

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
SLOAN SCHOOL OF MANAGEMENT

15.565 Integrating Information Systems:

Technology, Strategy, and Organizational Factors

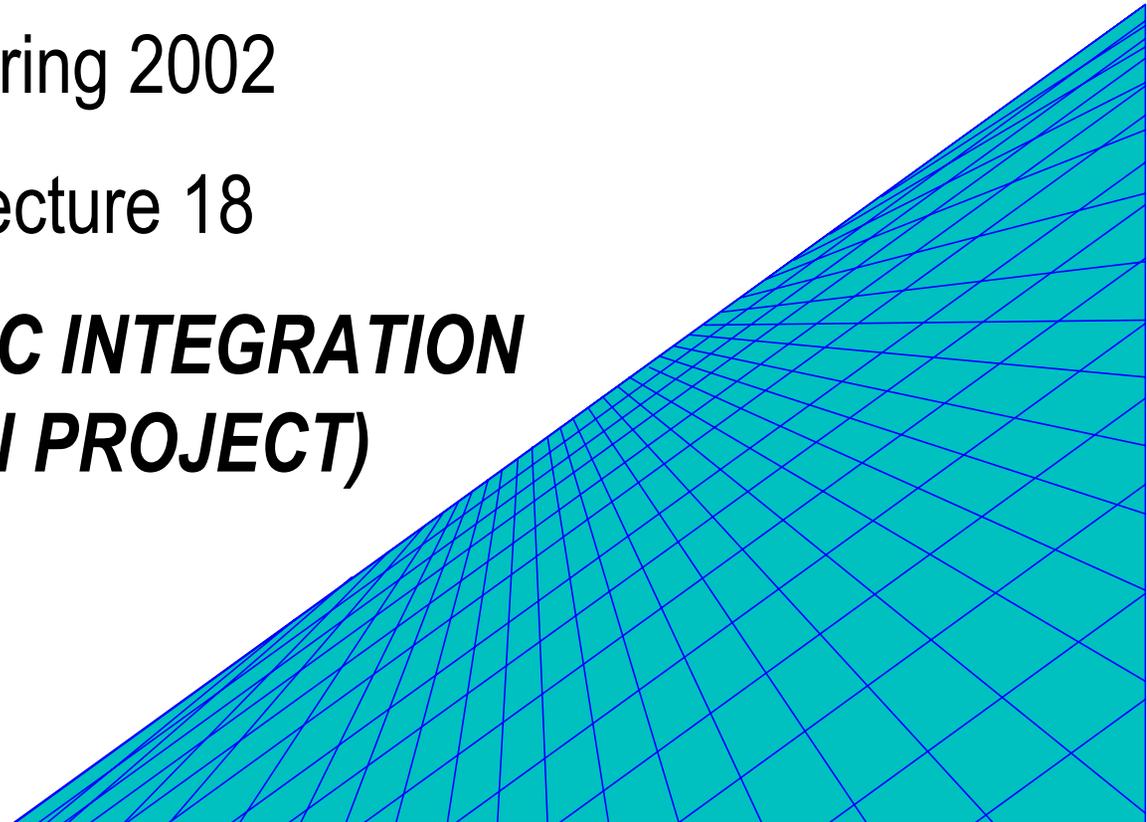
15.578 Global Information Systems:

Communications & Connectivity Among Information Systems

Spring 2002

Lecture 18

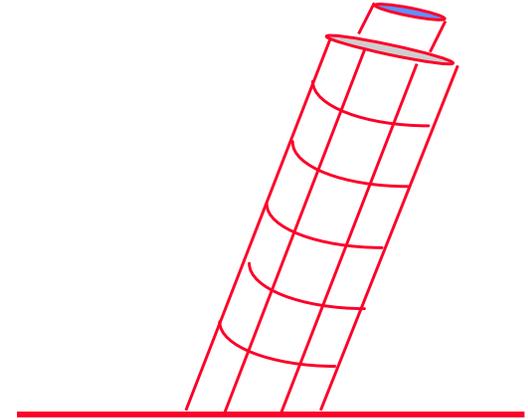
***SEMANTIC INTEGRATION
(COIN PROJECT)***



