institutional client contact list to attract new customers for their private client/portfolio management business. Or it may realize that profitability is down because the sales force is not incentivized to sell certain types of products. Therefore it is necessary to change their internal compensation system. Usually sales commissions are allocated as a percentage of the sales done by a salesman. But in order to incentivize the sales force, and to encourage them to sell bundled low-margin financial products along with the highly profitable ones, the FS firm may need to allocate compensation on the basis of revenue per customer instead of revenue per trade. This means the firm now needs to change the way profitability (or revenue) data is aggregated.

The short-term applicability of data aggregation may be even more important. Data is used to calculate key risk and exposure numbers within the FS firms. These numbers determine everything from the credit issued to a certain client, whether or not to call in loans, the time to issue margin calls, whether or not to underwrite a security offering of a client.

The risk management function is a critical internal area of a FS firm that uses data in numerous ways. Banking is based on taking calculated risks - higher the risk taken by the firm, higher the profit it hopes to make. In order to manage, monitor and profit from a firm's risk, all the firms positions need to be constantly monitored and analyzed. Risk arising from foreign exchange (FX), credit, certain financial products need to be monitored. In order to do this, complex mathematical models are used. These models depend on data feeds from the firm to calculate accurately risk and return levels. For example, to identify how much exposure (risk) a firm has to Indian Rupees, the firm needs to know how many Indian Rupees it holds and input it to a model (the model will run thousands of scenarios and provide a range of probabilistic risk the firm faces). In order to do this the firm has to first know the how much Indian Rupees it holds. This may not be as simple as calculating the amount of Rupees in the FX accounts. More complex financial deals (perhaps the IPO of an Indian software company that is underwritten jointly with an Indian bank) may have payment guarantees that are made in Indian Rupees instead of US Dollars. A loan given to a US firm building a plant in India may have payment guarantees in Indian Rupees. The firm needs to be able to access all this information when it wants to calculate its risk to the Indian Rupee under various scenarios, easy access to high quality data is important.

The above discussion presents a summary of why data is such a valuable asset to a FS firm and how the ability to aggregate and analyze this data is key to staying ahead of the competition.

3 The problems of aggregating data

The previous chapter provided an overview of the historical and current trends in technology as they related to the FS firm. We have also identified some historical and emerging data needs of FS firms. Here we will examine why it is so difficult to obtain the kind of information FS firms need. After all, the data is "all there" in their systems, right? And most of the systems have been interconnected to each other for years. Even if they aren't, it is not too difficult to connect and access various databases and use data mining tools to get the data/information needed.

We observe that the big issue with data is not that it is unavailable, but that once it is obtained (mined, extracted), it is often not what the extractor hoped to see. Data on the same topic that is obtained from various data sources can be contradictory to each other. Or the raw data obtained doesn't mean anything to the user because he has no way of interpreting it - no definitions of the data are given. This type of data heterogeneity is a constant problem in FS firms. The next sections will study this problem more formally.

We start by identifying what data heterogeneity is and spend some time classifying the numerous instances into several broad categories and provide examples of each category. Next we attempt to identify why such data heterogeneity exists. Finally we discuss the approach of standardization and how that applies as a solution to the problem of data heterogeneity. We will see that certain types of heterogeneity can be eliminated relatively successfully using standards, while others cannot.

3.1 Data Heterogeneity

3.1.1 What is heterogeneous data?

In very simple terms, heterogeneous data is exactly what it means on the surface – information or data that is supposed to mean and represent the same thing, but due to what ever reasons has more than one meaning, representation or interpretation. Since data is simply any form of "information", it can exist anywhere (in a firm's hand written general ledger, for example), but we limit our discussions to heterogeneous data that exists within one or more of a firm's information systems. In other words, we are going to limit our discussion to instances of information existing in a firm's data repositories and having contradictory meanings, multiple classifications or wrong definitions. Though we will limit most examples to the FS sector companies, the problems and example highlighted here are very applicable to most types of firms.

3.1.2 The problems cased by data heterogeneity

We know that FS firms are more and more dependent on internal and external data to stay competitive and to meet regulatory liabilities. A firm will need to aggregate, analyze and report its data daily, weekly, monthly or annually, depending on the purpose. At a given time a firm will want quick answers to questions such as the following in order to make tactical decision:

- What is our firm-wide Market Risk?
- What is my credit risk with AIG?
- Who is buying weather derivatives?
- Which [trading] desks are holding Malaysian Ringitts?
- What is my total exposure to Malaysian Ringitts?
- Generate a consolidated report for all of Joe's positions

- What products did Joe buy in Japan in June 2000?
- If the war in Lebanon escalates and all my clients crash, what is my exposure?
- What was total amount of interest-earning deposits we held at the close of business yesterday?

These are the questions a FS firm need quick answers to.

Often senior managers demand answers to these questions, some one (or groups of people) that work for them will be assigned the task of obtaining the correct information. He or she will then either

- a) access numerous databases (or executive support systems) directly and run queries to obtain the data directly, OR
- b) Contact business or IT personnel from various business units and request that they submit the information (each business unit is asked to provide their "part" of the data that will answer a question). Once this information is collected, the employee in charge of the job will attempt to aggregate/sum-up the data collected and present it to management.

Let us suppose that the manager wants an answer to the question "What is my credit risk with AIS?" (where AIS = Amalgamated Insurance Services, Inc.). Within a FS firm, it is often the case that that more than one business unit deals with AIS and its various divisions/groups/legal entities. Therefore, information has to be obtained from more than one business unit. But each business unit has their own unique way of dealing with AIG and often has their own way of recording information about their interaction with AIS. Depending on the particular database, AIS may be represented as "AIS", "Amalgamated Insurance Services, Inc.", "Amalgamated Insurance Services", "AIS Inc." or by numerous other representations. AIS may also be listed under "clients" in some transactions, while in other transactions AIS may be listed as "underwriter" or "insurer", depending on the particular relationship AIS had with the FS firm for that particular transaction. The person performing a data query has to be aware of all these finer connotations of data in order to be able to obtain the COMPLETE data about AIS from the databases. If he or she does not know that AIS is represented as "AIS Inc." in some databases, he will inadvertently miss some data when a query is performed to find all transactions where "company = "AIS". Even after the information is obtained, he or she has to be aware of the conventions used in each database. Some might report the trade volume in thousands of dollars, while others represent it in dollars. Therefore a trade of "value = 1000" means simply $1000 in one instance but means $1,000,000 in the other. The type of currency that is used is another important factor, especially in foreign trades where US Dollars may not be the default type of currency (the $1000 might mean Australian dollars).

The above is only an example. There are numerous examples and some of these will be presented below. But the key observation is that the existence of data heterogeneity makes it extremely difficult for FS firms to use their data in meaningful ways. Later we will look at why such Representation and other types of errors have entered into a firms data sources.

3.2 A classification of heterogeneous data types

Many of the heterogeneous data problems can be loosely classified into several broad categories for the purpose of analyzing them. They are, data heterogeneity caused by multiple representation, multiple classifications, multiple roles, multiple interpretations, time dependence of data and organizational structure. These are certainly not strictly defined, mutually independent classes of heterogeneous data (you will notice in fact there is overlap between the categories and some problems will fit the description of more than one category). But they do present common trends in the types of data-heterogeneity problems FS firms face. Categorizing makes it easier to understand the kinds of

problems encountered and to look for solutions that may help solve many problems in any given category/class.

3.2.1 Heterogeneity due to Multiple Representations

This is perhaps the type of data heterogeneity that is easiest to explain and to understand. It is also arguably, the type of data heterogeneity that is most commonly found in a firm's data sources. Even more importantly, this may be the easiest problem to solve (by easiest, we don't mean that it is the type of data heterogeneity that can be solved with the cheapest cost or the least amount of time. But the solution is often straightforward, even though it may be labor and resource intensive).

Heterogeneity is introduced into data through multiple representation whenever the same *value* or *content* of data is represented in more than one way within one or more databases.

The following example, drawn from a leading FS firm, shows how the country "Brazil" is represented in multiple ways within some of the firm's databases belonging to the Corporate and Institutional Client Group (CICG).

Figure 2: Multiple Country Codes in FS firm's New York databases

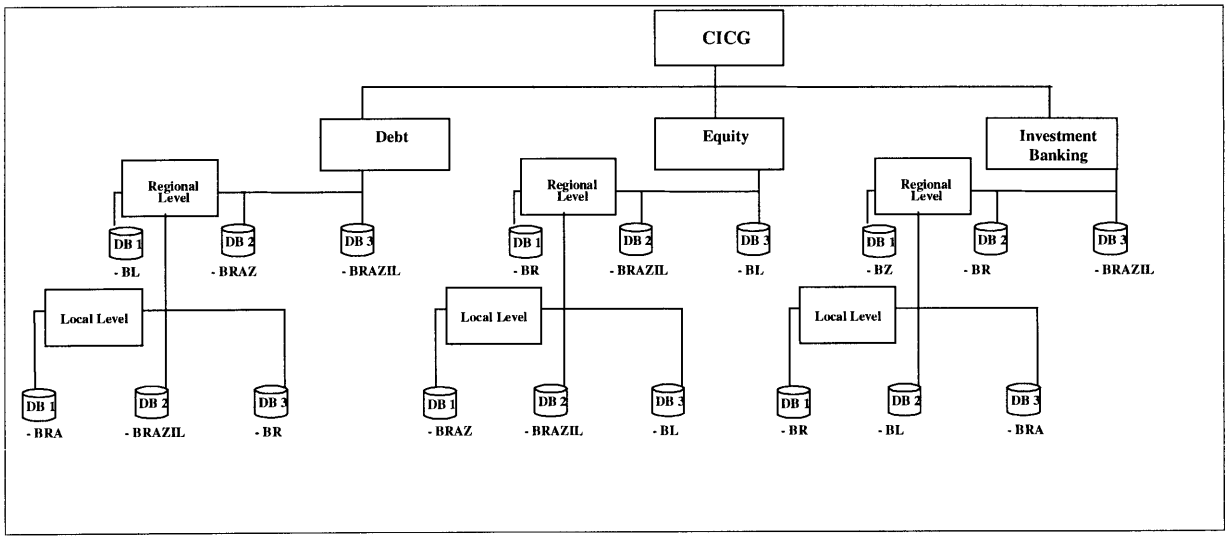


Figure 2: Multiple Country Codes in FS firm's New York databases, shows how multiple databases located in New York identify and represent a country (Brazil) in multiple ways. In the six databases belonging to 3 separate divisions of the FS firm, Debt uses 5 different codes (BL, BRAZ, BRAZIL, BRA and BR), Equity uses 4 codes (BR, BRAZIL, BL and BRAZ) and Investment Banking uses 5 codes (BZ, BR, BRAZIL, BL and BRA) to represent Brazil. The situation gets more complicated if all the other operating divisions across the globe are considered.

The Bank Holding Company (BHC) Act of the US (Section 4090.0 of the BHC Act[6]) requires all holding companies to file reports on country exposure (called the "Country Exposure Information Report", forms FFIEC 009 and 009a) in order to publicly disclose significant country exposure. The fact that these reports need to be filed each quarter means that multinational financial service

[6] Bank Holding Company Supervision Manual is available at http://www.federalreserve.gov

companies (such as the one that provided the above example) need to perform data aggregation on a regular basis. The penalties for false filing can be very high (even up to of $1 Million or 1% of the bank's assets[7]). But how can the firm accurately determine their exposure to Brazil's and Brazil's economy? A group of employees in charge of filing these reports need to query the numerous databases to find the information because more than business unit within the firm deal with clients that are in Brazil, have partnerships in Brazil, or hold positions in securities that are related to Brazil. But the employees cannot simply perform a search of the databases and look for "country = Brazil" or some similar field equal to "Brazil". They have to know the multiple ways "Brazil" is represented across the numerous databases and then search each database with the appropriate variation. Otherwise key transactions and trades can be left out. The situation is dangerous because abbreviations for Brazil are introduced into the firms systems in ad hoc fashion - so the employee performing a data query often does not know that a new representation has been introduced into the set of data sources since he queried the databases last month. Therefore he stands to miss information in his reporting.

Not only the exposure in one country, but the exposure of a bank to a single borrower is also regulated in most countries (depending on the country this could be anywhere between 25% to 65% [Goldstein96][8]). But as a borrower (especially large companies) does business with several different bankers in several different departments of the FS firm, the firm's exposure to the borrower is constantly changing. Having multiple representations of the borrower's name will cause problems as the bank tries to aggregate exposure information (an example would be when the borrower General Motors is represented as GENERAL MOTORS, GM NORTH AMERICA, GENERAL MOTORS NORTH AMERICA, GM, etc.).

3.2.2 Heterogeneity due to Multiple Classifications

Heterogeneity is introduced into data through multiple classifications whenever the same piece of information is classified in more than one way in one or more databases. The "piece of information" here refers to something as simple as a data field - for example, one database classifies a tomato as a vegetable while another database classifies it as a fruit. But they are both referring to the same thing/piece of information, namely a tomato. The relevance to the FS sector is best explained with the example below.

In late summer and early fall of 1998 the Russian financial crisis was beginning to escalate and the stock market reacted immediately to news. Concerns over the current and potential problems in Russia, as well as, its expanding effects on other emerging markets developed. Under such circumstances, FS firms wanted to know what their exposures to those [emerging] markets was – this knowledge would be necessary in order to make decisions that would either leverage the opportunity presented or try to minimize their investment losses. Consequently, Senior Management these firms asked their staff in Finance, Corporate Credit and Risk Management for reports measuring the exposure to crisis-ridden countries such as Russia, Mexico or Brazil.

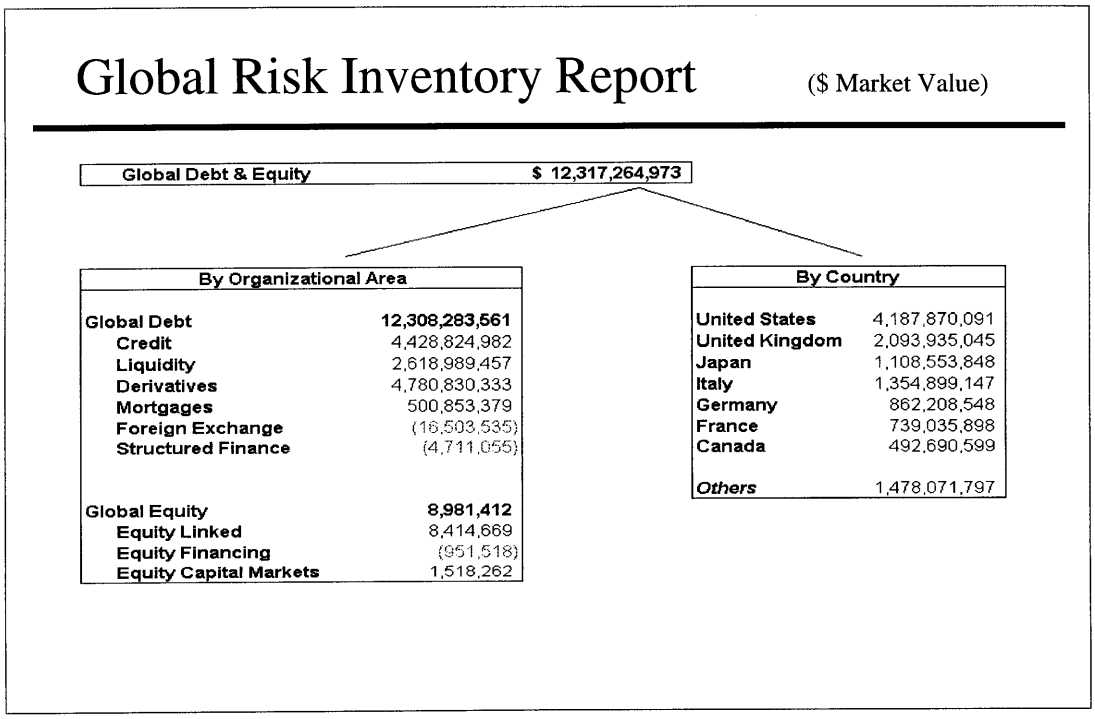
---

[7] Penalties for Errors in Reports Section 2250.0.1, BHC Act of the US, www.federalreserve.gov

[8] [Goldstein96, pp. 21 - Table: Rules on Maximum Exposure to a Single Borrower]

Figure 3 below illustrates one of the dimensions similar to that which the Senior Management one FS firm wanted to see information about its exposure. The right side shows the firms risk broken down by country. This information was collected from each individual country. The left side shows the firms risk as calculated and reported by various business units (the Credit group, Mortgages group, Foreign Exchange Trade Desk etc) which are drawn along product lines instead of along country lines. But the important thing to note is that the sum of the left and right is the same - Global Debt reports approximately $12.308 billion (sum of amounts reported by Credit, Liquidity, Derivatives and other business units), Global Equity reports approximately $0.008 Billion (the sum of amount reported by Capital Markets, Equity Financing and other business units) to give a sum of approximately 12.317 Billion. The same calculation performed by the country risk managers on the right side, sum up to be the same (12.317 Billion) figure. That is, no matter where the data was collected from, as long as it was aggregated properly, the firm would come up with the same figure for its exposure.

