Motivation

- Distributed databases makes many disparate sources available.
- The web is making even more semi-structured sources available.
  -- With XML and Web Wrapping, these can be treated as databases.
- Schema integration addresses the problem of syntactic inconsistencies.
  -- i.e., differing structures.
- How do we address semantic inconsistencies.
  -- i.e., differing meanings.
  (e.g., what does “price” really mean?)
Outline

• Story of the "Tower of Babel"
• Vision of the Future
  ⇒ Growing number of information sources accessible via “Information SuperHighway”
• Implications and Integration Challenges
  ⇒ Large-Scale Semantic Heterogeneity (More data, less understanding)
• Semantic Integration Approaches
  ⇒ Context Mediation Metadata Services

PROBLEM / NEED: EXAMPLE

• Want:
  – ski areas
  – more than 5 feet of snow

• What you get:

• Need to overcome two problems:
  1. Locate and extract snow depth information from Web pages (“data extraction”)
  2. Realize some inches ("), some centimeters (cm, cms) (“data interpretation”)
APPLICATIONS: Financial services, electronic commerce, asset visibility, in-transit visibility

Question: How much funding is left for Project A?
**Role Of Context**

---

**CONTEXT VARIATIONS:**
- GEOGRAPHIC (US vs. UK)
- FUNCTIONAL (CASH MGMT vs. LOANS)
- ORGANIZATIONAL (CITIBANK vs. CHASE)

---

**Types of Context**

- Representational
- Ontological
- Temporal

<table>
<thead>
<tr>
<th>Type</th>
<th>Example</th>
<th>Temporal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Representational</td>
<td>Currency: $ vs €</td>
<td>Francs before 2000, € thereafter</td>
</tr>
<tr>
<td></td>
<td>Scale factor: 1 vs 1000</td>
<td></td>
</tr>
<tr>
<td>Ontological</td>
<td>Revenue: Includes vs excludes interest</td>
<td>Revenue: Excludes interest before 1994</td>
</tr>
<tr>
<td></td>
<td></td>
<td>but incl. thereafter</td>
</tr>
</tbody>
</table>
Example : Context Differences
(from multiple web sources)

Daimler Benz (DCX) Financial Data

<table>
<thead>
<tr>
<th>P/E Ratio</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC</td>
<td>11.6</td>
</tr>
<tr>
<td>Bloomberg</td>
<td>5.57</td>
</tr>
<tr>
<td>DBC</td>
<td>19.19</td>
</tr>
<tr>
<td>MarketGuide</td>
<td>7.46</td>
</tr>
</tbody>
</table>

“What’s the P/E Ratio?”

August 21, 2001 – Wall Street Journal -- Page One Feature:
What’s the P/E Ratio? Well, Depends on Meaning of Earnings
By JONATHAN WEIL, Staff Reporter of THE WALL STREET JOURNAL

Few investors know it, but the U.S. stock market today is, by one way of looking at it, the most expensive it has ever been. How could that be, after the numbing slide since the market peaked in early 2000? It turns out that for all the pain, the stock market now is far out of whack with historical norms by one common measure, the price-to-earnings ratio.

The P/E ratio measures how companies’ share prices compare with their profits, showing how much value the market places on each dollar of a company’s earnings. The lower the P/E, as a rough rule of thumb, the cheaper the stock. Though this guide to value has lots of exceptions, it remains a venerable market benchmark. …

… Example of FMC: P/E = 1.58 from First Call
P/E = -9.62 according to WSJ calculation
The 1999 Overture

Unit-of-measure mixup tied to loss of $125 Million Mars Orbiter

“NASA’s Mars Climate Orbiter was lost because engineers did not make a simple conversion from English units to metric, an embarrassing lapse that sent the $125 million craft off course. . . .

. . . The navigators (JPL) assumed metric units of force per second, or newtons. In fact, the numbers were in pounds of force per second as supplied by Lockheed Martin (the contractor).”