Motivation

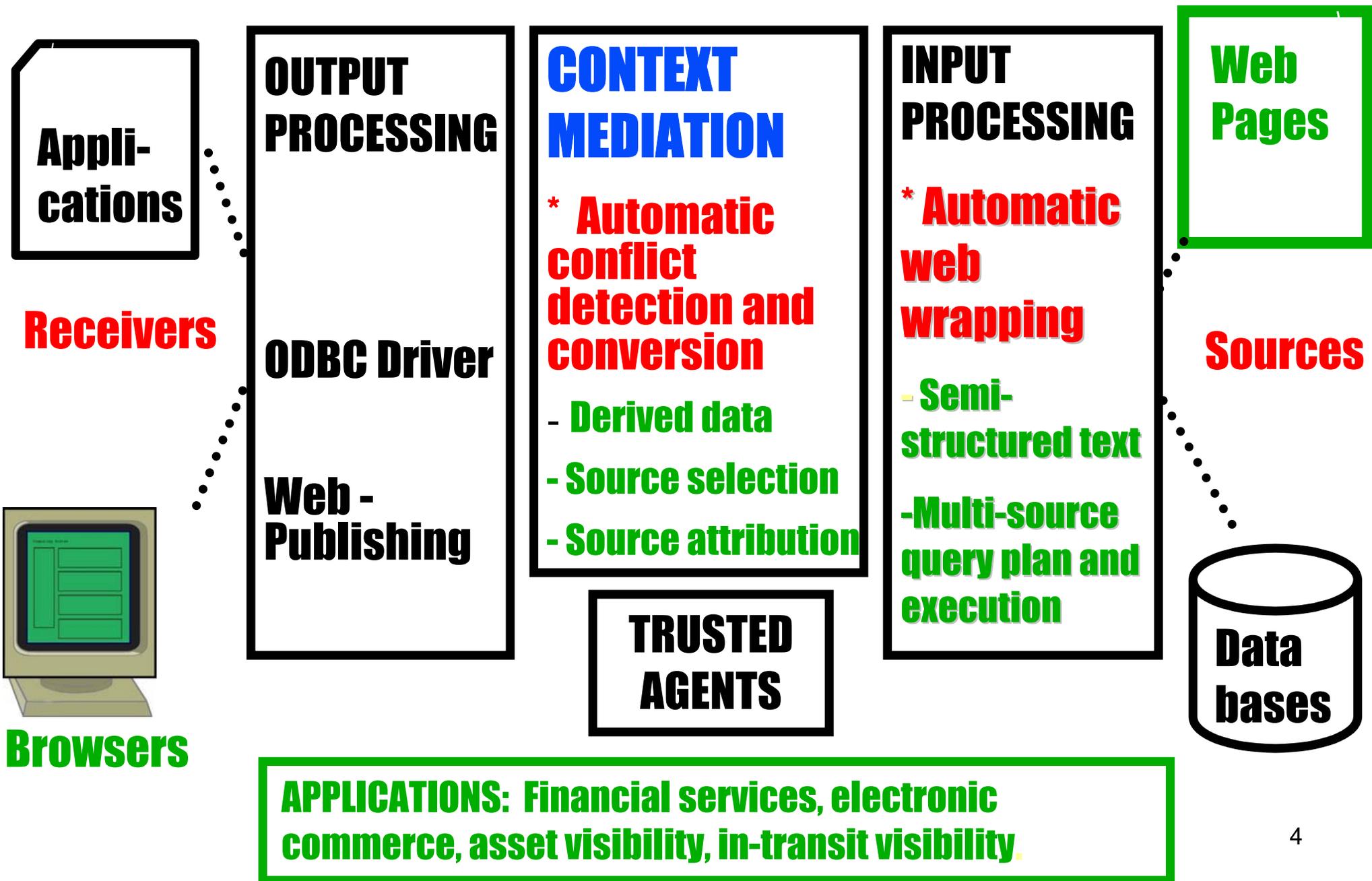
- **Distributed databases** makes many disparate sources available.
- The web is making even more semi-structured sources available.
 - With **XML and Web Wrapping**, these can be treated as databases.
- **Schema integration** addresses the problem of syntactic inconsistencies.
 - i.e., differing structures.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