Figure 3: Sample of Global Exposure information[9]



In attempting to calculate its exposure in certain countries the firm was hoping to receive reports similar to the one above, but at a country level. So for example Japan's country risk would be on the right, and the sum numbers reported by Japan's Credit group + Japan's Mortgages group + Japan's Foreign Exchange desk would be on the left, and they two numbers would be equal). This would enable the firm to identify how much it stands to lose if (for example) all Japanese companies it conducts business with go bankrupt. But as it turned out, the information that each division reported

[9] Obtained from Enterprise Risk Management Presentation; Microsoft PowerPoint Presentation provided to MIT by Chris Hayward, 1999 to illustrate the problem of aggregating data by multiple dimensions

was in conflict with each other, so that the left and right hand totals did not match. Therefore the job of aggregating the various reports that were provided by the sub divisions into one comprehensive report useful to upper management turned out to be surprisingly complex, and required manual "cleaning up", updating and interpretation of the data provided.

Why is this? Why would the different reports (from different divisions/business units) measure the firm's exposure to Japan differently? One of the key questions (and stumbling blocks) that was faced during this process arose due to the inability to properly identify a "Japanese Company". That is, when is a firm considered Japanese and when is it considered simply a branch/subsidiary of a US (or non-Japanese) firm? In specific, for the purpose of calculating Risk figures, 'what is a Japanese company?'

The reasons that one part of the FS firm classifies a company as being Japanese while another part of firm classifies the same company as being American (or anything else) are often complex. The most obvious is legal and regulatory factors - for example, even though the company is located in Japan and incorporated in Japan, it may be subsidiary of a US (or other country's) company. Therefore even if the Japanese subsidiary/branch/division goes bankrupt, the US headquarters may assume full liability, thus posing no significant risk for the FS firm in Japan. The reasons may also be due to firm's "rules of operations" or locally acceptable standards - employees of the FS firm in Japan probably think of the company in question as being "Japanese" and therefore classify it as such in the databases. Now for the first time when the US CEO wants to calculate his risks, he is obtaining a report from Japan that contains the data that is classified according to the Japanese rules/lingo. There are of course many other reasons why data heterogeneity is introduced due to multiple classifications.

An example of a classification issue (among other problems) was seen during the crash of Daiwa Bank in 1996. Daiwa Bank was a Japanese bank, that operated in the US as a commercial bank named Diawa Bank and several other subsidiaries - Daiwa Trust Company, Daiwa Securities America Limited. During the Senate Hearings that were conducted in the aftermath of the much publicized crash of Daiwa Bank, it was revealed that the FBSEA (Foreign Bank Supervision Enhancement Act) investigators were lax in its inspection of Daiwa Bank because it was a "foreign" bank. But the Daiwa Trust Company on the other hand, was subject to annual external audits just like a US firm[10]. Much of the controversy in this case revolved around how and why Daiwa wasn't audited and inspected more strictly, but relevant to our discussion is the fact that two parts of Daiwa were considered Japanese and US by two different authorities in the US.

3.2.3 Heterogeneity due to the Time Dependence of Data

Heterogeneity is introduced into data sometimes because of the varying nature of data over time. One example of this can be seen in the realm of stock ticker symbols. The symbol 'C' was owned by Chrysler Corporation until its merger with Daimler Corporation, when C was abandoned and 'DCX' was used. The symbol C was then taken over by the newly formed Citigroup. Databases that carry financial information about these two companies now need to reconcile the information – otherwise C could mean Citigroup for recent stock reports while it means Daimler Chrysler in another report that looks at more historical data.

Another example is IBM and Lotus, two companies that existed as separate public companies until IBM bought lotus. But before the purchase, both companies were in the high-tech (software, hardware and systems) business, and may have interacted with the same FS firm. The FS firm

[10] Congressional Hearing on Daiwa Bank, Institute of international Bankers, http://www.iib.org/ibf4.htm

represented IBM and Lotus as separate accounts and managed them separately. But after the purchase of Lotus by IBM, these accounts need to be merged and managed as one. Otherwise, a company executive trying to analyze his firm's historical transaction volume with, say, IBM, will not get the correct information.

Therefore, heterogeneity is introduced into data due to changing nature of information over time.

3.2.4 Heterogeneity due to Multiple Roles

Here we are dealing with the complex relationships and interactions a company (or its various legal entities) can have with another company (or its various legal entities). As an example, a FS firm may deal with Citibank on several levels – as a client (helping Citibank manage its assets through the firm's Asset Management division), as a partner (providing joint underwriting to a third party IPO), and even as a bank (using Citibank's banking services to direct deposit employee payroll checks). Depending on which business unit of the firm one talks to, records in the databases about Citibank will contain any of the 3 (client, partner, bank) under the tag/data field named "relationship". How to treat the relationship is specific to the context of each situation.

This problem was observed at a leading FS firm. The following are the multiple roles that the firm has (so far) identified (presented along with a brief definition of each role).

Figure 4: Multiple Roles played by one entity[11]

Role	Description
Advisee	Party that buys Advisory services from the firm.
Bank	The ultimate bank receiving funds for a party or an intermediary bank used in client settlement instructions.
Beneficial Owner	The individual or entity who is the owner of the underlying collateral for a particular transaction or order.
Broker/Dealer	Acts on behalf of a counterparty. Typically, a Broker/Dealer enters into a transaction with the firm without disclosing the counterparty at the time of a trade and later provides the name of the counterparty involved in the transaction.
Clearing Agent	Acts as agent in the settlement process for clients who are not using a depository for a given transaction.
Counterparty	Buying or selling, or paying or receiving party to a trade
Creditor	Party who lends money to the firm, for example through bank loans or other financing transactions
Custodian	The entity that receives/pays funds or delivers/receives securities on behalf of legal holder, beneficial owner or counterparty. Custodian also provides corporate action services.
Depository	Facilitate the delivery of securities between members by booking entries to reflect ownership instead of physically moving securities
Exchange	A central location where securities or futures trading takes place.

[11] Obtained from CFO-DD-01 - Definitions and Standards, 8/27/98

Guarantor	Any legal entity which backs the performance of a party, through a variety of facilities, including guarantee, letter of credit and provision of security.
Introducing Broker	A broker/dealer who utilizes the firm for clearing and execution purposes.
Investment Advisor	A party which manages accounts on behalf of its clients under contractual agreement (including Fund Managers).
Investor	Party who buys the firm's debt
Issuer	A Party that issues securities.

Supposing one firm (or individual) has dealings with our FS firm in ALL of the above roles? The FS firm needs to be able to differentiate each role and each instance of each role in order to be able to separate-out its dealings with this outside party. Given the amount of data heterogeneity that was introduced into the said firm's data sources due to the existence of multiple roles (such as above), a major "data cleaning" effort was required in order to be able to calculate the firm's risk with regard to each role of party - because depending on the purpose of the aggregation, the roles might need to be treated differently.

3.2.5 Heterogeneity due to Multiple Interpretations

Here we are talking about heterogeneity that is introduced into data due to the same information being interpreted or understood in more than one way.

For example, on any given day one can look at several web sites that provide financial information and find several (different) price-earnings ratios (P/E ratio) for the same company. Why is this? Because in the absence of a standard, the term "P/E Ratio" is interpreted differently by each web site. One web site might report the figures for the most recent quarter, another one for the past 4 quarters, while the other will use the price as a moving average for the past year. On August 21, 2000, the Microsoft web site moneycentral.msn.com reported the P/E Ratio for Daimler Chrysler Corporation (ticker symbol DCX) as $9.50. On the same date (at approximately the same time), www.excite.com reported DCX's P/E Ratio as $10.00, obviously contradicting the first web site's quote. Even if we agree that small differences in numbers can be due to different rounding off methods, a $0.50 difference in $10 is significant. Closer examination shows that moneycentral.msn.com interprets P/E Ratio to be "*the latest closing price divided by the latest 12 moth's earnings per share*[12]" while www.excite.com interprets P/E Ratio to be "*the stock's price divided by its earnings per share for a 12 month period*[13]". Notice that Excite's definition doesn't specify if they use the stocks current (last traded) price, the day's average price, or the day's (or previous day's) closing price. Due to these differences in interpretation, data heterogeneity has been introduced into these data sources that are attempting to report the exact same information.

A more extreme example was observed with regard to financial information reported on the stock of Yahoo (ticker symbol YHOO) on August 26, 2000. At approximately the time, www.excite.com reported the P/E Ratio to be $406.8 while biz.yahoo.com reported it as $419.86. The market cap [14]numbers were even more unequal. The first web site reported Yahoo's market cap to be 73.7 Billion while the second site reported it to be 76.8 Billion - a difference of over $3 Billion!. Though the reason for the discrepancy is no clear from the information on the web sites, we know at least that is

[12] P/E Ratio defined at http://moneycentral.msn.com/investor/glossary/glossary.asp?TermID=365
[13] P/E Ratio defined at http://quicken.excite.com/glossary/notetemplates/htmllist.dcg
[14] Market Cap is the Current Price multiplied by the Number of Shares Outstanding

not caused due to differences in incoming data (both these sites obtain their stock price quotes from Standard & Poors Comstock, Inc. and the price at the moment of investigation was $134 1/4 on both sites. More over, both sites reported the number of Shares Outstanding to be 549,333 Million). Therefore we can presume with reasonable accuracy that the difference is due to the way the two sites interpret (and hence calculate) the above figures.

3.2.6 Heterogeneity due to Organization Structure

This is the type of heterogeneity that is introduced into a firms data sources due the organizational structure of its clients. For example, this type of heterogeneity includes the way the firm identifies Hewlett Packard from Hewlett Packard Puerto Rico. Or the way it identifies (and represents within its databases) IBM and Lotus Corporation (which was an independent company before it was bought by IBM). In calculating the firm's exposure to Hewlett Packard, it may be vital to know that the firm has already extended a large credit line to Hewlett Packard Puerto Rico. On the other hand, Hewlett Packard Puerto Rico may have a credit rating of AAA while Hewlett Packard has a rating of only AA, thus enabling the firm to consider extending lower interest rates to the Puerto Rican subsidiary. A memo from a FS firm that we interviewed underscores a similar issue[15]:

> "Counterparty credit information is critical to the valuation and risk management of credit derivatives trades. Counterparty ratings and country risks are input to the valuation models used to calculate the trade value. A trade with a BBB counterparty is executed at a different level than one with a BB rated counterparty. Misinformation at the time of execution as to counterparty will ultimately result in a hit to trading P&L. Accurate counterparty information is critical."

Similar to issuing credit, being able to calculate a firm's risk depends on the ability to identify uniquely each party that is liable for each trade, should anything go wrong. But when trades are executed without properly identifying which role a client is playing in each particular instance, calculating risk figures is almost impossible. The excerpt from the next memo[16] again underscores this issue.

> "Identification of the correct legal counterparty is critical for Global Debt and Credit Derivatives credit exposure reporting. If the Firm does not accurately identify the counterpart of each derivative transaction, [we] incur the following risks:
>
> - Underestimating credit risk and reserves,
> - Undercollateralizing credit exposures,
> - Increased funding cost from overcollateralization, and
> - Improper equity allocation.
>
> These risks could lead to a loss in the event of a counterparty default. As part of the release of our new credit risk management system, Brutus, we discovered approximately 3,100 transactions with $2 billion MTM exposure that contained more than one counterparty representation."

3.3 Why is heterogeneity introduced into FS firm data?

The reasons for the existence of heterogeneous data are numerous. Here we present some of them.

---

[15] From " CounterParty Repository (CPR): Business Case"

[16] Interoffice memorandum, From: Molly Mathes, Global Derivatives Credit/Collateral; To: Distribution List; June 17, 1998; Subject: Counterparty Repository Initiatives

3.3.1.1 Lack of data standards

Given the globalized nature of the FS industry today, it is hard to imagine that most firms operated in stand-alone mode not so long ago. In fact, most of the systems were developed and were operated in silos – with little interconnection to other systems. Therefore it was not a high priority to develop common data standards that would enhance the exchange of information across systems. Lack of standards is equally important at the firm level. As some of the above examples highlighted, different business units in the same firm have different standards. Different representation standards caused Brazil to be represented in numerous ways. Different classification standards caused a company to be classified both as Japanese and American. Different interpretation standards cause the "P/E ratio" to be interpreted differently.

3.3.1.2 Non-rigorous process of populating the databases and lack of data ownership

A strict process for data entry and storage is needed if a firm wants to achieve a minimal level of data quality. If control data entry is not done, heterogeneity is inevitable. Consider the example where the trader at each trading desk has the flexibility of entering a ticker name for a country as part of entering required information for a trade they just performed. The first time "Brazil" needs to be entered into a data field, the trader can essentially (randomly) pick any of the previously mentioned representation (BRAZIL, BRAZ, BR etc) unless there is some controls set in place. A control could be as simple as presenting him with a standardized list of country tickers so that he can select Brazil from the list. Unless the process is controlled, not only will he randomly chose, but the next trader who enters a Brazil-related trade will choose another representation. With the volume of trades that are entered and processed daily, the FS firm will soon have a collection of representations for Brazil. Clearly defined data ownership responsibility is also key to the avoidance of data heterogeneity. Often, the person entering the data is not the ultimate user of it. In the example of the trader above, though he enters the data, it is unlikely he will ever use that data again. Instead, his superior or more often some other senior manager in a totally different business unit will need to analyze data related to Brazil. This person is the one who will run into the problems of multiple representation. Unless there is some incentive for the trader to enter correct information (for example by basing a part of his/her of compensation on the percentage of trade tickets that were entered accurately), it is difficult to ensure data quality.

Thus, changing the data population process and creating data ownership will help alleviate at least the specific representation problem we identified above.

3.3.1.3 Having to satisfy different data needs of the organization and regulatory agencies

The above causes point to short falls that can be remedied. But we need to clearly understand that heterogeneous data doesn't come into existence *purely* because of the sloppiness of the personnel and the process (though of course this *could* be the case sometimes) or the lack of standards. Sometimes, it is actually *necessary* for a firm to maintain the same data in multiple formats in order to meet the data needs of its many divisions and the regulatory agencies it is governed by.

For example, consider the numerous reports (electronic or paper) that a global FS firm files with the multitude of government authorities/agencies that regulate it. The Securities and Exchange Commission (SEC), Federal Deposit Insurance Corporate (FDIC), Federal Reserve Bank of the US and their equivalent organizations in every country that the firm operates in, requires daily, weekly, monthly and annual reporting on certain key banking data. None of these regulatory bodies require the *same* data, nor do they have the same rules for calculating key figures. Even in instances that they do require the same data, often the required format might be different depending on the country. For

these and many other reasons it may be normal and necessary for a firm to maintain the same information in many formats, many levels of aggregation and many geographically specific naming conventions.