Primark Web Examples

Top 25 US Co. by Net Sales (Disclosure)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Company</th>
<th>Net Sales (000's)</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>General Motors Corp</td>
<td>168,828,600</td>
<td>12/31/95</td>
</tr>
<tr>
<td>2</td>
<td>Ford Motor Co</td>
<td>137,437,000</td>
<td>12/31/95</td>
</tr>
<tr>
<td>3</td>
<td>Exxon Corp</td>
<td>121,804,000</td>
<td>12/31/95</td>
</tr>
<tr>
<td>4</td>
<td>Wal Mart Stores Inc</td>
<td>93,627,000</td>
<td>01/31/96</td>
</tr>
<tr>
<td>5</td>
<td>AT&amp;T</td>
<td>79,609,000</td>
<td>12/31/95</td>
</tr>
<tr>
<td>6</td>
<td>Mobil Corp</td>
<td>73,413,000</td>
<td>12/31/95</td>
</tr>
<tr>
<td>7</td>
<td>International Business</td>
<td>71,904,000</td>
<td>12/31/95</td>
</tr>
<tr>
<td>8</td>
<td>General Electric Co</td>
<td>70,028,000</td>
<td>?</td>
</tr>
</tbody>
</table>

Top 25 International Co. by Net Sales (Worldscope)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Company</th>
<th>Net Sales (000's)</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mitsubishi Corporation</td>
<td>165,848,468</td>
<td>03/31/96</td>
</tr>
<tr>
<td>2</td>
<td>General Motors Corp</td>
<td>163,861,100</td>
<td>12/31/95</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>8</td>
<td>Exxon Corp</td>
<td>107,893,000</td>
<td>12/31/95</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>16</td>
<td>International Business M</td>
<td>71,940,000</td>
<td>12/31/95</td>
</tr>
<tr>
<td>17</td>
<td>General Electric Co</td>
<td>68,948,000</td>
<td>12/31/95</td>
</tr>
<tr>
<td>20</td>
<td>Mobil Corp</td>
<td>64,767,000</td>
<td>12/34/95</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Primark was a company that owned:
- Disclosure
- Worldscope
- Datastream
- Information services
Another Context Example
(Basis for Demo)

<table>
<thead>
<tr>
<th>Company Name</th>
<th>Net Income</th>
<th>Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAIMLER-BENZ AG</td>
<td>346,577</td>
<td>56,268,168</td>
</tr>
<tr>
<td>DAIMLER BENZ CORP</td>
<td>615,000,000</td>
<td>97,737,000,000</td>
</tr>
</tbody>
</table>

O&A DEM-USD Exchange Rate
1.00 German Mark = 0.58 US Dollar as 12/31/93

Some Context Differences
Context Definitions

<table>
<thead>
<tr>
<th></th>
<th>Disclosure</th>
<th>Worldscope</th>
<th>DataStream</th>
</tr>
</thead>
<tbody>
<tr>
<td>Currency Used</td>
<td>Country of Incorporation</td>
<td>USD</td>
<td>Country of Incorporation</td>
</tr>
<tr>
<td>Currency Conversion</td>
<td>Money Amount As Of Date</td>
<td>Money Amount As Of Date</td>
<td>Money Amount As Of Date</td>
</tr>
<tr>
<td>Currency Symbols</td>
<td>3 Letters</td>
<td>3 Letters</td>
<td>2 Letters</td>
</tr>
<tr>
<td>Scale Factor</td>
<td>1</td>
<td>1000</td>
<td>1000</td>
</tr>
<tr>
<td>Company Names</td>
<td>Disclosure Names</td>
<td>Worldscope Names</td>
<td>DataStream Names</td>
</tr>
<tr>
<td>Date Style</td>
<td>American with '/' as separator</td>
<td>American with '/' as separator</td>
<td>European with '-' as separator</td>
</tr>
</tbody>
</table>

Olsen (OANDA) Web Source uses 3 Letter Currency Symbols and European Date Style with '/' as a separator
**Domain Model**

- **number**
- **exchange-Rate**
- **string**
- **currency-Type**
- **scaleFactor**
- **currency**
- **country-Name**
- **fyEnding**
- **date**
- **OfficialCurrency**

Some currency context possibilities:
- Currency is stated explicitly as part of record
- Currency not stated, but the same for all (e.g., US $)
- Currency not stated or constant, but inferred by country

---

**COIN System Architecture**

- **SERVER PROCESSES**
  - WWW Gateway
  - Wrapper
  - Web-site

- **MEDIATOR PROCESSES**
  - COIN Repository
  - Context Mediation
  - Optimizer
  - Executioner

- **CLIENT PROCESSES**
  - ODBC-compliant Apps (e.g., Microsoft Excel)
  - ODBC-Driver
  - CGI-scripts
  - Web Client

17

18
System Demonstration
Single Source Queries with Mediation


Query: Find out from Disclosure what Net Income for DAIMLER-BENZ was. Use Datastream Context.