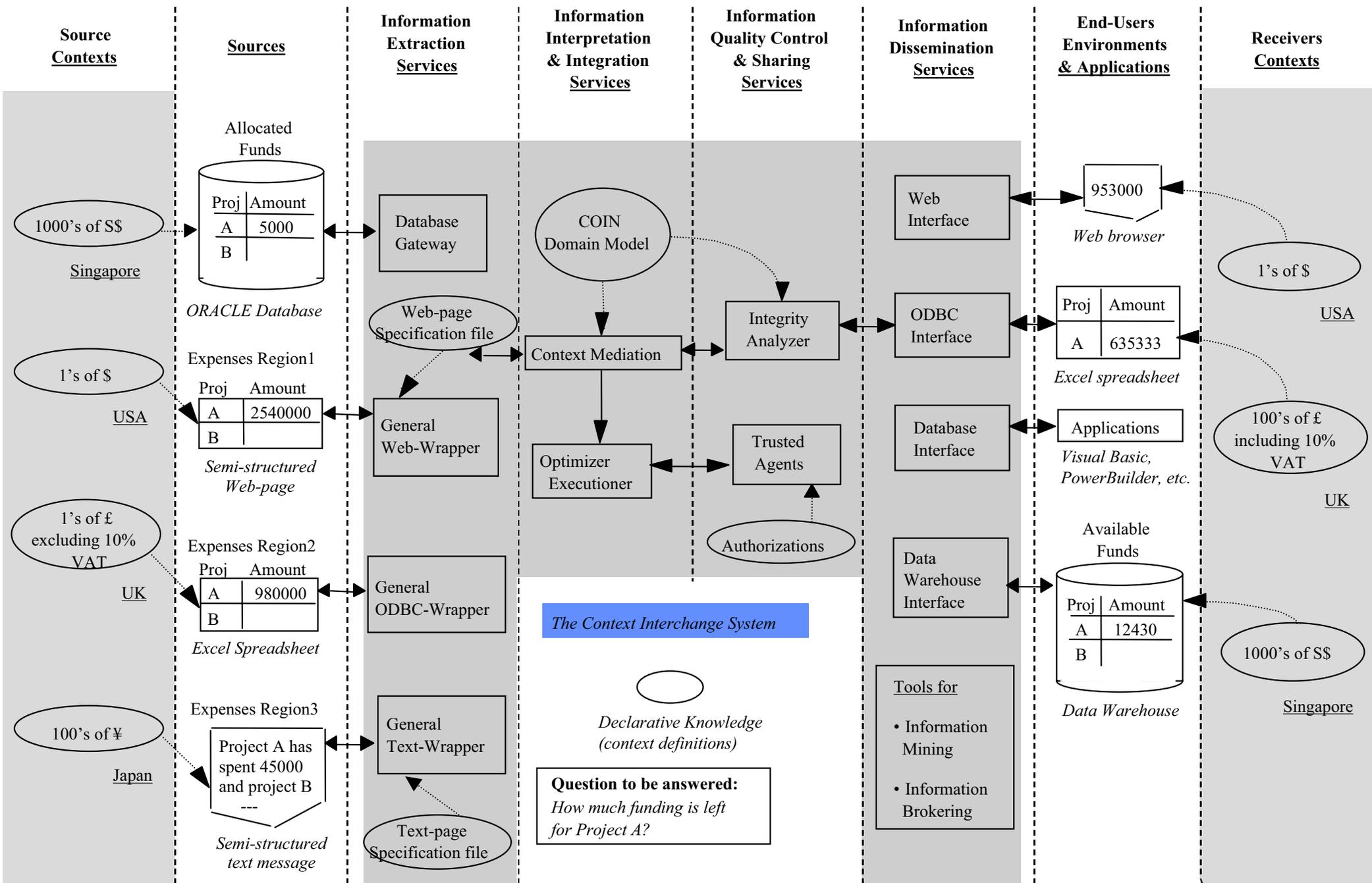
Outline



- **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**

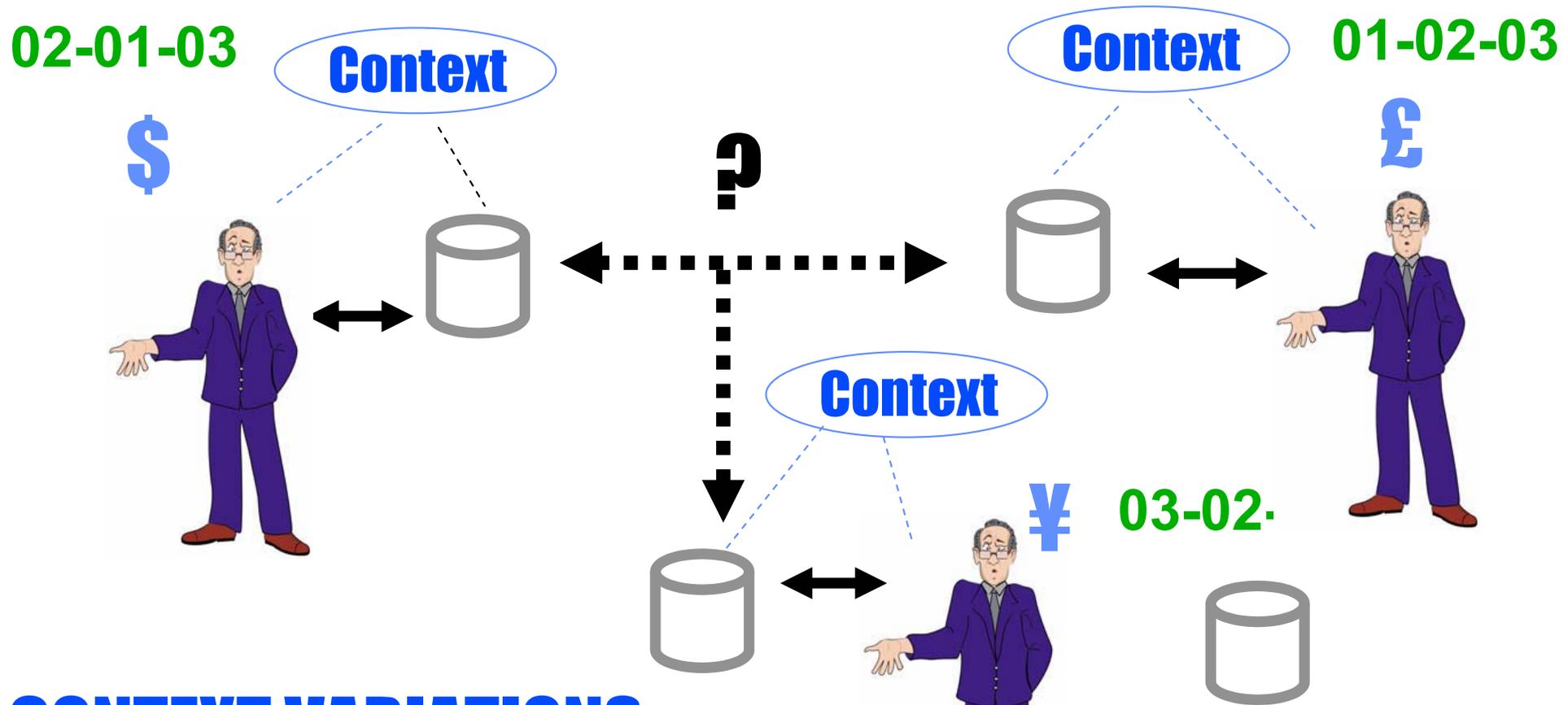
MIT Sloan COntext INterchange (COIN) Project





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)**

Data:
Databases
Web data
E-mail

Example : Context Differences

(from multiple web sources)

Daimler Benz (DAI) Key Ratios

	<u>P/E Ratio</u>	<u>EPS</u>	<u>Dividend</u>
ABC	11.6		0.29
Bloomberg	5.57	15.32	8.127
DBC	19.19	4.36	0.899
MarketGuide	7.46	10.83	0.47