3.3.1.4 Historical reasons

The analysis of computing in Chapter 1 showed that most corporations initially developed their numerous systems individually, without much focus on communication between systems. Then when interconnectivity became an issue, many tools and solutions were found to enable systems to be connected to each other, with the focus being on the systems being able to understand the communication protocol and the (relatively few) pieces of data that was being passed back and forth. But suddenly when these firms wanted to access and analyze the DATA from these systems, these silo systems had to be made truly compatible. Commonality of data representation, classification and interpretation were never focused on before and were now required elements. Therefore we see that the natural evolution of computing and data within firms contributes greatly to the introduction of heterogeneous data - changing such legacy systems can be costly, time consuming and risky.

3.3.1.5 The growth (through mergers, acquisition) of FS firms

Like any other industry, the FS sector has seen the consolidation of many smaller banks, forming behemoth FS firms. It may be possible that an FS firm has, through meticulous data planning, avoided all of the above mentioned heterogeneous data types. But if ever it merges with another, or when one firm takes over another, the 'new' firm inherits systems from both parties that are often guaranteed to be incompatible with each other. Connecting these systems physically is often the easy part. But trying have meaningful connectivity logically involves dealing all the above types of data heterogeneity.
Therefore we see that the natural evolution of business cause heterogeneity to be introduced into a firm's data sources, even if it has sound data management practices.

3.4 Dealing with data heterogeneity through standardization

3.4.1 Common standardization approaches

Given the above causes for data heterogeneity, it is perhaps already clear that having standardized definitions and standardized representations is a good way to avoid the problem. In a perfect world, all data elements would be clearly defined and all formats and representations of data would be standardized. This would make data aggregation and analysis relatively simple. Most firms that chose the data standardization approach have chosen to do one or more of the following:

- Standardize the definition of key data elements within the organization
- Standardize the data entry/control process so that data quality is maintained
- Create authoritative data sources that all organizational units refer to (for that particular data)
- Standardize the communication protocols within systems or databases
- Standardize how data will be represented in databases

Following the above and many other paths, successful standardization efforts have been accomplished at the inter as well as intra firm level.

3.4.2 A common standardization process

The standardization processes that are followed by most firms usually have in common the steps identified below:

1. Identify the heterogeneous data and the specific databases or systems they occur in
2. Establish an organization wide standard and attempt to define what the particular data will mean (or how it will be represented/classified) from that point in time onwards
3. Incorporate the "new" definition/representation into the data-sources either by (a) clearing up all the data-sources until they are compliant to the standard (i.e. "clean-up" the existing data), or (b) establishing a centralized authoritative data source for the particular data element.

3.4.3 The problems with attempting standardization

So why is that we don't see all FS firms scurrying to undertake some data standardization project? The fact is that along with the successes there exist many standardization projects that have been failed, by any measure we chose to judge it by. The following may be some factors that deter standardization as a solution, or have rendered some past standardization effects ineffective.

3.4.3.1 Organizational Issues

One of the most (if not the most) important stumbling blocks standardization attempts run into are related to human and organizational problems. Standardization initiatives often run into this problem in step 2 of the standardization process. Setting up of standards implies that one way of "doing things" is accepted while all others are discarded. That is, the multiple definitions, representations, classifications used by the different business units will have to be discarded and a single definition, representation, classification needs to be accepted in order for the project to proceed. In this type of situation there is a high incentive for each business unit to fight for "their" definition, representation, classification to be used as the standard for the whole organization. If successful, it means that the "winning" department can avoid having to allocate a budget to clean up their data-sources, since they are already compliant with the standard. This is a situation that almost invites turf wars and long drawn out negotiations about the specifics of the standard.

Supposing a standardization attempt gets through this first hurdle, the battle is not over. Once a small group of employees have worked hard to create the new standards, numerous others are now needed to implement the recommended standard. But often the rest of the organization does not see the real benefits of standardization (or the cost of non-standardization). They may view the attempt as just another "flavor of the month" attempt of management that is only going to distract them from their "real" jobs - the one that pays their performance bonus. Therefore the implementation efforts meet with resistance or ends up with a group of employees who are unenthusiastic about the attempt. Sometimes this type of reaction is due to the lack of communication from the initial team, or due to the lack of commitment from senior management or due to the fact that employee incentives are not tied to standardization attempts.

3.4.3.2 Cost

Though an effective solution to the data heterogeneity problem in many situations, standardization is extremely expensive. It is also complex as it is costly and often it is difficult to see the results. The costs are incurred mainly during the implementation phase - data "clean-up" often requires many personnel to spend numerous hours on each data-source. So even if a firm manages to avoid the organizational issues mentioned above, budget constraints will often dampen standardization attempts. This occurs often when the firm makes each business unit pay for the standardization efforts. In fact, Forrester reports 58% of the time, firms make users pay for standards [Deutsch98].

But if the benefits cannot be VERY clearly laid out, it is unlikely that business units will pay for the high costs.

Even if the individual business units are not paying, standardization still represents considerable cost when compared to the chances of failure. There is certainly more than one example of a failed standardization attempt that has cost large sums of money. For example, [Alverstrand95] points out failed national and international standardization attempts . [Schmidt98] provides a case study that contains many failed standards for facsimile terminals.

3.4.3.3 Time

Standardization attempts take time - identifying the problem, coming up with standards and implementing them is not a simple task. Many employees need to spend considerable lengths of time devoted to standardization attempts. The standardized data is only available to the firm's use after a considerable period of time.
Therefore standardization may not be the best method of dealing with data heterogeneity, especially in FS firms where the need to see immediate benefits of standardization is high.

3.4.3.4 Ability to keep up with changing business environment

Even if these organizational politics can be avoided, and budgetary constraints are non-existent, standardization is still a colossal task. But no matter how much money is spent, standards cannot keep up with the future needs or the fast changing phase of the financial services industry. It isn't surprising that by the time a large standardization effort has ended, the rules of the game have changed so much that the standards are no longer up-to-date, valid or even useful. And not even the most careful standards committee can predict the future to be able to select a standard today to meet future data needs. After all, it is at least theoretically possible to create excellent standards that meet all the current data needs of a firm. But changes in law, mergers, acquisition and takeovers happen often. Each time some such change occurs, the standards have to be re-visited. Either a new project needs to be undertaken to update the old standards or the old standards need to be discarded altogether. This inability of standards to evolve easily is a major shortfall of the standardization approach. We will see in a later chapter a type of technology that may be able to solve some of the issues of standard evolution.

3.4.3.5 Difficulty in enforcing standards

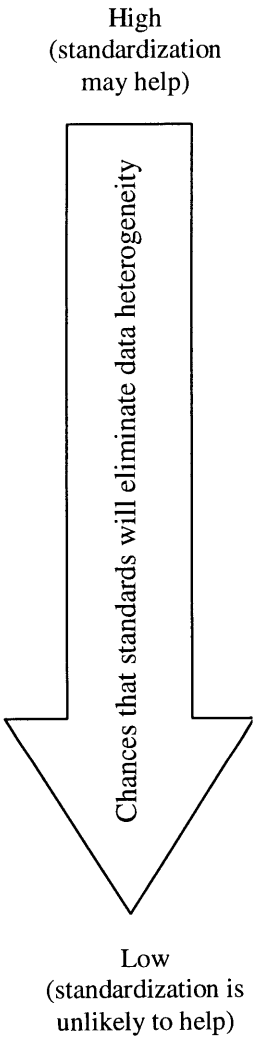
Standards are useful only if they are implemented and enforced. But enforcement of standards tends to be quite difficult. With global organization with operational units that are highly dispersed geographically, it is difficult to convince all employees about the virtues of standards. In most firms, the standardization committee issues a "directive" or a set of "rules and regulations" that include standards that all developers of new data-sources should follow. But it is difficult to check if these standards are adhered to, unless each project is forced to be reviewed by a "standards committee", just like it is reviewed by a budgetary and technology purposes. But even if this is done, the business needs for speed and flexibility can override standards - a business unit may decide to authorize a non-standard-compliant database to be built because it needs to store the data immediately. A survey by Forrester Research showed that 69% of CEO's will "allow a violation of standards if there is strong business justification [Meringer97]".

3.4.4 Standardization as a solution to dealing with heterogeneous data categories

Suppose a firm was able to avoid all the problems listed above and successfully complete a data standardization effort. Does this mean that data heterogeneity will be eliminated from all its data sources? Will some types of heterogeneity still remain? At least can all the questions listed in section 3.1.2 (titled "The problems cased by data heterogeneity") be answered after standardization?

Figure 5 below is an attempt to analyze which category of heterogeneous data (presented earlier) can potentially be eliminated by a standards based approach. Note that here we are assuming that all the problems inherent with standardization (high costs, organizational politics, time constraints) do not exists - therefore we are purely looking at a 'super-successful', hypothetical standards effort.

Figure 5: Ability of standardization to solve heterogeneity



Type of Heterogeneity

Multiple Representation:
For example, standards could mandate and enforce that Brazil will be represented in BRZ in all the firms databases

Multiple Roles
For example, the firm can standardize multiple roles/relationships and establish ways of identifying each. As a simplistic example, Citibank will be identified by unique ID 988712xx where the xx varies depending on Citibank's role (00 = bank, 01 = client, 11 = partner etc.)

Multiple Interpretations
For example the firm can mandate and enforce the rule that when calculating P/E Rations, P will always be the closing stock price from the previous day and that E will always be the earnings per share for the most recent 12 months.

Organizational Relationships
For example, establish a hierarchical accounts so that the account number indicates which division of each client is dealing with the firm. But different interpretations may be needed depending on purpose.

Multiple Classification
Because classification is too dependant on the context the data is being used for and the context that it was created in. for example a company may be "Japanese" for tax purposes, but "American" for certain risk, liability or bankrupcy purposes

Time Dependence
What changes will occur in the future are not predictable. Therefore the changes in Data (which are representations of facts) in the future cannot be planned, estimate or predicted).

We have to point out that the above picture is drawn assuming that the particular standards can be established instantaneously and that the business environment is changing. But the moment the environment does change, through a merger of two companies, perhaps, most standards need to be updated. This take a considerable amount of time and often the any previous investments the firms

have individually made in creating standards are invalidate because everything needs to be reworked. In other words, standards, even if they "work" for some of the categories listed above, are not able to evolve easily.

4 Inter-firm (industry wide) standardization attempts

This chapter is devoted to industry-wide standardization attempts. Though the focus is certainly on the FS sector, the standards discussed are not necessarily limited to banks. Some of the described standards are more than just data standards - they provide a complete, standardized, infrastructure to facilitate communication and economic activity between those who adhere to the standard.

Out of the numerous standards that exist today, a relatively small sub set is presented here. They are selected because of their applicability and importance to FS firms. A large amount of commerce that happens daily depends on these standards. Almost all of these were started to create a particular need that was common to more than one firm. Identification of such a need was catalytic in bringing a diverse group of people together to work towards finding a workable standard. Once a standard was established, a smaller subset of institutions has been in charge of monitoring it and keeping it up-to-date.

4.1 CUSIP[17]

4.1.1 History and Purpose:

The Committee on Uniform Securities Identification Procedures (CUSIP) Service Bureau was formed to study and to create a solution to the problem of not being able to accurately and efficiently perform the clearance of securities due to the lack of a unique number to identify each security. With the participation of many players of the FS industry who felt the genuine need for such an identification scheme, CUSIP numbers were introduced to uniquely identify each security. CUSIP identifiers have been used since the 1960's to identify more than four million securities.

CUSIP numbers are currently assigned by the CUSIP Service Bureau, operated by Standard & Poor's for the American Bankers Association[18]. The Bureau exists for the primary purpose of uniquely identifying issuers and issues of securities and financial instruments within a standard framework, and disseminating this data to the financial marketplace via various media. CUSIP numbers and standardized descriptions are used by virtually all sectors of the financial industry, and are critical for the accurate and efficient clearance and settlement of securities as well as back-office processing. The numbering systems was designed with the goal of allowing it to be flexible enough to meet future as well as current operational requirements. Thus it is adaptable to the internal systems of all users, to communication systems, to automated document readers etc. and the structure of the system allows each user to assign numbers to securities or other assets carried by him but not covered by him.

4.1.2 What is covered by CUSIP

1. Any Corporate, Municipal, Government and Private Placement security (debt or equity) issued in the U.S. or Canada
2. Mortgage-backed futures contracts (known was TBA[19]'s)

---

[17] www.cusip.com

[18] Control of all CUSIP numbers technically belongs to the CUSIP Board of Trustees.

[19] A TBA type CUSIP incorporates the security's mortgage type, coupon maturity, settlement month etc. within the number itself

3. Securities actively traded on an international basis which are underwritten (debt issues) or domiciled (equities) outside the US or Canada. A CINS Number (CUSIP International Numbering System) was developed in 1998 to identify these international issues. It carries the same 9 digit format of a CUSIP number (6 characters for the issuer, 2 characters for the Issue and the last for the check digit). CINS numbers are listed in the International Securities Identification Directory (ISID), which allow users of CINS to cross references to all other major international numbering systems.

4.1.3 Example

1. CUSIP = 837649128

   The first six positions (837649) are the **issuer number**. A single alphabetical file has been developed for corporate, municipal, and governmental issuers, and an issuer number of six digits has been assigned to each in alphabetical sequence.

   The next two positions (12) are the **issue number.** If these two characters are purely numeric, it's an equity security. Of there is at least one alphabetic character in these two positions, it's a fixed income security.

   The last position is the **check digit**, which is based on Modulus 10 Double Add Double technique as follows:

   ```
     8   3   7   6   4   9   1   2
   x 1 x 2 x 1 x 2 x 1 x 2 x 1 x 2
   --- --- --- ---- --- --- --- ---
     8   6   7  12   4  18   1   4
   ```

   Thus, 8 + 6 + 7 + 1 + 2 + 4 + 1 + 8 + 1 + 4 = 42

   The complement of the last digit of the sum becomes the check digit. The complement of 2 is 8; therefore, the CUSIP number with issuer number 837649 and issue number 12 would appear as 837649 12 8 with the check digit.

   Any alphabetical character (A-Z) are assigned numerical value for this calculation (e.g. the letter A=10, B=11, C=12 etc.)

2. CINS = Z 23456 78 9

   The first position "Z" denotes the country of origin or geographical region of the issuer.

   The next five positions (positions 2 to 6) identifies the issuer or company name and is used only for securities issued by that issuer.

   The next two positions (7-8) identify the different issues for a particular issuer. For example, the code may reflect ordinary shares, preferred stocks, warrants, or debt issues.

   The last is the check digit, as before with CUSIP.

4.1.4 Future directions and issues

As thousands of new securities are issued each day, CUSIP has started to run out of unique numbers. Therefore some of the numbers have started to be recycled (after a long period of time has elapsed of course). But this could cause problems if extremely old securities need to be investigated and analyzed.

The second problem with CUSIP is that it is currently limited to North American securities. Even CINS is essentially a number created to help the American FS community to identify the limited international securities it trades. But as the barriers for foreign investments fall and cross boarder securities sales become common place, CUSIP runs the risk of being marginalized. Standard & Poors (who maintain CUSIP) has joined with Telekurs Financial to create a definitive, international

securities identification system called ISID*Plus*[20]. The ISIDPlus database will contain a CINS or CUSIP number for every security traded outside of North America, cross-referenced to every other national security identification number assigned to that security. If ISID*Plus* takes off, most other identifiers such as CUSIP will be rendered useless.

Another issue is that CUSIP uses an alphabetical sequence to identify the issuer number. But as the firms merge and change their names, all previous CUSIPs belonging to that firm need to re-issued in order to keep up with the changes. For example when Bell Atalantic became Verizon wireless, all CUSIPs belong to the company needed to be changed so that the 'issuer number' part would correctly point to Verizon, instead of Bell Atlantic.