Capabilities Demonstrated:
Ability to perform Scale Factor Conversion, Date Format Conversion, Company Name Conversion.
## Conflict Detection and Mediation

### Mediated Query in Datalog

- **Date convert**
- **Scale factor convert**
- **Name convert**
Mediated SQL Query & Result

**Mediated SQL Query**

```sql
SELECT datesform_date2, discaf_net_income*0.01
FROM datesform
WHERE datesform_company_name = 'DAIMLER-BENZ AG'
AND datesform_AS_OF_DATE = '01/05/94'
AND datesform_company_name = discaf_company_name
AND datesform_company_name = dstreamaf_company_name
AND datesform_company_name = quotes.Cname;
```

**Final results – from Disclosure but in Datastream context**

<table>
<thead>
<tr>
<th>Result</th>
<th>DiscAF.LATEST_ANNUAL_DATA.DiscAF.NET_INCOME</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>31-12-93</td>
</tr>
<tr>
<td>Amount</td>
<td>100000</td>
</tr>
</tbody>
</table>

More Complex Example (4 sources: DB + Web)

**eCOIN Demo for TASC Financial Example**

(Tasc11)

```sql
SELECT WorldAF.TOTAL_ASSETS, DiscAF.NET_SALES,
DiscAF.NET_INCOME, DStreamAF.TOTAL_EXTRAORD_ITEMS_PRE_TAX,
quotes.Last
FROM WorldAF, DiscAF, DStreamAF, quotes
WHERE WorldAF.COMPANY_NAME = "DAIMLER-BENZ AG"
AND DStreamAF.AS_OF_DATE = '01/05/94'
AND WorldAF.COMPANY_NAME = DStreamAF.NAME
AND WorldAF.COMPANY_NAME = DiscAF.COMPANY_NAME
AND WorldAF.COMPANY_NAME = quotes.Cname;
```
## Conflict Table (1st part)

<table>
<thead>
<tr>
<th>SemanticType</th>
<th>Column</th>
<th>Source</th>
<th>Modifier</th>
<th>Value in source context</th>
<th>Value in target context</th>
<th>Conversion Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>company_financials</td>
<td>IV</td>
<td>DiscAF(Name, FYEnd, ShareType, Income, Assets, Incorp)</td>
<td>currency</td>
<td>disclosure : V01</td>
<td>worldscope : USD</td>
<td>atr(V13, FYEnd, V12, value (V12, V11, V10), dateEnd, (V0, V8, V7, V6), value(V0, V9, V11, V5), value(V0, V11, V4), value(V0, V11, V10), value(V7, V11, V3), V2 is V1 * V3)</td>
</tr>
<tr>
<td>company_financials</td>
<td>Income</td>
<td>DiscAF(Name, FYEnd, ShareType, Income, Assets, Incorp)</td>
<td>scaleFactor</td>
<td>disclosure : 1</td>
<td>worldscope : 1000</td>
<td>V5 is V4 / V3, V2 is V1 * V3</td>
</tr>
<tr>
<td>company_name</td>
<td>Name</td>
<td>DiscAF(Name, FYEnd, ShareType, Income, Assets, Incorp)</td>
<td>format</td>
<td>disclosure : ds_name</td>
<td>worldscope : ws_name</td>
<td>name_map (V4, V3, V2, V1)</td>
</tr>
<tr>
<td>company_name</td>
<td>CName</td>
<td>DiscAF(CName, Name)</td>
<td>format</td>
<td>value : ws_name</td>
<td>worldscope : ws_name</td>
<td>name_map (V4, V3, V2, V1)</td>
</tr>
</tbody>
</table>