“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

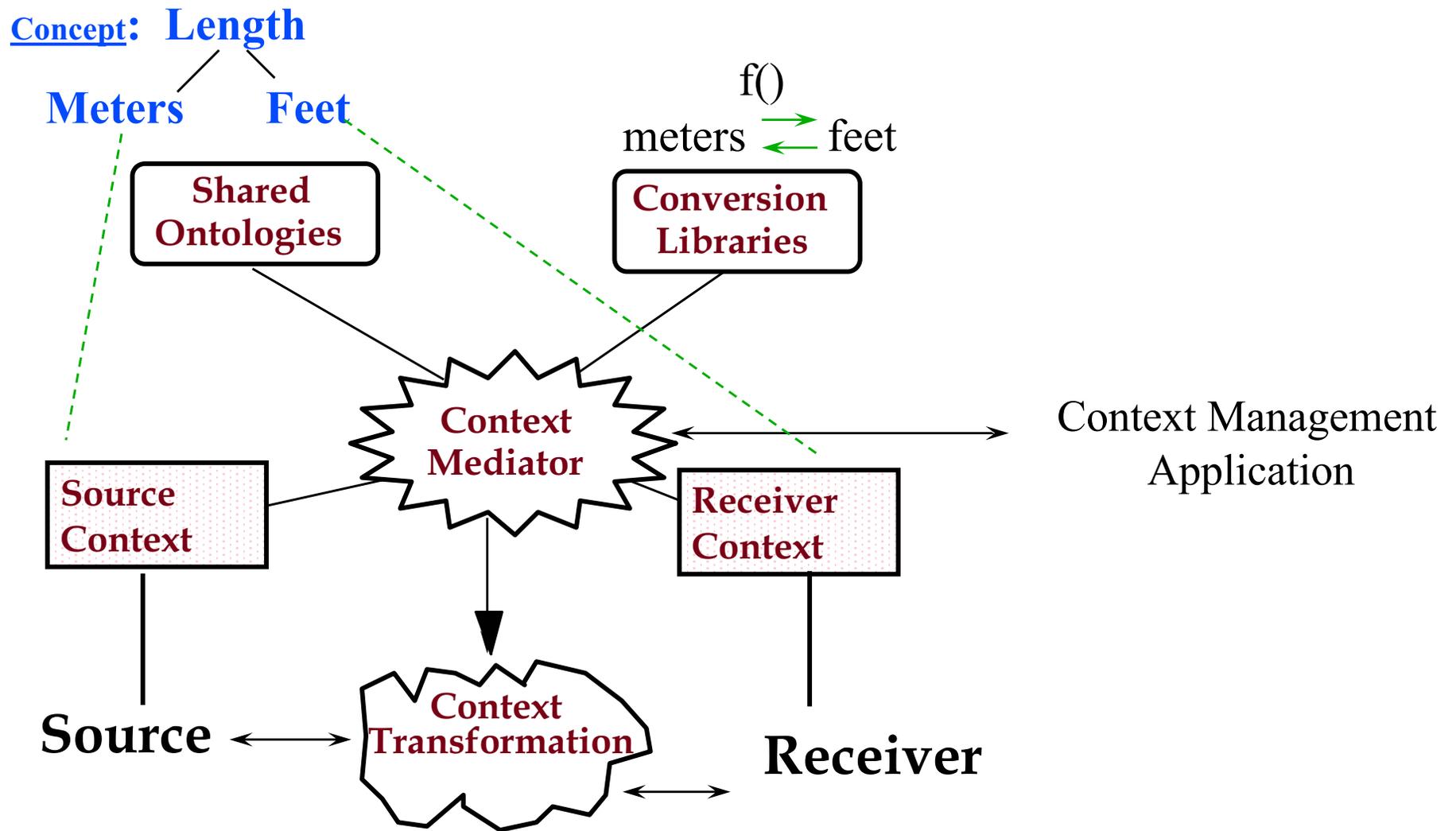
Unit-of-measure mixup tied to loss of \$125 Million Mars Orbiter

“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. ...