4.2 SWIFT[21]

4.2.1 History and Purpose

In the early 1970's six European and North American banks commissioned a feasibility study called the Message Switching Project. The Society for Worldwide Interbank Financial Telecommunications (SWIFT) was formed in 1973 and it developed messaging formats that would facilitate the exchange of financial data worldwide securely, cost effectively and reliably through its global network. SWIFT is more than a data standard - it is also a complete messaging protocol and architecture.

It is currently widely accepted across financial services firms - in 1999 the one-billionth SWIFT messages was transmitted. More than 7000 customers in 191 countries worldwide use the SWIFT network in daily operations.
This wide acceptance has had the effect of encouraging other vendors of financial services and payment products to make their products compatible with SWIFT. Among these SWIFT-enabled products are:

- Payment products such as PayPlus$ (by FundTech), ProPay (by Provida), Infinity Message Manager (by SunGard)
- Securities processing and enabling products such as GobalPlus (by SunGard Asset Management), GEOS (by Software Daten Services)
- Reconciliation products such as Intellimatch (by MicroBank), FTMS (by Management Data), SMART Stream Reconcilliations (by Geac)
- Electronic Banking products such as BankQuest (by Bottomline Technologies)
- Middleware such as ProSwitch (by Provida), Mercator (by TSI Software)

4.2.2 What is covered by SWIFT

All types of financial transactions such as the following are covered:
Operations:

1. All foreign exchange (FX) transactions (options allocations, transfers, hold orders, interest rate swaps, multi party FX transactions, customer and bank payment instructions)
2. Loan deposit, confirmation orders
3. Derivative Re-couponing and terminating messages
4. Settlement and Clearing messages, orders (advice of payment/non-payment, advice of acceptance, advice of collection etc.)
5. Trade confirmation messages

---

[20] http://www.isidplus.com
[21] www.swift.com

6. Many other exotic financial messages.

Each message will contain information about business elements similar to the fields below (the example is for a MT100 Single Customer Credit Transfer message). Depending on the type of SWIFT message, a subset of the following fields or many other fields will be used.) :

- Transaction Reference Number
- Value Date, Currency Code and Amount
- Ordering Customer
- Ordering Institution
- Senders correspondent
- Receiver's correspondent
- Intermediary
- Account with institution
- Beneficiary Customer
- Details of Payment
- Details of Charges

In turn, each field will have a separate subset of standardized definitions and instruction. For example, the specification for the field "Currency Code Amount" specifies that:

- The format should be in the order of (Date)(Currency)(Amount)
- The (Date) should be in YYMMDD format
- (Currency) should be in ISO 4217 currency code
- Integer part of Amount must contain at least one digit
- Etc...

Fields such as "Ordering Customer" may have many sub-fields/structures (name, account number, address). Regional and country customizations are possible for certain fields[22], subject to the approval of the specific local regulatory bodies.

4.2.3 Future directions and issues

With the new focus on STP (Straight Through Processing), SWIFT is likely to become even more popular than before. STP is an effort to have all electronic financial transactions cleared and processed in "one-go", without resort to manual intervention (thus reducing the re-work/re-entry of data by humans and slowing down of the money flow across institutions). For example, a simple error such as omitting the "/" character that is required in the ACCOUNT NUMBER field (field 59) in a MT100 SWIFT message[23] means that human intervention is needed to correct or clarify this error, thus disabling any hope of STP for the particular transaction. Towards this end, SWIFT is developing new messaging standards (such as adding extra message failure protocols) that will enhance the existing protocols and facilitate STP [SWIFT99].

4.3 FIX and FIXML[24]

The Financial Information Exchange (FIX) is a messaging standard developed specifically for real-time electronic exchange of securities transactions. FIX is a public domain specification owned and maintained by Fix Protocol, Ltd.

The FIX effort was initiated in 1992 by a group of institutions and brokers interested in streamlining their trading processes. These firms felt that they, and the industry as a whole, could benefit from efficiencies derived through the electronic communication of indications, orders and executions. The

---

[22] Securities Market Practice, publication by SWIFT (www.swift.com)

[23] MT100 is the name for a Single Customer Transfer payment message and is the most widely used type of SWIFT message

[24] www.fixprotocol.org

result is FIX; an open message standard controlled by no single entity, which can be structured to match the business requirements of each firm.
FIXML is the XML derived grammar of the FIX protocol. A FIXML implementation will have message format validation, more expressive structure, and leverage off existing standards.

4.3.1 What is covered by FIX

FIX is a messaging standard and protocol and it allows for the transmission of numerous types of messages, some of which include:

- Advertisement
- Indications of Interest
- News
- Email
- Quote Requests, Quotes, Quote Acknowledgements
- Market Data Request, Market Data Snapshot, Market Data Request Rejects
- Tradiing Session Status
- New Order, Order Cancel, Order Replace
- Settlement Instructions
- Bid Requests, Bid Response

4.3.2 Future directions and issues

FIX's flexible tag-value message format is a double-edged sword. It imposes no structural constraints on a message, so all but the most trivial validation must happen at the application level. The increasing number and complexity of FIX application messages creates ambiguities that are handled differently by diverse systems. This type of flexibility, though useful at times may be problems as interconnectivity between firms increase and the focus intesifies on the the ability to process and settle finances automatically. For full automation, a more structured approach for messaging is needed and it is possible that FIX will not be able to meet this need. But the FIXML is an effort that is underway to provide some much needed structure to FIX.

4.4 XBRL (eXtensible Business Reporting Language)[25]

XBRL was created for the purpose of preparation and exchange of business reports and data.
The XBRL initiative was initially funded by the AICPA (American Institute of Certified Public Accountants). Soon many other companies (such as the Big-4 accounting firms, FreeEDGAR.com Inc, Microsoft, Multex, Morgan Stanley Dean Witter, Oracle, IBM) and industry/professional consortiums (such as Canadian and Australian Institutes of Charted Accountants etc) joined the effort. Today, most key players of the financial information supply chain are involved in developing the standard.

XBRL hopes to simplify the exchange of financial data by introducing a universal system of XML tags that will identify the function of each piece of financial data. Currently several different - and incompatible - formats are used to publish this data.

XBRL is:

- A standards-based method with which users can prepare, publish (in a variety of formats), exchange and analyze financial statements and the information they contain.

[25] www.xfrml.org

- Freely licensed, and permits the automatic exchange and reliable extraction of financial information across all software formats and technologies, including the Internet.
- Ultimately benefits all users of the financial information supply chain: public and private companies, the accounting profession, regulators, analysts, the investment community, capital markets and lenders, as well as key third parties such as software developers and data aggregators.
- Does not require a company to disclose any additional information beyond that which they normally disclose under existing accounting standards. Does not require a change to existing accounting standards.
- Improves access to financial information/speeds up access
- Reduces need to enter financial information more than one time, reducing the risk of data entry error and eliminating the need to manually key information for various formats, (printed financial statement, an HTML document for a company's Web site, an EDGAR filing document, a raw XML file or other specialized reporting formats such as credit reports and loan documents) thereby lowering a company's cost to prepare and distribute its financial statements while improving investor or analyst access to information.
- Leverages efficiencies of the Internet as today's primary source of financial information by making Web browser searches more accurate and relevant. (More than 80% of major US public companies provide some type of financial disclosure on the Internet.)
- XBRL meets the needs of today's investors and other users of financial information by providing accurate and reliable information to help them make informed financial decisions.

4.4.1 Future directions and issues

The currently released standard only covers specifications for publishing companies' financial statements. Other specifications will cover additional types of business repots - such as regulatory reports including Securities and Exchange Commissions EDGAR files, tax filings and business event reports such as press releases - will be issued within the next 18 to 24 months. It is believed that this feature (of being able to file taxes and SEC reports) will be a major selling point for XBRL and may expedite its adoption as an industry standard.

Meanwhile, the XBRL Project Committee is working with vendors such as SAP who are already working to integrate XBRL directly into their software, so when a customer wants to run their financial statements, XBRL is an option [Trombly00-1].

The real value of XBRL may be yet to come. If it succeeds in its stated goal of establishing standards for *calculating* common financials, problems such as the P/E Ratio problem we identified in Section 3 could be avoided.

But bringing relatively diverse global accounting practices to meet one standard is easier said than done. Therefore XBRL is still a limited solution.

4.5 Bar Codes

To alleviate the problem of collecting accurate product information, the consumer product industry developed UPC (Unified Product Code) or bar code. It has given manufacturers and retailers the ability to manage their supply chain and inventory levels more efficiently and track how successful each line of product is in the market. The bar code is one the most successful standards created - almost every product you buy today contains a unique bar code. The ability to scan a bar code

enables it to be sent directly to computerized processing systems - this allows for instant updates on the levels of stock at a retailer and eliminates the need to manually enter data about daily sales etc. The code itself contains two parts. The first part identifies the manufacturer. It uses a manufacturer number to do so and the Uniform Code Council assigns this number. The second part is customizable and each manufacturer can use his numbering to identify his products.

Because bar codes are so widely accepted and because they truly facilitate the process of data aggregation (and hence improve the ability to make strategic decisions) we believe it will continue to be one of the best examples of standardization.

4.6 DUNS Number

The DUNS (Data Universal Numbering System) has been created and is maintained by Dun & Bradstreet. It is unique 9-digit number that allows the identification and linkage of more than 57 million companies worldwide. It is an internationally recognized common company identifier that is used in EDI (Electronic Data Interchange) and electronic commerce transactions. The numbering system allows the building of corporate family relationships so that identification of the parent, subsidies, headquarters and branches of firms is possible. Due to this feature, the DUNS numbering systems is widely used by numerous industry sectors. When a FS firm wants to aggregate information from all the various divisions of a major client, the DUNS numbering hierarchy can be useful.

It is also advantageous to use a numbering system such as DUNS instead of creating an internal party identification system because DUNS is constantly maintained and updated by Dun and Bradstreet, thus eliminating the need for the FS firm to be aware of each new branch a client opens up (for example).

But DUNS doesn't identify partnerships, join ventures or other types of special relationships between firms. This is a limitation for some FS firms that routinely enter into partnerships with other FS firms to jointly fund many transactions.

4.7 Some other Standards

The above standards are relatively well established in that they have been in use for a considerable period of time and are accepted by a relatively large number of organizations (probably because they help solve the specific problem it was supposed to solve in the particular industry).
There are also newer, emerging standards that are being created daily. These are driven primarily electronic business environment. The need to communicate, transact and process information quickly, accurately and purely-electronically has never been more important. Therefore many different, often competing, standards are beginning to be formulated by various industry groups and other organizations. The following are only two of many such emerging standards.

ebXML (Electronic Business XML): This is an attempt to define a common structure for interoperable messages sent between companies and among industries[26]. Unlike most other messaging frameworks, however, ebXML is creating an end-to-end architecture that includes not only messaging but also generic business process models, a core set of common data components and distributed repositories for storing industry or company requirements.

[26] www.ebxml.org

BizTalk: BizTalk is an industry initiative started by Microsoft and supported by a wide range of organizations, from technology vendors like SAP and CommerceOne to technology users like Ariba[27].

The above two standards as well as many of the other newer standards have one thing in common - they make use of XML (eXtensible Markup Language) as a foundation. XML is a markup language for documents containing structured information. It is a meta-language for describing markup languages. That is, XML provides a facility to define tags and the structural relationships between them. Since there's no predefined tag set, there can't be any preconceived semantics. All of the semantics of an XML document will either be defined by the applications that process them or by stylesheets.

[27] www.biztalk.org

5 An Intra-firm standardization attempt: The Enterprise Data Standardization Initiative (EDSI)

5.1 Background

Just as the FS industry as a whole continues to carry out standardization attempts (often quite successfully), individual firms too have attempted to achieve standardization of data and systems. In this section we will present an example of a (ongoing) standardization effort at a major Capital Markets/Financial Services firm[28].

As one of the largest and most profitable players in the FS sector, the said firm faces the challenge of keeping all its multitude of divisions, groups and legal entities operating in complete unison and working towards achieving its profitability goals. The firm's senior management has to be able to react to the changing needs of its customers and regulators. It also needs to be able to accurately measure its strengths and weaknesses in order to be able to take proactive, strategic measures to increase profits and market share. As discussed in Chapter 2, the ability to manage, aggregate and process data is vitally important if the firm hopes to fulfill these needs.

As one of the older FS firms, this firm has inherited the typical problems that are associated with numerous systems that often duplicate data. It is not uncommon to have separate entities at the firm dealing with the same client and storing that information in separate, unrelated databases, using two different formats. In other words, due to the reasons explained in Chapter 3, the firm has inherited the heterogeneous data problems that were discussed earlier.

The firm's senior management has been aware of the problem since heterogeneous data has caused problems in the past - not only when data is aggregated for strategic decision making, but also when trying to streamline day-to-day operations. The following quote from an internal memo highlights this issue[29].

> "We have suffered numerous monetary losses due to the existence of duplicate counterparties in our current counterparty databases. As an example, we had three transactions with a single counterparty, represented by three unique counterparty identifiers in our current counterparty database. When the counterparty chose to amend their settlement instructions, we amended the settlement instructions for one of the three representations. When it came to settle on the other two transactions, we paid the counterparty at their previous instructions, which were no longer valid. It took two days to redirect the funds to the counterparty's amended settlement instructions, and we incurred interest charges as well as losing operational respect"

Another hint at internal data problems was seen during the Asian economic crisis of 1998. Companies like the firm we interviewed needed to assess their exposure to Asian markets. But it took considerable effort and time to obtain an accurate idea about the firm's exposure to this market because the various heterogeneous data standards across business units made it difficult to aggregate risk/exposure figures quickly. The figures that were reported were unfortunately far from meeting the level of accuracy that the firm needed.

---

[28] We will refer to them as 'the firm' in our discussions.

[29] Interoffice memorandum, From: Molly Mathes, Global Derivatives Credit/Collateral; To: Distribution List; June 17, 1998; Subject: Counterparty Repository Initiatives.

5.2 Enterprise Data Standards Initiative (EDSI) [30]

Events or situations like the ones above highlighted the need to be able to aggregate information accurately. Although the need exists to accurately aggregate all information, the firm realizes that it vital to at least be able to measure such things as the exposure to a certain market, a client or a product. With this goal in mind, they have undertaken several standardization projects over the years. The Counter-Party Repository (CPR) Initiative was undertaken to identify uniquely each counterparty (any entity that the firm did business with) and avoid redundancy of multiple representation for the same counterparty. The CFO Committee on Data Entity Definitions was put together in the Summer of 1998 and was charged with identifying key data elements, defining them, identifying and defining several key attributes for each data element, recommending ownership for this data and communicating these definitions across the organization. These are just two recent examples of standardization efforts. The stated objectives of these two relatively small projects have been met, but there is much more work to be done because only a subset of the firms data was tergeted in these projects. There has also been numerous other attempts at standardization initiatives at this firm.

Here we present the status of a current, major standardization initiative at the firm called the Enterprise Data Standardization Initiative (EDSI).

5.3 The Business Drivers

Past and present problems are driving the EDSI initiative as mentioned above. An internal presentation provides us with anecdotal evidence of the need for EDSI:

> 'We are drowning in our own data and are unable to transform our "data" into "information". Due to inconsistencies in data identification, understanding our risk exposure within and across [the firm's] entities, in a timely and efficient manner, is near impossible. Improperly calculating collateral requirements and margins result in unnecessary expense and regulatory issues.'

The key business drivers can be categorized into three areas.

5.3.1 Regulatory and Financial Exposure

One of the key factors that are driving the EDSI is a legal and regulatory factor. As we discussed in Chapter 2, the repealing of the Glass Steagall Act and the introduction of HR10 is bringing about major regulatory changes in the banking world. The multitude of regulatory controls that were previously limited to banks now apply to securities firms. Not only that, numerous new regulations have been introduced on top of previously existing ones. HR10 and the fact that the firm now *wants* to be a bank means that it will be regulated as one. One executive estimates that the number of regulations has increased "three fold" since HR10 was enacted.