## Conflict Table (2nd part)

<table>
<thead>
<tr>
<th>SemanticType</th>
<th>Column</th>
<th>Source</th>
<th>Modifier</th>
<th>Value in source context</th>
<th>Value in target context</th>
<th>Conversion Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>company_financials</td>
<td>Income</td>
<td>DStreamAF (Account, Name, Title, Department, Currency)</td>
<td>currency</td>
<td>datastream : USD</td>
<td>worldscope : USD</td>
<td>atr(V13, FYEnd, V12, value (V12, V11, V10), dateEnd, (V0, V8, V7, V6), value(V0, V9, V11, V5), value(V0, V11, V4), value(V0, V11, V10), value(V7, V11, V3), V2 is V1 * V3)</td>
</tr>
<tr>
<td>date</td>
<td>AsOfDate</td>
<td>DStreamAF (Account, Name, Title, Department, Currency)</td>
<td>dateStart</td>
<td>datastream : European Style</td>
<td>worldscope : American Style</td>
<td>datetimestr(V4, V3, V2, V1)</td>
</tr>
<tr>
<td>company_name</td>
<td>Name</td>
<td>DStreamAF (Account, Name, Title, Department, Currency)</td>
<td>format</td>
<td>datastream : ds_name</td>
<td>worldscope : ws_name</td>
<td>name_map (V4, V3, V2, V1)</td>
</tr>
<tr>
<td>currency_type</td>
<td>Currency</td>
<td>DStreamAF (Account, Name, Title, Department, Currency)</td>
<td>currencySync</td>
<td>datastream : 2char</td>
<td>worldscope : 2char</td>
<td>Currency_map (V2, V1)</td>
</tr>
<tr>
<td>currency_type</td>
<td>USD</td>
<td>DStreamAF (Account, Name, Title, Department, Currency)</td>
<td>currencySync</td>
<td>datastream : 2char</td>
<td>worldscope : 2char</td>
<td>Currency_map (V2, V1)</td>
</tr>
</tbody>
</table>
Generated SQL (1st Part)

```sql
select worldcaf.total_assets, discaf.net_sales, ((discaf.net_income*0.001)*olsen.rate),
(dstreamaf2.total_extraord_items_pre_tax*olsen2.rate), quotes.Last
from (select date1, 'European Style -', '01/05/94', 'American Style /
from datesform
where format1='European Style -'
and date2='01/05/94'
and format2='American Style /
from datesform,
(select dt_names, 'DAIMLER-BENZ AG'
from name_map_d_ws
where ws_names='DAIMLER-BENZ AG' name_map_d_ws,
(select da_names, 'DAIMLER-BENZ AG'
from name_map_d_ws
where ws_names='DAIMLER-BENZ AG' name_map_d_ws,
(select DAIMLER-BENZ AG, ticker, exc
from ticker_lookup2
where comp_name='DAIMLER-BENZ AG' ticker_lookup2,
(select DAIMLER-BENZ AG, latest_annual_financial_date, current_outstanding_shares, net_income, sales, total_assets,
country_of_incorp
from worldcaf
where company_name='DAIMLER-BENZ AG' worldcaf,
(select country, currency
from currencytypes
where currency <> 'USD'
currencytypes,
(select exchanged, 'USD', rate, date
from olsen
where expressed='USD'
olsen,
(select company_name, latest_annual_data, current_shares_outstanding, net_income, net_sales, total_assets, location_of_incorp
from discaf
where company_name='DAIMLER-BENZ AG' discaf,
(select as_of_date, name, total_sales, total_extraord_items_pre_tax,
earned_for_ordinary, currency
from dstreamaf)dstreamaf,
(select as_of_date, name, total_sales, total_extraord_items_pre_tax, earned_for_ordinary, currency
from dstreamaf2)dstreamaf2,
(select char3_currency, char2_currency
from currency_map
where char3_currency <> 'USD'
currency_map,
(select country, currency
from currencytypes
where currency <> 'USD'
currencytypes2,
(select exchanged, 'USD', rate, '01/05/94'
from olsen
where expressed='USD'
and date='01/05/94'
olsen2,
(select Cname, Last
from quotes)
quotes
where currencytypes.country = discaf.location_of_incorp
and currencytypes.currency = olsen.exchanged
and dstreamaf.currency = dstreamaf2.currency
and dstreamaf2.currency = currency_map.char2_currency
and olsen.date = dstreamaf.as_of_date
dstreamaf.date = dstreamaf2.as_of_date
dstreamaf.exchanged = dstreamaf2.exchanged
dstreamaf2.currency = dstreamaf2.currency
and currencytypes2.currency = dstreamaf2.currency
and currencytypes2.currency = dstreamaf.currency
and name_map_d_ws.dt.name = dstreamaf2.name
and name_map_d_ws.ds.name = dstreamaf.name
and name_map_d_ws.ds.name = dstreamaf.name
ticker_lookup2.Cname = quotes.Cname
and datesform.date1 = dstreamaf2.as_of_date
dstreamaf.currency = US$'
union
select worldcaf2.total_assets, discaf2.net_sales, ((discaf2.net_income*0.001)*olsen1.rate),
dstreamaf2.total_extraord_items_pre_tax, quoted2.Last
from (select date1, 'European Style -', '01/05/94', 'American Style /
from datesform
where format1='European Style -'
and date2='01/05/94'
and format2='American Style /
from datesform,
(select dt_names, 'DAIMLER-BENZ AG'
from name_map_d_ws
where ws_names='DAIMLER-BENZ AG' name_map_d_ws,
(select da_names, 'DAIMLER-BENZ AG'
from name_map_d_ws
where ws_names='DAIMLER-BENZ AG' name_map_d_ws,
(select DAIMLER-BENZ AG, ticker, exc
from ticker_lookup2
where comp_name='DAIMLER-BENZ AG' ticker_lookup2,
(select DAIMLER-BENZ AG, latest_annual_financial_date, current_outstanding_shares, net_income, sales, total_assets,
country_of_incorp
from worldcaf
where company_name='DAIMLER-BENZ AG' worldcaf,
(select country, currency
from currencytypes
where currency <> 'USD'
currencytypes3,
(select exchanged, 'USD', rate, '01/05/94'
from olsen
where expressed='USD'
and date='01/05/94'
olsen3,
(select Cname, Last
from quotes)
quotes2
where currencytypes2.country = discaf.location_of_incorp
and currencytypes2.currency = olsen.exchanged
and dstreamaf2.currency = dstreamaf2.currency
and dstreamaf2.currency = currency_map.char2_currency
and dstreamaf2.currency = currency_map.char2_currency
and olsen2.date = dstreamaf2.as_of_date
dstreamaf2.date = discaf2.as_of_date
dstreamaf2.currency = discaf2.currency
and currencytypes3.currency = discaf2.currency
and currencytypes3.currency = dstreamaf2.currency
and name_map_d_ws2.dt.name = discaf2.name
and name_map_d_ws2.ds.name = dstreamaf2.name
ticker_lookup22.Cname = quotes2.Cname
and datesform2.date1 = discaf2.as_of_date
dstreamaf2.currency = US$'
union
select worldcaf22.total_assets, discaf22.net_sales, ((discaf22.net_income*0.001)*olsen2.rate),
dstreamaf22.total_extraord_items_pre_tax, quoted22.Last
```