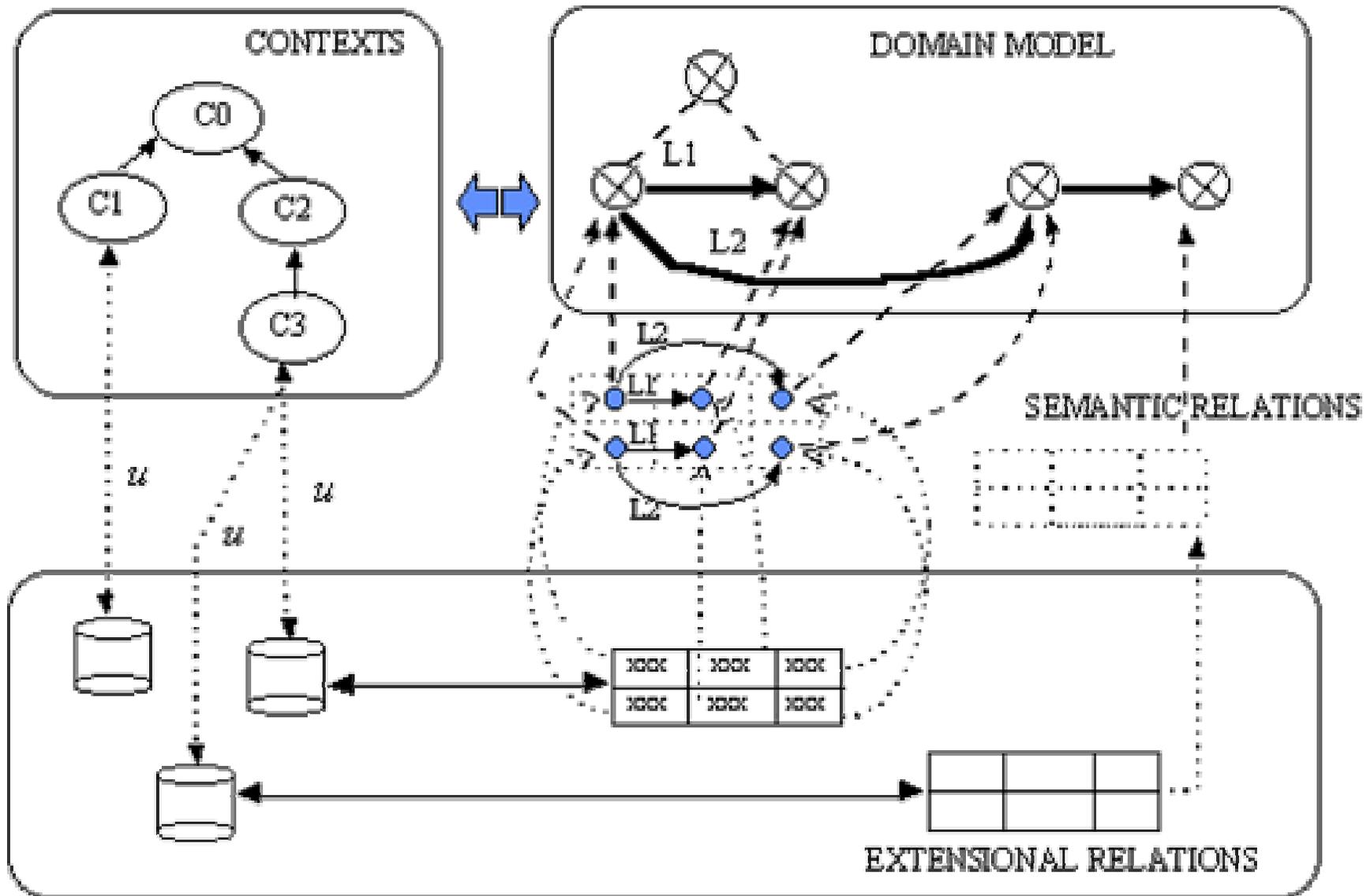
... The navigators (JPL) assumed metric units of force per second, or newtons. In fact, the numbers were in pounds of force per second as supplied by Lockheed Martin (the contractor).”

Source: Kathy Sawyer, *Boston Globe*, October 1, 1999, page 1.

The Context Interchange Approach



COIN Elevation Axioms



Primark Web Examples

Top 25 US Co. by Net Sales (Disclosure)

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

Top 25 International Co. by Net Sales (Worldscope)

Rank	Company	Net Sales (000's)	Date
1	Mitsubishi Corporation	165,848,468	03/31/96
2	General Motors Corp	163,861,100	12/31/95
...
8	Exxon Corp	107,893,000	12/31/95
...
16	International Business M	71,940,000	12/31/95
17	General Electric Co	69,948,000	12/31/95
20	Mobil Corp	64,767,000	12/31/95
...

Primark was a company that owned:

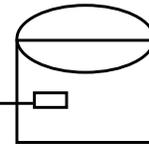
- Disclosure
- Worldscope
- Datastream

Information services

Another Context Example

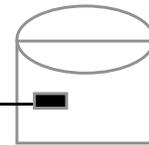
Company Name	DAIMLER-BENZ
Net Income	614,995
Sales	97,736,992