Unfortunately, it is expected that the existing systems at the firm will have difficulty meeting these regulatory requirements. The proper calculation of risk and exposure numbers (at the level and depth required by the regulations) is expected to be difficult. For example, it needs to consolidate/aggregate and report (to regulating bodies), information about core classes of data - for example, it needs to file reports that are consolidated along products or clients. But the lack of standard common identifiers and redundancies make this type of consolidation time consuming. Similarly, calculating Value at

[30] Information presented in this section is based on interviews done at the as well as copies of several presentations given to the firm's senior management by the EDSI project task force.

Risk and Capital at Risk numbers is difficult because it is not possible to accurately consolidate information for all of the firm's counterparties[31]. This may be due to the fact that one client (also known as a 'counterparty') may be represented in multiple ways in one or more databases. Or it may be because two sub units/divisions of a client will have two separate operating accounts with the firm, without there existing any link between the two accounts (which would enable the proper calculation of the firm's exposure to this particular client).

Due to its data problems, the firm might risk reporting inaccurate information to regulators. Detection of such mistakes could mean punishment, usually in the form of fines, but could also mean the revoking of an operating license in extreme situations. But apart from the potential inaccuracies it might report, the firm thinks that there are some other important numbers that is unable even calculate. An example is the Fischer Template that is part of the soon-to-be-mandatory SEC Risk Reporting Requirements. According executives, the first 14 pages of this template would have to be left blank today because of difficulties the firm will face when attempting to estimate risk under various scenarios.

The firm's far sightedness and business savvy in anticipating upcoming regulatory changes discussed here is perhaps the biggest driver of EDSI.

5.3.2 Client Relationship Exposure

The firm's inability to aggregate client positions across legal and operating entities can cause problems when it tries to accurately evaluate a client's holistic financial relationship with the firm. For example, suppose the firm performs business with several units of General Motors (GMAC, GM, General Motors Venezuela etc.). At present, each business unit at the firm may deal separately with each unit at GM, without coordination. But how can the firm assess if it is too exposed to GM in the hypothetical event of the auto industry experiencing a major business downturn? How can it correctly calculate the total amount of credit it has extended to GM as a whole when each business unit uses separate systems that don't communicate with each other store this information in different places? How can it advice and help GM as a whole to re-structure its debt unless it knows exactly where GM's money is as a company (instead of knowing about where GMAC, GM, GM Venezuela individually invest money)?

This type of problem can be expected not only with corporate clients, but also with individual customers who deal with separate divisions of the firm. The probability of issuing incorrect margin calls is greatly increased because the firm has some problems correctly measuring its exposure to one client – this can affect the client relationship greatly.

Even in other facets of client interaction, the ability to identify all of a client's relationships with the firm is vital – anecdotal evidence can be seen even in the portion of an internal memo[32] quoted here:

> "As part of the initiative to report and withhold interest from swap transactions with certain counterparties, we performed a mass mailing to our counterparty base. Because of duplicate counterparties, it was extremely embarrassing to hear back from numerous counterparties that the same package of information was mailed to them two and three times"

---

[31] Counterparty is defined as any person or legal entity that performs business with the firm or any of the firm's legal entities.

[32] Interoffice memorandum, From: Molly Mathes, Global Derivatives Credit/Collateral; To: Distribution List; June 17, 1998; Subject: Counterparty Repository Initiatives.

Maintaining the excellent reputation it so far has among its customers and peers is important to the firm. Therefore it knows that even this type of situation of multiple mailings, though seemingly trivial, is something that needs to be remedied.

5.3.3 Costs

There are several types of costs that the firm incurs due to the existence of data heterogeneity.

5.3.3.1 Operational Costs

The first one is the operational costs - the large number of disparate systems, lack of process control and high manual data entry and maintenance has resulted in exceptionally high operational costs relative to competitors.

In a survey done at the start of EDSI, respondents indicated that they spend anywhere from 10 to 50 percent of their time 'cleaning up' data when they need to perform aggregation or consolidation of data.

Having disparate systems also takes it toll when new systems are being built – the new system has to deal with the different data standards of each system it has to communicate with. This increases development costs, reduces productivity (employees are spending more time worrying about how the systems are going to understand each other instead of trying optimize the performance of the system being built) and reduces employee moral.

5.3.3.2 Cost of regulatory misconduct

The firm is anticipating the upcoming regulatory changes and realizes that it could potentially face heave fines in the event of being unable to file the correct reports with authorities. As we pointed out earlier, the firm believes that they may have issues in meeting at least parts of these regulations. The potential fines present a considerable cost to the firm. Coming under regulatory radar once for a mis-filing could increase the chances of unnecessary future scrutiny. Therefore the firm is looking to EDSI to preempt these potential costs of regulatory misconduct.

5.3.3.3 Cost of missed opportunities

This is the price the firm pays by not being able to analyze its data quickly and accurately enough to identify and make use of new market opportunities. Though it is difficult to accurately estimate this figure, it is clear as we discussed in Chapter 2 that the FS sector is more of an information business than ever before. Therefore, the players with the ability to manipulate and analyze the most amounts of data in the shortest time have the capability to offer new products to their clients, thus increasing revenues. The firm knows that it needs this ability if it expects to maintain its existing lead in the FS sector and it is hoped that EDSI will help.

Another type of missed opportunity is the inability to obtain the benefits of Straight Through Processing (STP). Standard Foreign Exchange (FX) transactions are done in the millions of dollars among large FS firms. The loss of overnight interest that is experienced due to processing delays can add up to be considerable. Therefore the firm hopes that EDSI will help increase the percentage of transactions that will experience STP.

As important as the avoidance of such situations as the ones mentioned earlier (such as the Asian economic downturn and the multiple counterparty example sited above), EDSI is driven by another key factor. That is the understanding that the whole face of banking is changing and that it is turning into a truly information based business. The firm depends heavily upon accurate and timely information to compete and stay ahead in the market. The firm's business plan is a simple one - to take its capital, combine it with raw data, and add to it expertise and intelligence, and produce value added information products and services for its clients. Unfortunately over the years, it has neglected its dependency on information. Today, like most of the FS firms, it is in an Information Business without information management. Luckily this firm is farsighted enough to realize the FS sector's increasing dependence on data. There for it started EDSI as part of its effort to bring proper information management practices into the organization and to enable easy data access, as much as to improve efficiencies and cut costs.

5.4 The EDSI approach

5.4.1 Goal and purpose of the EDSI

The internally published mission statement for EDSI reads:
"*EDSI's mission is to provide the necessary architecture, databases and governance structure to support the creation, use and distribution of information across the entire enterprise*".

The ultimate goal is to be able to develop and information architecture with consistent content and standard messaging so that data from any system anywhere can be understood by every system everywhere.

5.4.2 Scope of the Effort

Start Small, then expand enterprise wide.
An enterprise wide standardization initiative is a colossal effort. Therefore the firm has selected several critical areas of the business that has the chance of seeing immediate benefits from the EDSI. These business areas will be the first customers of the newly connected, standardized systems.

At present a pilot project has been undertaken. This will serve as a proof of concept for the overall approach as well provided the immediate benefits to the selected customers. Depending on the success of the pilot phase, EDSI standards will be extended to all divisions.

Customers
One of the (if not the most) important clients selected for the pilot is Corporate Risk Management (we will refer to them as Risk). This is the business unit that is in charge of performing the necessary analysis to accurately asses the firm's financial exposures. It is hoped that EDSI will enable Risk to avoid most of the problems highlighted above by allowing the aggregation of data in *various ways* that were previously not possible. Furthermore, access to *various levels* of aggregated data will facilitate Risk's ability to calculate exposure under different scenarios using financial modes.

Similar to Risk, the following are some of the other business units that will eventually benefit from EDSI:

- Sales - Improved data standards will enable sales to better calculate collateral and margin requirements and better manage client relationships.

- Legal - Will provide the building blocks for a better legal master repository. This will enable them to identify the detailed legal nuances of each transaction and each client, thus making reporting procedures more streamlined.
- Credit - Will enable the Credit unit to accurately allocate credit to the firm's customers

All the clients fall within the firm's Corporate & Institutional Client Group area - in other words these are key areas that support the Investment Banking side of the firm (as opposed to Asset Management, Private Client etc.)

Within the next sections, we will use Corporate Risk Management for examples that explain some of the functionality of EDSI.

5.4.3 EDSI's overall approach to standardization

There are many approaches that can be taken towards achieving the system interconnectivity and standardization (we will review some of the theory related to this in a section that follows). The firm's approach has revolved around the following three core elements:

- Establish Core Data Repositories - identify key data elements, bring them to one centralized source and use the centralized source as a reference when ever this data is needed
- Create a common language among systems - establish standard definitions and values for all shared data
- Create a standard infrastructure - develop the infrastructure (organizational and architectural) needed to support these standards

5.4.4 Establishing Core Data Repositories

The following three databases are three of the key EDSI data repositories.

1. The Product Master Environment (PME) database - PME is scheduled to be the firm's principle source of securities data, containing a rich cross reference to all street/industry identifiers (such as CUSIP, ISIN, SEDOL etc.)
2. Common Party Entity Repository (CoPeR) - CoPeR is scheduled to be the firm's principal source for client and counterparty data. CoPeR will feed downstream databases and will contain substantial public information sourced from Dun & Bradstreet, including legal hierarchies, branches, subsidiaries, key employees etc.
3. RiskMap - RiskMap is scheduled to be the firm's principal source of data on internal trading books. It is planned to incorporate the trading books into CoPeR for consistency.

5.4.5 Establishing a Common Language

Here, EDSI is attempting to establish data standards so that all systems that communicate with each other are able to understand the data they receive, send and process. Quite literally thinking about this as a common "language", the firm envisions each communication about a trading activity that is done between machines to contain a subject, verb, direct object and a verb modifier.

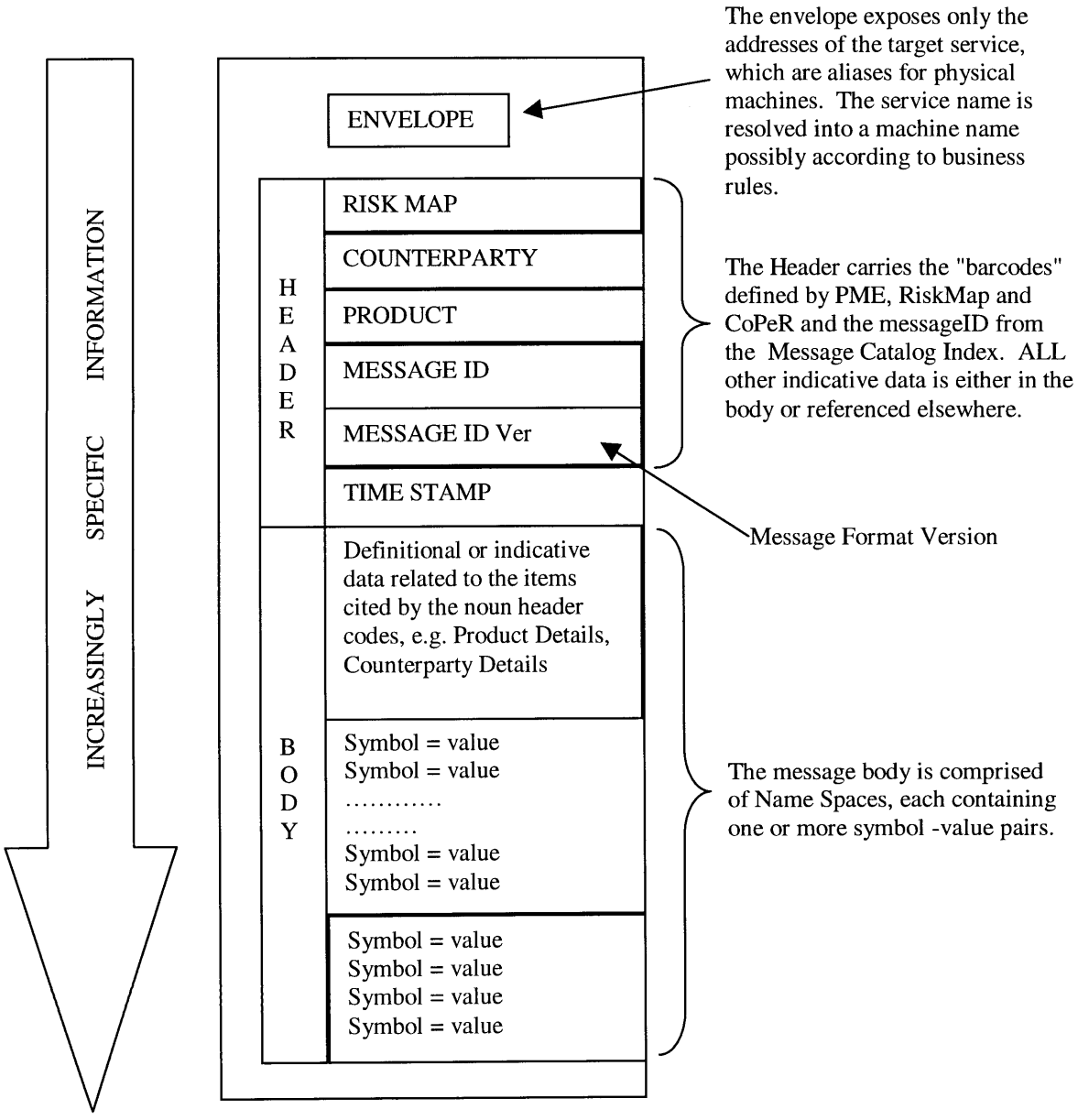
Language Element	Subject	Verb	Direct Object	Verb Modifier
Content	Internal Party (e.g. which trader or trading desk initiates the deal)	Action (e.g. buy, sell, transfer etc.)	Product/Service (e.g. which security is being traded)	Counterparty (e.g. who is the client)
Core database used	RiskMap		PME	CoPeR

With this kind of "language" common to each system and each message that is communicated between systems, it is hoped that all machines (systems) can understand trading activity without confusion - because the system will always know who initiated the trade, whether it's a buy or sell, what product is being traded and who the product is being traded to/from.

5.4.6 Creating the Standard Infrastructure

Once a common language, and standardized data sources are created, the firm needs a method to have the information travel between systems. For this, EDSI proposes a messaging format as follows:

Figure 6: Proposed Trade Message Format



5.4.7 How it all comes together

In this section we will present an example of how each of the key databases, messages and systems will come together to create the new, standardized environment envisioned by EDSI.

Suppose a trader obtains a new institutional client. Let this client be "MIT Lincoln Lab". Upon finding their investment needs, risk profile and other information, it is necessary that the trader add MIT Lincoln lab into the firm's systems and start executing trades on behalf of Lincoln Lab.