Generated SQL (Continued - Partial)

```sql
```

etc
Final Result

A multidatabase query to get information from four different databases with different contexts. The user chose one of the contexts.

- From WorldAsf.
- Where WorldAsf, COMPANY_NAME = "CRABO-BB" and WorldAsf, ASSETS_DATE = '01/01/74' and WorldAsf, COMPANY_NAME = "Stellaris.UNIKE"

Parallel Execution

Retrieving data From Web source

Execution Trace (1st Part - Partials)

...
The 1805 Overture

In 1805, the Austrian and Russian Emperors agreed to join forces against Napoleon. The Russians promised that their forces would be in the field in Bavaria by Oct. 20. The Austrian staff planned its campaign based on that date in the Gregorian calendar. Russia, however, still used the ancient Julian calendar, which lagged 10 days behind.

The calendar difference allowed Napoleon to surround Austrian General Mack's army at Ulm and force its surrender on Oct. 21, well before the Russian forces could reach him, ultimately setting the stage for Austerlitz.

Summary

• Tremendous opportunity to gather and integrate information from many diverse sources
• But ... need to overcome many context challenges
• Context-type “metadata” plays a critical role
• COIN technology can be an important aid for semantically meaningful information integration:
  - Scalable
  - Extensible
  - Application Domain Merging
  - Reuse and extension of ontologies and contexts