Context Mediation Services



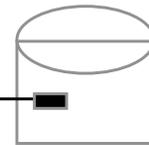
Datastream

Company Name	DAIMLER-BENZ AG
Net Income	346,577
Sales	56,268,168



WorldScope

Company Name	DAIMLER BENZ CORP
Net Income	615,000,000
Sales	97,737,000,000

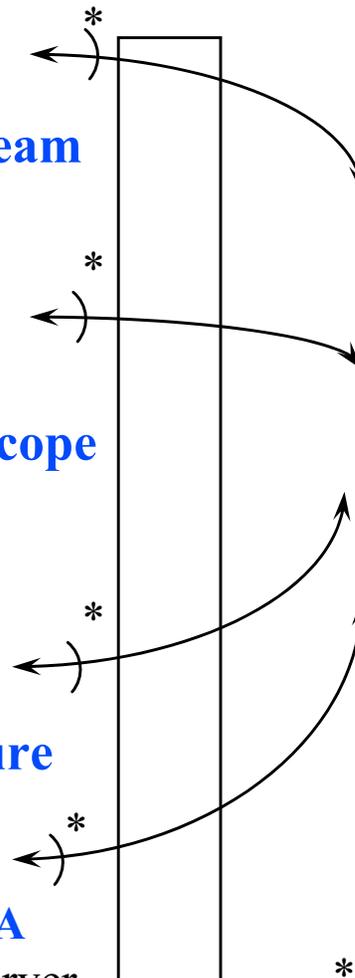


Disclosure

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

OANDA

Web Server



Users & Appl.
Systems

* Wrapper Services

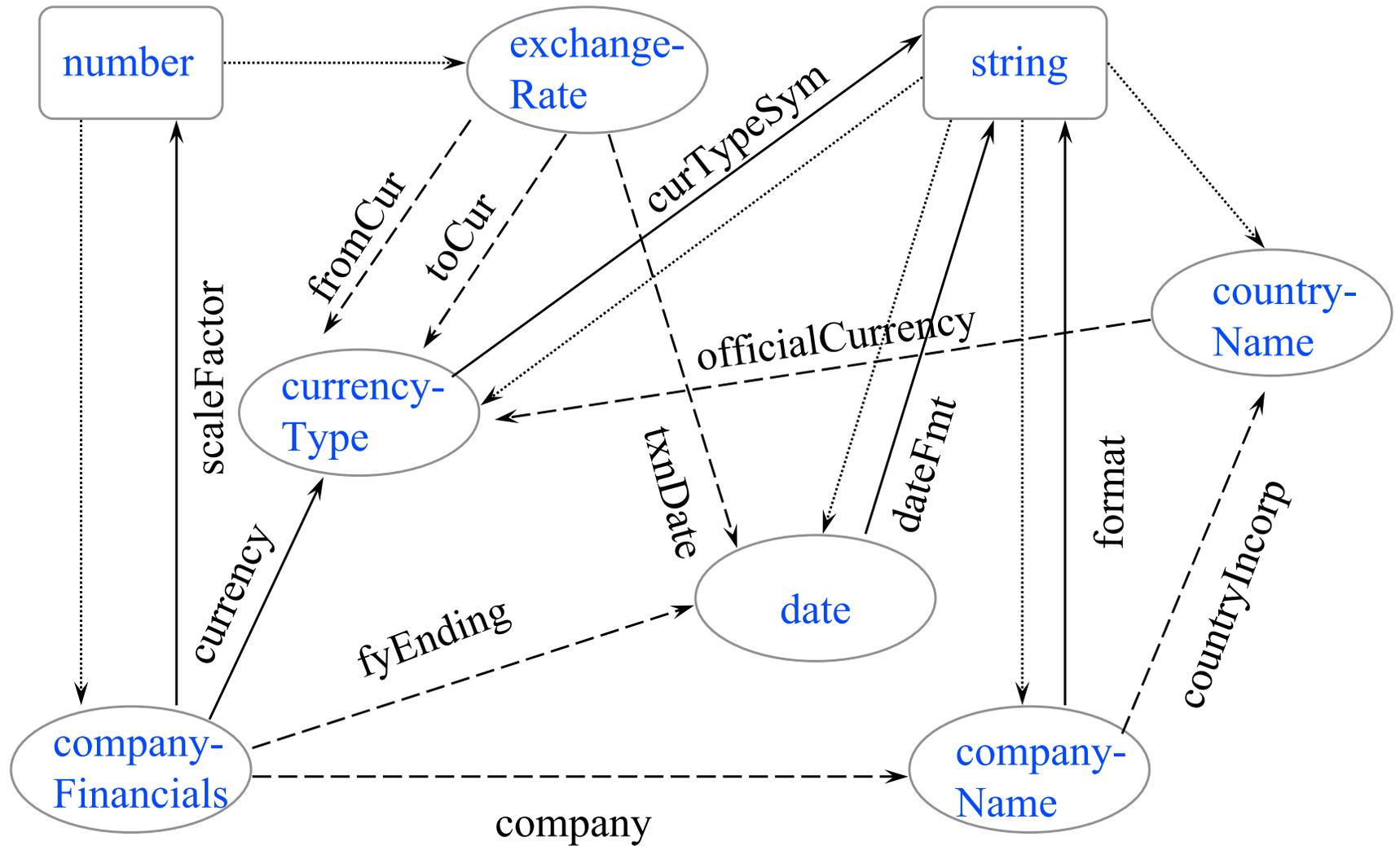
Some Context Differences

Context Definitions

	Disclosure	Worldscope	DataStream
Currency Used	Country of Incorporation	USD	Country of Incorporation