With the new EDSI standards and processes, the basic steps will be as follows:

1. The trader initiates a request for new a CoPeR ID for the new client, MIT Lincoln Lab. When doing so, he will provide as much information as he can (or has) to CoPeR about the new client. For example, it may be likely that he mentions that the new client is a subsidiary of MIT that is a separate legal entity.
2. The trading system will refer to the CoPeR database (described above) for a unique ID.
3. If MIT Lincoln Lab doesn't exist within CoPeR, a new CoPeR ID has to be issued. Before doing so, a check will be performed to see if MIT is already a client of the firm. If it is, a hierarchically linked CoPeR ID will be issued to MIT, so that in the future, the system knows that MIT and MIT Lincoln Labs are connected in a way that is specified. If the trader didn't provide any information about MIT/MIT Lincoln Lab connection, the personnel in charge of CoPeR will perform a check nevertheless, sometimes even calling up Lincoln Lab and running through a questionnaire about its ownership, subsidies etc. If Lincoln Lab is already a client, the trader will be notified of the existing CoPeR ID, and he can proceed with the trade.
4. The trader can now execute trades on behalf of the new client MIT Lincoln Lab. Each time a trade is made, the particular trading desk the trade is being made from will be captured. Therefore it is easy to calculate each trader's commissions and performance. Each trade will also force the trader to choose a product from the PME database. This way, the trader chooses a product from a standardized list of available products, and avoids having to enter information in the "comments" field of a trade message.
5. Once all the information for the trade is encapsulated, the trading system will refer to library of standard messages to identify which type of message is appropriate, given the type of product and trade that is being performed. If this is a completely new type of message that is required, the new message format will be created and it will be stored for future use. It will also refer to the standardized business events to obtain the routing information for the message.
6. The message will be routed to the appropriate systems. The message may pass through several systems before its execution is complete. If communication with external parties is needed (perhaps to transfer payments to the client), appropriate messaging (such as SWIFT) will be used. The translation between internal messaging formats and external messaging formats have been defined early in the EDSI process.
7. If credit checks are needed to complete the trade, the appropriate centralized Credit database/Credit-processing system will be referenced. Since the counterparty identification numbers are standardized, it is now clear that MIT Lincoln Lab is a subsidiary of MIT. Therefore accurate estimates of credit can be performed immediately, unlike before. But more complex structures may need to be defined by the firm for cases when the company in question is a wholly owned subsidiary of another firm vs. a minority owned subsidiary. The firm also needs to define such things as how it will determine if 51% means "majority" shareholder or not (often this varies by country).
8. Once a trade is completed, RiskMap will be updated to reflect the changes. Now, at any moment the firm can refer to RiskMap if it wants to know its exposure to MIT or MIT Lincoln Labs. It can also use RiskMap at the end of the month to find out how much commission needs to be paid to trader Joe who handles the MIT Lincoln Lab account. It can also cross-reference between

RiskMap and PME to find out how much exposure it has to Municipal Bonds issued by the Massachusetts (and bought on behalf of MIT Lincoln Lab).

5.4.8 Measures of success

At the end of the pilot, the EDSI hopes to be able to establish (or evaluate) the following:

- Validate the concept of Information Authoritative sources (PME, CoPeR etc.)
- Validate the concept of information "bar codes" (show that the integrity of the unique data identifiers for Product, Counterparty and Organizational Data can be maintained consistently throughout the entire data flow)
- Validate the proposed information architecture - show that the proposed architecture satisfies all the information sharing requirements between systems and that the proposed messaging satisfies downstream data requirements. Most important is demonstrating that the model-based approach to information architecture yields a framework that satisfies the data consolidation requirements of the business drivers mentioned above.
- Validate scenario generation architecture that Risk requires

5.5 An evaluation of the firm's approach, decision and process

In this section we will look at some existing theories/ideas on data management and analyze the firm's approach with reference to them.

5.5.1 Some Data Management Approaches

An organization can take many approaches to data management and the existing literature presents many ways of thinking about data management. Goodhue, Quillard and Rockart [Goodhue90] collect these various approaches and categorize them as follows:

- **Approaches with a technical focus.** These include tools and techniques such as database management systems, data dictionaries, data entity relationship modeling
- **Approaches with a focus on organizational responsibilities.** These include the establishment of organizational units such as database administration and data administration, and the formulation of administrative policies and procedures covering areas such as data ownership, access and security.
- **Approaches with a focus on business-related planning.** These include planning processes and methods such as Strategic Data Planning that link the acquisition and use of data with business objectives.

There is no evidence that one category provides a completely adequate approach. But the literature does provide empirical evidence suggested that most data administration groups have had little or no success in correcting key data management problems. For example, [Coulson82, pp 6] acknowledges that many efforts to straighten out data management problems thorough the implementation of a data dictionary have failed. Because the ultimate goal is not to put in place tools or to create organizational units, but rather to link data need with the needs of the business, much attention has focused on the third type of approach. These planning approaches however, require significant resource commitments and are often not easy to undertake.

5.5.2 Thinking about data management at the organization level

Goodhue/Quillard/Rockart contend that there are several "paths" to improving the management of data. Which path is selected depends heavily on organizational considerations. Successful efforts studied in the literature are very diverse in business motivation, organizational scope, planning method and type of result obtained. The four key areas presented below are a set of choices that organizations make as they undertake a data management effort.

- The identification of a **business objective**. In the successful companies, data management actions are almost always justified not by conceptual or technical arguments, but by one for four compelling business needs: coordination, flexibility, improved managerial information, efficiency
- The **scope** of the data management project. The successful firms explicitly define and limit the scope of their efforts. Some focus on a functional area (such as finance), others on a division, while some are corporate-wide efforts.
- The **data planning** method. Top-down, in-depth strategic data modeling is not the only data planning process observed. In fact, there are a number of obstacles in accomplishing a large-scale strategic data planning effort. The planning processes vary widely in terms of their formality, detail and emphasis on data models. The rage of options varies from strategic data modeling, to more limited planning approaches, to no planning whatsoever.
- The **"product"** of the data management effort. Much of the data management literature centers on the implementation of subject area databases. But five distinct "products", which were the end results of the data management project teams' work are observed. These products are: subject area databases for operational systems, common systems, information databases, data access services and architectural foundations for future systems.

In the following sections we will examine each of the key areas (product, data planning, scope, business objective) – first as they are presented in the literature and second as they apply (or not apply) to EDSI.

5.5.3 Data Management Products

The most common "product" in existing literature is a set of subject area databases used by multiple operational systems. There are also other products such as common systems, information databases, data access services and architectural foundations for the future. This section discusses each of these products.

5.5.3.1 Subject Area Databases

Subject area databases contain data, which is organized around important business entities or subject areas, such as customers and products, rather than around individual applications such as order processing or manufacturing scheduling. Many different operational applications may share (both access and update) data from a single set of Subject Area Databases. In the realm of data management, creating Subject Area Databases has been a very common approach.

EDSI has taken steps very similar to this. The PME (Product Master Environment) and CoPeR (Common Party Repository) are two key databases that contain subject specific knowledge about financial products and organizations/clients respectively. EDSI has chosen to eliminate redundant/incorrect data collections that previously existed in multiple locations about each of these subjects. These databases will serve as the one central source for each subject from now on. A special EDSI group will be set up to 'own' these databases and to maintain their quality.

5.5.3.2 Common Systems

A second type of data management "product" is the operational data files or databases which are developed for common systems. Common systems are applications developed by a single, most often central, organization to be used by multiple organizational units. Physically, there can be one or multiple copies of the system. As opposed to the subject area databases discussed above (which are developed to be shared by a range of applications), a common system approach tends to focus on replacing existing redundant systems in one specific application area. Many firms already have common systems in place, often developed not for data management purposes, but rather to ensure common procedures or to lower IT/IS costs.

We see numerous examples of Common Systems in the EDSI project – examples are the front office trading systems (RAM, MTS), middle office systems (GMAC, GEDT, Risk) and back office systems (for transaction processing, clearing & settlement etc). But major re-designs of these systems are not part of EDSI at this moment other than to ensure that these systems will be able to "understand" the newly formatted trade messages and that they comply with the EDSI defined data standards and processes.

5.5.3.3 Information Databases

Most Information Databases can be defined as Subject Area Databases for managerial information. Whereas the first two products provide data to be used by transaction processing systems and for monitoring real-time operations, Information Databases are aggregations of data, which are primarily used for staff analysis and line management information.

An example of this is IPS (Institutional Profile System), a database used to record data of transactions made by traders for each of their clients. This database is used by management primarily for the purpose of calculating compensation – the trader gets paid a percentage of the "production credit" (the amount of revenue he has generated from a client). Historically this system has carried a large amount of redundant data. Previous to EDSI it was impossible to identify which trader initiated each trade for which client. Therefore the traders depended on IPS to record this information. As such, was not uncommon to find more than one "account" for IBM, opened by several different traders. In their standardization effort, EDSI hopes to streamline this system because with the new messaging format described earlier, it is possible to identify which trader initiated each trade. But there is resistance by the traders to the replacement of IPS - since EDSI is still an unproven concept, they are unwilling to exchange the trusted and true system for something they think will eliminate their ability to differentiate their trading ability.

5.5.3.4 Data Access Services

The first three products discussed emphasize the development of databases with pertinent, accurate and consistent data. But managerial access to existing data, even without attempting to upgrade the quality or structure of the data, has been identified as important. Though the particulars of the efforts differ, many firms centers this effort around a small cadre of personnel whose goal is to better understand what data is available in current systems and to put in place mechanisms to deliver this data. These efforts seem to be widely applauded by managers who are finally able to "get their hands on" existing data. They also have the promise of surfacing questions concerning definitional inconsistencies as managers attempt to make use of the data provided. But not surprisingly, Data Access Services appear more helpful in companies where data is of reasonable quality than where the data is a mess.

EDSI is very much aware of the need to build in the correct organizational support structure to ease the access to data. At a fundamental level, the need to be able to correctly analyze client exposure, foreign exchange exposure and other concerns, are all a part of management support, and will be hopefully be achieved by EDSI. However, management support/management information systems are not a part of the study for this thesis.

5.5.3.5 Architectural Foundations for the Future

Most firms tend to focus on a limited set of data serving a portion of the corporation. However, there clearly is a danger that by approaching data management in a function by function, business unit by business unit, or subject area by subject area manner, a company leaves itself open to real problems if, in the future it desires to integrate data across these boundaries.

To attempt to avoid these future incompatibility problems, some organizations have focused on developing architectural foundations. By architectural foundations, the literature means policies and standards which, when adhered to, will lead to a well-structured and consistent data environment. Managers allocating resources for architectural foundations are investing in the future without necessarily having immediate benefits in mind.

There are two different types of architectural foundations that need to be emphasized: (1) wide scope strategic data models, and (2) common data definitions.

One view of a firm's data architecture is a corporate-wide strategic data model to serve as an underlying blueprint for all future systems development. Many newer derivations of IBM's BSP (Business Systems Planning)[33] methodology, James Martin's Strategic Data Modeling Approach, and others have been common methodologies that have been adopted by companies. Proponents of these approaches claim that a strategic data model provides an architectural foundation that will lead to consistency of information, more easily integratable systems, and improved productivity in system development and maintenance.

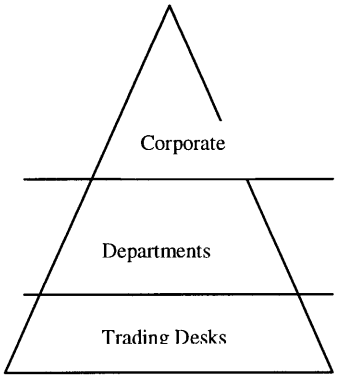
The second approach to data architecture is the standardization of data definitions. The choice of which data elements should have corporate wide standard data definitions is an important architectural issue. In any firm there are some data elements which are so basic to the operations of the business and which are the basis of so much shared communication, that it is critical for all parts of the organization to refer to these elements in the same way. Presumably these data elements should be given global, mandatory definitions. Below corporate level it may make sense for a particular divisions to standardize on certain additional data elements, just within that divisions. Thus standardization of data definitions can be seen as a hierarchical process.

This second approach (of creating data definitions) has been clearly adopted by the EDSI team. It has taken a hierarchical view of the data when attempting to standardize it.

---

[33] http://oz.plymouth.edu/~harding/ibmbsp.html and http://oz.plymouth.edu/~harding/bsp.html provide a good summary of BSP. Or refer to Business Systems Planning, IBM Manual #GE20-0527-3, July 1981

Figure 7: Three Levels of Data at the firm



Level 1- Generic Data such as Counterparty (along with the n1 key elements that define it) and Product (along with n2 key elements that define it)

Level 2 - Data (and processes) consistent within a department

Level 3- Data (and process) consistent within a trading desk

In other worlds, EDSI has identified that (for example) counterparty and product information is fundamental to the daily operations of the firm. Therefore, corporate wide standards have been established for them. Each system, business unit and group must conform to these published standards. On the other hand, EDSI realizes that there are many data items that are very specific to a particular business unit, department or even a trading desk (for example, some desks trade very specialized financial products that contain detailed levels of information that are not used or not needed by any other trading desk). Such variations are allowed, as long as overall conformity is met.

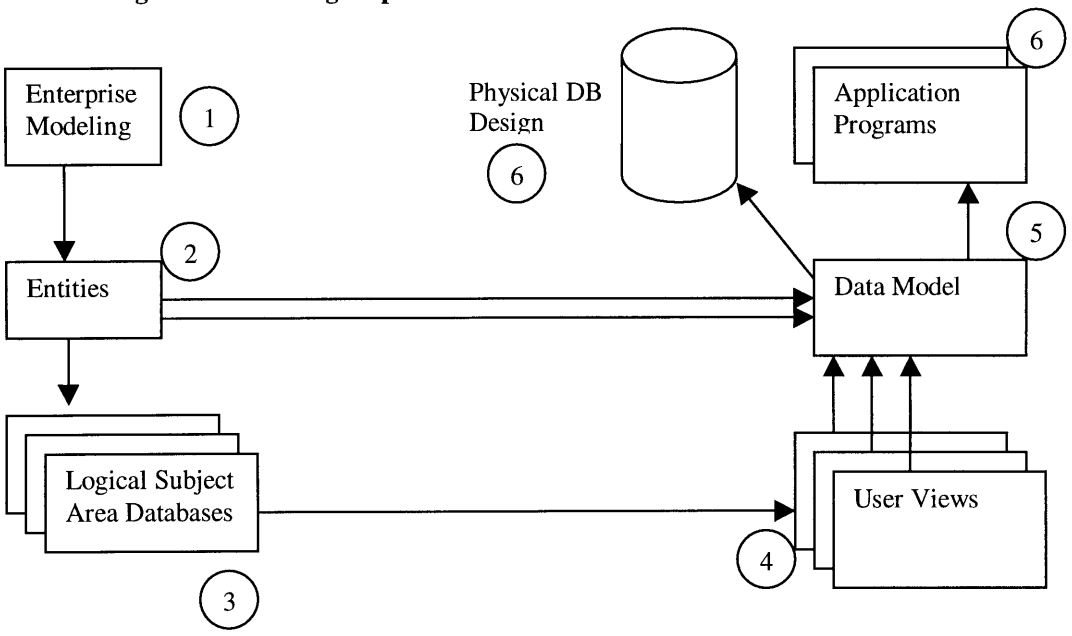
5.5.3.6 Data Planning Processes

Data planning has often been synonymous with large-scale strategic data planning and modeling. There are however other less all-encompassing planning approaches which can be equally or more effective. Goodhue, Quillard, Rockart categorized the processes they observed into four types: the strategic data modeling, targeting, and 80-20 approaches, and no explicit planning. These are planning processes that assist the organization to identify the target for data management action and to chose the action (or "product" discussed earlier) to pursue. The following sections summarize each approach very briefly:

1. Strategic Data Modeling:

The following diagram [Goodhue90, page 17] presents a general overview of the data-oriented strategic planning approach.

Figure 8: Strategic Data Planning Steps



The left side of the diagram shows the top-down planning approach, leading to the bottom-up design shown on the right side of the diagram.

The process begins with the development of an enterprise or business model (box 1). The enterprise model depicts the functional areas of the firm, and the processes that are necessary to run the business. The next step is to identify corporate data entities and to link them to processes or activities (box 2). Data requirements are thus mapped onto the enterprise model, leading to the identification of subject areas for which databases need to be implemented (box 3).

In general only selected portions of the enterprise model and subject area databases are selected for bottom-up design. Building the logical data model is the first step. The data model (box 5) results from a synthesis of detailed end-user and management data views (box 4), with the results of the previous top-down entity analysis (box 2). Database design and subsequent design of application programs (box 6) proceed from the logical data model.

The underlying assumption of the strategic data planning methodologies – that it is impossible to plan effectively if one does not know what the business is, what it does and what data is uses – is difficult to contest. However few firms chose this approach and even fewer have succeeded in implementing this model in actual systems.

2. Targeting High Impact Areas

Most corporations that skip or abbreviate the top-down planning methodology do not act without a plan. There are a variety of alternatives. The most common process is the "targeting" of a particular function or other business area. In some companies, important problem or opportunity areas are quite clear without an extensive analysis.

3. 80/20 Planning Approach

Sometimes there is a desire to get the major benefits of a global data planning without having to invest the amounts of time and money necessary to carry out a full-scale strategic data planning process. The aim in these cases is to zero in quickly on the key "products" to be implemented

(bottom-up), while reducing the amount of effort spent in a global planning (top-down) phase. This type of planning is termed an "80/20" approach, after the adage that for many undertakings, 80 percent of the benefits can be achieved with 20 percent of the total work.

4. No Explicit Planning

There are also data management actions that can be taken without doing any data oriented planning. For example if a decision is made to provide better access to existing data, without addressing changes in the form of that data, then no data planning methodology is needed.

Of the above approaches, EDSI's approach is closes to that of approaches 2 and 3 (Targeting and 80/20). It resembles targeting because the Risk Management function has been very clearly identified (for reasons explained earlier) as a high impact area that could benefit from the data management efforts. On the other hand, even after selecting the Risk Management function, EDSI has taken a 80/20 approach in deciding on which specific data to standardize and clean – the customer information, product information and the risk database are the three key products chosen by EDSI because the cleaning and standardization of these three products are believed to solve a majority of the data problems the firm experiences today.

5.5.3.7 Bounded Scope

Almost no firm attempts to manage all the data used by the corporation. The focus the data management is limited in one or more ways. An important factor in the success of a data management effort is that the scope of the effort be carefully selected.
Some firms focus on a functional area while others start divisional efforts. Others undertake corporate data planning efforts, but keep the focus of the effort to a small set of standard data definitions.

This approach is certainly true of the firm – as mentioned above, even though EDSI is a corporate level effort, the Risk Management area has been selected as the immediate target. More over, as shown in Figure 7: Three Levels of Data at the firm, they are attempting to standardize only key data elements, leaving the rest to individual functions/business units.

5.5.3.8 Business Objectives

The successful data management processes have been aimed, for the most part, at solving a clear and specific business problem or exploiting an opportunity. The kinds of business problems and opportunities that motivate executives to consider more intensive management of data resources are:

- Coordination within or across business units: the perceived need for better coordination often drives the development of common data definitions
- Organizational flexibility: the desire for greater organizational flexibility to allow either an internal restructuring of the organization or a refocusing of the organization due to changes in the external environment (market place, regulatory environment etc).
- Improved information for managers: improved access to data and improved data quality is key to management's ability to analyze organizational data
- IT/IS efficiency: Improved data management is considered to be a way of addressing the need to increase development productivity and reducing maintenance costs.

Several of the above business objectives play a somewhat indirect role in the motivation for EDSI. As discussed earlier in this chapter, EDSI is driven by the need to meet regulatory standards, the need to better manage client exposure and relationships and the need to limit costs. But we can see that

achieving the above business objectives mentioned in the literature (such as organizational flexibility, IT/IS efficiency), the firm will be able to achieve it's stated goals of meeting regulatory standards, managing client exposure and limiting costs.

5.6 Critical factors that will determine the success of EDSI

There is a substantial body of literature on information systems implementation. Out of this, Goodhue, DeLong, Rockart [Goodhue92] have identified a list of critical factors that will determine the success of standardization efforts. Most of these are summarized in [Bader99] - in this section we will selectively use some of the important propositions made by [Bader99] to analyze briefly the health of EDSI.

- ***Top management must perceive data integration as critical to the strategic goals of the organization:*** So far, evidence at the firm suggests that there is considerable management buy-in for EDSI, at least for pilot phase. The fact EDSI is driven by the very real business need of meeting regulatory compliance and improving customer relationships and risk is some evidence that the business sees the value of EDSI. The other evidence comes (indirectly) by the fact that EDSI is charged as business expense, instead of a "technology" expense. This is indication that the business side of the organization is at least perceives EDSI to be a valuable effort. The fact that EDSI first undertook a pilot of the effort is also commendable – this not only makes it easier to obtain management buy in, but also allows EDSI to evolve if it is evident that business needs are not being met. We should note that this is by no means the first time that such large standardization efforts have been undertaken at the firm. Therefore the EDSI team needs to constantly evaluate the viability of their project.
- ***Alternative approaches should be considered for developing an integrated architecture:*** It is not clear what alternatives were considered when deciding on the approaches and methods used in EDSI. An exploration of newer technologies may have been suitable during the initial planning stages, especially since the current approach has some (however small) resemblance to past efforts such as the PRI (Party Repository Initiative). In the next chapter we will discuss one alternative that may be able to meet some of the potential limitations of EDSI.
- ***Architectures must be enforced in order to have an effect on data integration:*** It is clear that the EDSI team realizes the importance of enforcing their standards. Towards this end, they have taken steps to publish clear guidelines about EDSI, the goals, and standards through the organization. But more importantly they have started developing new business functions – groups of employees that will serve as reference points and guides on the Architecture, Vocabulary and Data Services.
- ***Imitation is an efficient way of devising architectures rather than reinventing them:*** The fact that the firm is willing to learn and use existing best practices is evidenced in the fact that they have examined existing architecture and standardization principles (both in industry as well as academia). They are also aware that they should use industry-wide, accepted standards and tools wherever possible. In other words, "use XML instead of creating a firm-ML"
- ***Do not spend too much time bringing novice data modelers up to the learning curve:*** EDSI seems to have gathered an experienced team of veterans to lead the EDSI effort. Therefore it can be perceived with some certainty that they will avoid the problems some organizations faced by having to bring everyone up-to-speed.
- ***Individuals time requirements may add as a screening device that selects less desirable participants:*** So far this doesn't seem be a problem with EDSI – at least at the highest levels of leadership, EDSI is the full time job function for set of highly qualified individuals. But unless

the individual business units are convinced of EDSI's value, it is likely that they will assign to EDSI the people within their units that are 'unimportant' (and hence can be spared) .

- ***New knowledge about the purpose of data integration initiatives is difficult to be diffused within the organization:*** As evidenced by the numerous brochures and books available in bulk in each office, it appears that the EDSI leaders are making a concerted effort to communicate the purpose of the effort to the organization. Even during interviews, they indicate that the major part of their day is spent "marketing EDSI" to the business. It certainly looks as if EDSI is take the right steps, instead of assuming that every one will be automatically be convinced of the virtues of standardization just because they are required to comply with the standards.
- ***Education and communications alone can probably not justify the cost of a data integration effort:*** Here we see that there is much room for improvement. As mentioned in the above bullet point, EDSI is taking definitive steps to educate the enterprise to the virtues of EDSI. But we believe that they are not facing "the tough questions" because right now the initiative is being driven by a business unit (Risk) that very clearly has a viable business need for the standards. But these standards (or the requirement to comply with these standards) will soon start effecting other business units – and they will have to pay the cost of having to be EDSI-compliant. At that time, we believe that EDSI will need to clearly elaborate the benefits (in dollar terms) of compliance, in order to justify the costs these other business units will incur. So far we do not see a clear case being made for this – perhaps not because EDSI is unwilling, but because estimating such benefits/costs is truly difficult. For example, how can one estimate the cost to the firm that arises due to missed opportunity – the inability to react to a market need that was caused due to the existence of data heterogeneity? But as pointed out earlier, there are some costs that CAN be estimated – for example, the EDSI team can put together the total amount of fines the firm paid to regulators because legal/regulatory reporting standards were not met. Unless actions similar to this are taken, EDSI will face an uphill battle in taking the pilot project enterprise-wide.

In summary, we see that EDSI is taking best practices from theory, other firms and industry to engage in a well-thought out data planning effort. While it is doing many things "right", we believe that there is still room for improvement. Some potential problems (such as business units' unwillingness to accept and pay for EDSI) were highlighted above as we discussed the critical factors that would determine the success of EDSI. But one area of particular concern is the need to have flexibility to meet future business needs. At present, the EDSI has identified key data elements and standardized them. But in the future if the firm decides to acquire another FS firm, or decides to merge with one, a newer subset of 'key data elements' may be involved. Or if regulations change again and authorities want different reporting formats, newer levels and types of data aggregation may be needed. When such situations arise, the standards will need to be changed. Therefore EDSI, like many other standardization efforts, faces the challenge of keeping it standards evolving in order to match the evolving business needs of the firm.

6 Alternatives to standardization

6.1 Why alternatives are needed

We saw above in Chapter 3 that there are certain kinds of data heterogeneity issues that aren't easily solved with standardization. In particular, the subtle differences in data (such as the question of whether or not sales tax is included in a quoted price for a book and whether or not the price is expressed in US Dollars or UK Pounds) are very hard to solve by establishing standards. Therefore the need for alternative methods becomes important.

The inability to deal with certain types of data heterogeneity is not the only problem with standards. One of the biggest drawbacks of standards is its inability to respond easily to the changing business and information environment. For example, a company can go through a expensive yet successful standardization effort as long as it succeeds in creating and enforcing good standards. As long each user agrees with the set standards, there will not be problems. But what happens when the company buys another company? Or mergers with one? The good standards that worked previously may not work with equally good standards established by the other firm. Or the other firm may not have standards at all. In any case, all established standards may need to be re-examined. This usually leads to yet another standardization attempt. That is, the evolution of the business lifecycle leads to more standardization activities for a firm that is interested in establishing standards. This can be a continuous and large expense as companies evolve continuously.

Therefore, research has been done to find a "better" or at least alternative set of solutions problems like the above. Apart from facilitating the better access to day, these projects focus on the important issue of 'understanding' the 'real meaning' of data. In other words, they attempt to deal with *context*. A context is the collection of implicit assumptions about the context definition (i.e. meaning) and context characteristics (i.e. quality) of the information. The problem is the existence of heterogeneous contexts, whereby each source of information and potential receiver of that information may operate with a different context. Under these conditions, the types of problems mentioned above are created and large-scale semantic heterogeneity occurs. Therefore the capability to reconcile semantic conflicts among sources and receivers (called "semantic interoperability") is needed among data sources.

In the following section we provide a brief overview of a project at MIT's Sloan School of Management that is attempting to provide a better semantic integration of disparate data sources.

6.2 COIN (COntext INterchange Project) at MIT[34]

With advances in technology, in particular the widespread use of the Internet, the number of data sources has increased exponentially. The advances in technology have helped to connect these systems together. New data access methods have eased access to them to such an extent that the distinction between a locally stored data sources (traditional databases) and data accessed over the Internet is disappearing, at least in the eyes of the use. He or she wants seamless access to all data sources and expects each system to understand requests stated in their own terms, using their own concepts of how the word is structured. In other words, the use wants the data in his or her context.

[34] This section primarily based on [Bressen97] and [Madnick97]

In order to provide this type of functionality, architecture with the following types of features is needed:

- A way to access remote and local data sources with ease
- A way to understand/interpret the context of this data
- A way to understand/interpret the context of the user's data query
- A way to resolve differences in contexts/semantics

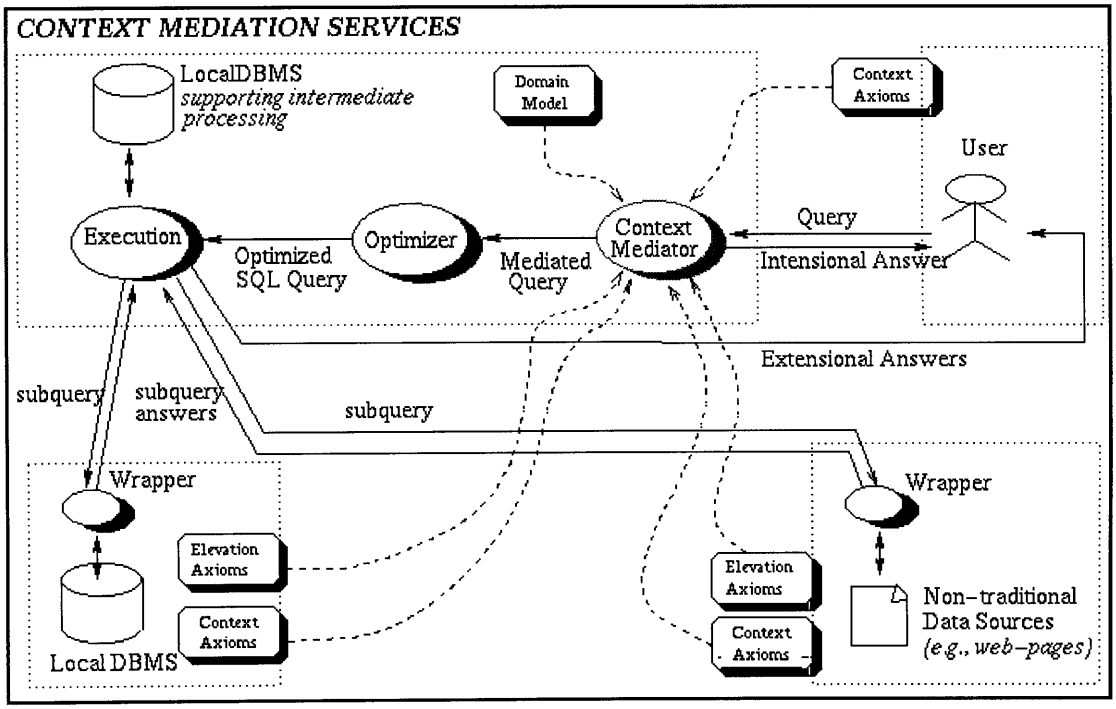
COIN addresses the need for semantic interoperability by providing a Context Mediator that detects and reconciles semantic conflicts among distributed data sources and receivers. It does this through reasoning about the contexts associated with systems engaged in the data exchange.

6.2.1 The COIN Architecture

We will use the following very simple (but useful, and often used) example to explore the features of COIN.

Suppose a the travel agent for a small business is looking up hotel fairs in a city - let's pick Paris, France to be precise. The user may already have a collection of Paris hotels, relevant prices etc. in a database that is situated in his office. But he may also prefer to search the Internet for hotel prices. He submits a "question" or a "search" in order to find this information. There are many formats, graphical user interfaces, browsers that he can use in order to do this search. Let us see what happens to this query in the context of COIN, using Figure 9: COIN Architecture which is a diagram of the basic components (architecture).

Figure 9: COIN Architecture[35]



[35] Available at http://context.mit.edu/~coin/description/coin_model.html

6.2.1.1 Data sources and wrappers

Let us suppose that the two data sources available are the Local DBMS (on the bottom left-hand side of the figure) and the web page ("non-traditional data source") on the bottom right hand side. These represent the two general classes of information sources - structured data sources such as traditional DBMS and on-line information services that are often either semi-structured pre-structured or unstructured.

Semi-structured format of web pages includes information presented in tables, lists, trees or other structuring organizations for which the structure is not fully known in advance and must be analyzed on the fly during the search/query process in order to locate the data. Most stock exchange quote services offer semi-structured format data. If we are lucky, the web page will be of a pre-structured format. This is the case when a page uses data representation compliant with a standard such as the Open financial Connectivity standard. In the worst case, the web data sources will be unstructured plain text.

No matter what type of data source, COIN wrapping technology is able to take advantage of the Hypertext structure of web sources and of underlying structure provided by HTML. It treats a web service as a collection of static and dynamic pages connected by transitions.

6.2.1.2 Mediation Services

The next key element is the mediation service. The mediator transforms a query written in the terms known to the user or application program (i.e. according to the user's or programmer's assumptions and knowledge) into one or more queries in the terms of the component sources. In other words, the mediator takes the users request for data (which is expressed in the user's context) and converts it into a 'language' that the data source understand (the source context) and extracts the information.

COIN allows queries 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 makes use of a Domain Model and Context Definitions for this purpose

The **Domain Model** is a collection of definitions. In the case our Paris hotel example, it would define notions such as money, room size, single/double bed types etc.

Context Axioms define the different interpretations of the information in the different sources or the receiver's point of view. Going back to our example, the travel agent has a context C1 which assumes money amounts are in US Dollars without and the source has context C2 which expresses money in French Francs and includes taxes. This information is captured in the set of Context Axioms. The mediator re-writes the user's query using the proper currency conversions (using other sources to get information about exchange rates etc).

Using the above type of architecture components described (briefly) above, COIN is able to resolve contextual differences between data sources and help the user extract the information he wants, in the format he expects. First the user's 'question' ("find a hotel in Paris for less than 200 Dollars per day with double beds" etc.) is transformed into a SQL query. The query is optimized (if possible) and executed. During the execution, the query may be broken down into several sub-queries, each of which accesses a different data source. Conflicts between contexts of these data sources are resolved by the Mediator using the domain axioms. Conflicts in context between the retrieved data and the user's view are also resolved in a similar way. For example, currency conversion procedure is used to first convert the quoted prices from Franc's to dollars and then it is compared (to be <$200). Other important issues such as the components of the quoted price (does it or does it not include tax?) are

also resolved. The data is then aggregated (if needed, for example to provide the user a list ordered in ascending order) and presented to the user.

6.2.2 Benefits of COIN

6.2.2.1 Is able to deal with more categories of heterogeneity

We see from the above that COIN, because it is able to deal with contextual issues, is able to deal with a larger subset of heterogeneous data than standardization.
A good example would be the problem of having "Price" interpreted differently for two financial products. But context axioms will specify that 'price' is indicated in millions of US Dollars in one system, while it is in hundreds of Australian dollars in the other. This type of heterogeneity (what we called "heterogeneity due to multiple Interpretations) is not possible with a standardization approach.

6.2.2.2 Is able to deal with internal and external data sources

One problem with standards is the difficulty to enforce them externally. Even if a firm is successful in creating data standards internally, it is not realistic to expect that it can try to convince the 'rest of the world' (i.e. all other data sources) to adopt their standards. With the reliance on information from the Internet and other outside sources, these firms will run into the same problems it was trying to solve in the first place by adopting standards internally.

The advantage of COIN is that it is able to deal with internal or external sources with ease. The wrapping technology allows it to extract information out of HTML based web pages. With the adoption of XML (which is more structured than HTML), COIN will also be able to extract more useful information easily.

6.2.2.3 Can respond to changes quicker

If the structure of the data source changed, COIN it can still extract useful information since it identifies context conflicts on the fly (during the execution) of the query.

Changes in an organization can cause havoc even on the perfect standardized systems. For example, even if two separate FS firms had individually standardized systems, if the two firms merge, much effort will be needed to combine these. Time lost in identifying the specific heterogeneities across the two systems is too critical to business success. But using a COIN based approach, the individual systems don't need to be revamped. Instead the mediator can identify the differences between the systems and make aggregation possible relatively instantaneously.

6.3 How is COIN different from "standardization" ?

As we pointed out in the beginning of this section as well as Chapter 3, the inability to deal with the evolution of standards is one of the major problems of the standardization approach. For example, let us suppose that all the divisions of a FS firm agreed to adhere to the standard that "P/E Ratio" would be calculated as follows:

- P/E Ratio = Price/Earnings and expressed in 10's of US Dollars where:
- Price = Price of a Share as reported in real time via Standard & Poor's Comstock
- Earnings = Earnings per Share for the most recent 12 months, where:
  - Earnings figure is obtained from source X (such as an Annual Report)
  - Number of Shares Outstanding is defined as all outstanding common stock of the company.

- The "market" in this case is the NASDAQ

This type of 'standard' foe the calculation of P/E Ratio will help the firm eliminate some of the data heterogeneity problems that are faced when an analyst wants to compare two companies.

But now suppose the firm buys a regional (West Coast based) brokerage house that specializes in the locally situated, highly successful, small to medium sized businesses. After the purchase of the regional firm, the analysts of the newly formed FS firm need to be able to analyze the set of portfolios that were previously managed by the acquired firm. But it would not be surprising that the regional firm used quite a different interpretation of P/E Ratio, and calculated it as follows:

- P/E Ratio = Price/Earnings still, but is expressed in 1's of US Dollars (instead of *10's of dollars*)
- Price = Price of a share as reported at the close of the market the previous day and (instead of *the real time price*)
- Earnings = Earnings per Share but for the most recent financial year (instead of *the most recent 12 months*)
    - Earnings figures are obtained from source Y (which is *different from source X*)
    - Number of Shares Outstanding is defined as all outstanding common stock + preferred stock (instead of *just common stock*)
    - The "market" in this case is San Francisco Stock Exchange and the 'close of the market' is defined as 4.15 p.m. Pacific Time.

There are clear differences in the way the two firms interpret the P/E Ratio. After the take over, the new firm has the choice of accepting one of the previous interpretations or creating a newer, better interpretation for "P/E Ratio" altogether. No matter which is chosen, the result is that a new standard needs to established. This means a standards project needs to be started, new data elements need to be defined, 'old' systems need to updated so that they 'understand' the new standard etc. Along with this effort comes all the previously discussed problems of standardization - some of which these firms experienced individually anyway when establishing their individual standards for the P/E Ratio. The costs are high, the effort is time consuming, and success is not guaranteed. In the meantime, if a second takeover occurs, things need to be started over again.

This type of situation, which could easily happen given the number of mergers and takeovers that are going on in the FS sector, is exactly where COIN's advantage is highlighted. Because of its use of Context Axioms defined above, the mediator knows that the data sources of the two banks (which are now merged) use different interpretations. For example, it can find out that one uses 1's of US Dollars while the other uses 10's of US Dollars is apparent in the axioms. Therefore, the bank doesn't have to "clean" the data of the bank it just bought. Instead, the mediator is able to perform the necessary conversions depending on how the analyst (the user in this case) wants to view the P/E Ratio figures when comparing his clients. It takes minimal time to do this conversions based on context where as a data standards effort would take months.

This ability to evolve and encompass different standards with the minimal time and cost is one of the crucial areas where architectures such as COIN can surpass standardization as a method of dealing with heterogeneity.

Even in the FS firm we studied in Chapter 5, we saw that the problem of evolving-standards (due to evolving business needs) existed. For example, the PME (Product Master Environment) standardizes the identification of all products that are traded by the firm. The attributes of each product contain such industry references as SEDOL and CUSIP numbers. But suppose that the CUSIP committee decides to change its definition of CUSIP - it now increases it into 15 digits instead of 9, splits the CUSIP into 4 parts instead of 3 and defines new algorithms to identify each part. The firm will have

to re-define its data structures (and standards) in order to be able to deal with this change - in other words, all 'old' CUSIP's will need to be converted to the new format. But if a COIN-like approach is followed, the notion of CUSIP (including about the above-mentioned changes) can be contained in the Domain Model. Therefore, 'old' CUSIPs and new CUSIPs can be left alone in the numerous databases without re-work. The mediator will interpret as appropriate when data from the heterogeneous sources need to be accessed.

7 Conclusions

Through our discussion in this thesis we have established first that the ability to access, aggregate and analyze data is vital to the survival of most FS firms. But a major problem that is faced by these firms is the inability to 'understand' the data once it is extracted. The difficulty in data aggregation occurs due to heterogeneity that is introduced into the data for various reasons. We grouped these heterogeneity problems into several categories in order to be able to study them and to identify broad solutions.

Heterogeneity is introduced into a firms data sources due to various reasons. Not all of them are due to the 'bad' management of data. We saw the natural evolution of businesses as well as the numerous regulatory agencies cause a firm to develop legitimate data heterogeneity among its systems. But we saw also that the ability to deal with these heterogeneous data sources in order to access, analyze and aggregate the data in these sources is vital to the survival of a FS firm. Therefore some kind of method is needed to deal with heterogeneity.

We saw that "standardization" was used as a popular method to dealing with issues of data heterogeneity. There are many successful standardization attempts both at the industry as well as firm level. We saw several examples of industry-level standards that have been created successfully, though the level of success is arguable. We also looked in detail at one FS firm's standardization effort. We analyzed their effort using some published theory on data management and standardization. We saw that the firm is taking many right steps towards its goal but pointed out that there are also some issues it needs to deal with if it hopes to achieve success.

We looked at the draw backs of the standardization approach and saw that in particular, it was unable to deal with evolution of standards satisfactorily. When we presented the alternative approach developed at MIT, we saw that the ability to deal with evolution was one of COIN's selling points. Therefore we contend that firms should consider alternatives such as our example if it hopes to achieve the expected results - that is, not just the ability to have standards, but to have standards that can evolve, respond to business needs fast and are cost effective to establish.

This thesis, is in a sense, is a work in progress because standards as well as ways of dealing with data heterogeneity are continuously evolving. Therefore the following ideas briefly present areas where there is much scope for further investigation and study:

a) Study of the "standards war" that is going on in the electronic commerce and FS sector. Detailed investigation may help predict which ones are likely to come out ahead. Even more interesting to investigate would be a firm's options - should firms such as the one we interviewed jump on one bandwagon? Should they stop creating internal standards (such as the Standard Trading Messages) and wait for industry to come up with something that works for all, thus saving future re-work?
b) Another interesting, hands on approach would be to attempt to implement COIN (or a prototype of this) in the "real world" of FS. That is, use data from actual legacy FS systems and attempt to implement the context interchange system and then test whether or not data aggregation is easier. In particular, try using COIN as an alternative (to at least a part of) EDSI and attempt to compile some of the regulatory or risk management reports that FS firms are currently unable to produce.
c) Investigation of the costs of implantation of EDSI (or similar approaches) vs. COIN.

It is hoped that the background provided in this thesis will enable the reader to understand the key issues involved and engaged his/her interest sufficiently in order to do further investigation.

8 Bibliography

[Alverstrand95] Alverstrand, Harald T., *"The Internet Standardization Process"*, presentation given at COST A3 Workshop, November 22-24, 1995

[Bader99] Bader, John ; Hayward, Chris ; Madnick, Stuart; Siegal, Michael; Razzo, Johnathan; "*An Analysis of Data Standardization Across a Capital Markets/Financial Services Firm*", MIT Sloan School, Working Paper, 1999

[Bowers95] Bowers, Tab and Devine, Ted, "*The next upheaval in the US payment systems*", The McKinsey Quarterly, 1995 Number 4, pp. 74-84

[Bressen97] Bressen, Stephane, "*Semantic Integration of Disparate Information sources over the Internet using Constraint Propagation*", 1997

[Brynjolfsson00] Brynjolfsson, Erik and Hitt, Lorin M, "*Computing Productivity: Firm Level Evidence*", MIT Working Paper, April 2000

[Callinan00] Callinan, Patrick, "*Benchmark Data Overview*", The Forrester Report, Q2, 2000

[Coulson82] Coulson, Christopher J., *People Just Aren't Using Data Dictionaries*, Computerworld, August, 1982

[DeLong86] DeLong, David W. and Rockart, John F. *"Identifying the Attributes of successful Executive Support System Implementation"*, Management in the 1990's, CISR Working Paper #132, Sloan School of Management, 1986

[Deutsch98] Deutsch; Waverly; Cameron, Bobby; Hermsdorf, Leslie; "*IT Standards and Catalytic IT*", The Forrester Report, December 1998

[Goldstein96] Goldstein, Morris and Turner, Philip, "*Banking Crisis in Emerging Economies: Origins an Policy Options*", Bank of International Settlements, Biz Economic Papers No. 46 - October 1996, available at www.bis.org/publ/econ46.pdf

[Goodhue90] Goodhue, Dale; Quillard, Judith; Rockart, John F. "*The Management of Data: Preliminary Search Results*", Management in the 1990s, CISR Working Paper #140, Sloan School of Management, 1990

[Goodhue92] Goodhue, Dale L., Wybo, Michael D., and Kirsh, Laurie J., "*The Impact of Data Integration on the Costs and Benefits of Information Systems*", MIS Quarterly, September 1992, pp. 292-311.

[Hamilton99] Hamilton, James, *"Gramm-Leach-Bliley Act Creates Financial Dynamic for the Next Century"*, www.bankinfo.com, 12/01/99

[Kasrel98] Kasrel, Bruce; Doyle, Bill; Metzger, Tell; "*Making Financial Sites Work*", The Forrester Report, July 1998

[Madnick97] Madnick, Stuart E. "*Metadata Jones and the Tower of Babel: The Challenge of Large Scale Semantic Heterogeneity*" 1997

[Malone91] Malone, Thomas W. and Rockart, John F. *Computers, Networks and The Corporation*, CISR Working Paper #232, Center for Information Systems Research, MIT, Aug 1991

[Marino96] Marino, Jory J. "*Glass-Steagall repeal will ignite hiring war for Wall St. data*

executives", American Banker, Tuesday January 23, 1996.

[Meringer97] Meringer, Julie; Deutsch, Waverly; Manousoff, Lucie; "*Standards Meet the Internet"*, The Forrester Report, November 1997

[MSDW99] *The Internet and Financial Services*, Morgan Stanley Dean Witter, Equity Research North America, August 1999

[Porter85] Porter, Michael and Miller, Victor A. "*How Information Gives you Competitive Advantage*", Harvard Business Review, July – Aug 1985

[PSA96] *"SEC considering Revisions to MBS/ABS Disclosure and Reporting Requirements"*, Bond Markets section, PSA Now, September 1996

[Rockart86] Rockart, John F. "*The Role of the Executive in the New Computer Era*": The Rise of Managerial Computing: The best of the Center for Information Systems Research Sloan School of Management, MIT, Edited by John F. Rockart and Chritine V. Bullen, Dow Jones-Irwin, 1986

[Rockart91] Rockart, John F. and Benjamin, Robert, *"The Information Technology Function of the 1990 Function of the 1990s: A unique Hybrid"*, CISR Working Paper #225, Center for Information Systems Research, Sloan School of Management, MIT, 1991

[SBC98] *"Information Technology in Retail Banking"*, SBC Warburg Dillon Read, April 1998

[Schadler97] Schadler, Ted and Dolberg, Stan, "*Data Warehouse Strategies"*, The Forrester Report, September 1997

[Schmidt98] Schmidt, Susan K. and Werle, Raymund, "*Coordinating Technology: Studies in the International Standardization Telecommunications"*, MIT Press, 1998

[SWIFT00] *"Merrill Lynch Taps SWIFT for ETC"*, SWIFT Solutions #6 May 2000, available at www.swift.com/solutions/articles6/6_merrill.htm

[SWIFT99] *Global STP Benchmarking*, SWIFT Solutions, #5, November 1999

[Trombly00-1] Trombly, Maria *Big names back new XML-based financial standard*, Computerworld, 07 April 2000 (www.computerworld.com)

[Wilson95] Wilson, Gregory, "*Getting Beyond Glass-Steagall"*, The McKinsey Quarterly, 1995 Number 2, pp. 109-115

[Rockart88] Rockart, John F. and Short, James E. *"Information Technology and the New Organization: Towards More Effective Management of Interdependence"*, Management in the 1990s, CISR Working Paper #180, Sloan School of Management, MIT, 1988

[Trombly00-2] Trombly, Maria *Finance Players Back XML-Based Standard: Consortium to push report specification*, Computerworld, 17 April 2000 (www.computerworld.com)

Other Data Sources

www.xfrml.org for details, history, latest developments of XBRL

www.cusip.com for details, history and listing of CUSIP and CINS numbers

www.swift.com for details, history latest developments of the SWIFT messaging protocol.

www.fixprotocol.org for details about the FIX protocol

www.bog.frb.fed.us for BHC Act Regulatory Requirements

www.psa.com for news about the Public Securities Association

www.ebxml.org

www.biztalk.org