Currency Conversion	Money Amount As_Of_Date	Money Amount As_Of_Date	Money Amount As_Of_Date
Currency Symbols	3 Letters	3 Letters	2 Letters
Scale Factor	1	1000	1000
Company Names	Disclosure Names	Worldscope Names	DataStream Names
Date Style	American with '/' as separator	American with '/' as separator	European with '-' as separator

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

Domain Model



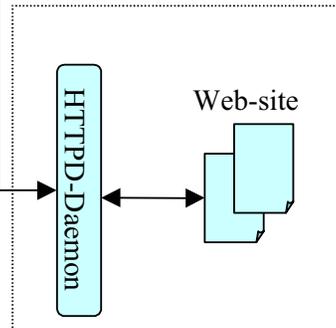
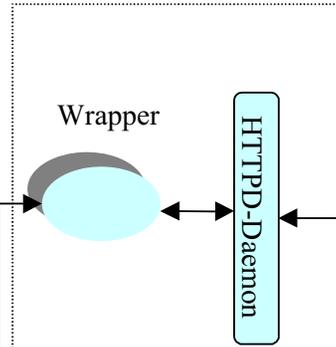
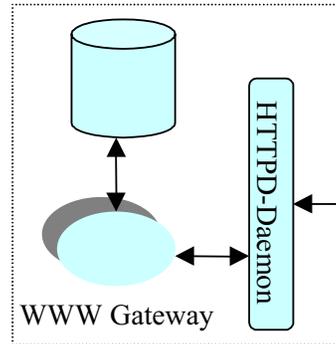
Some currency context possibilities:

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

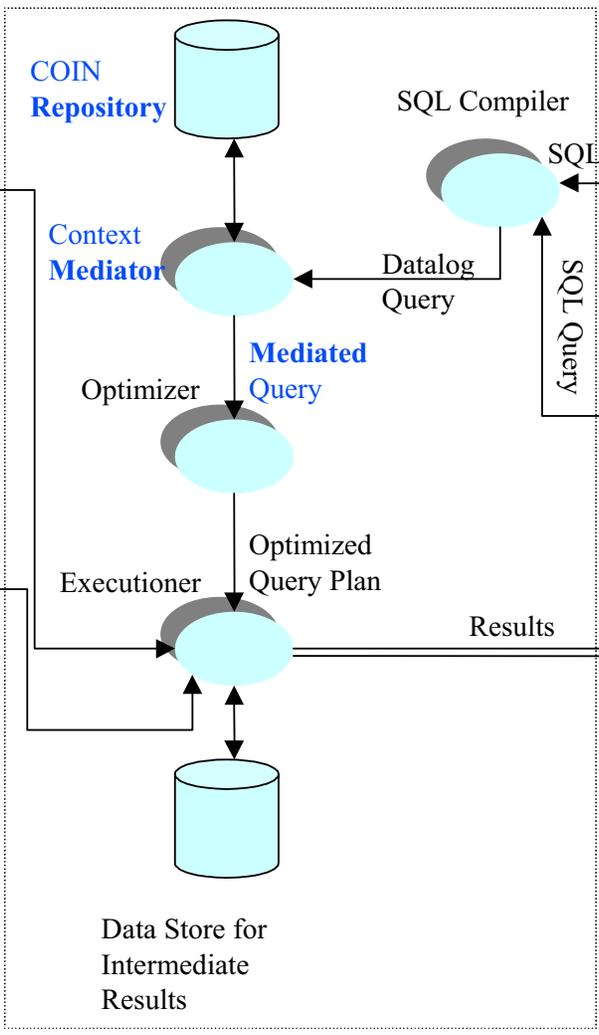


COIN System Architecture

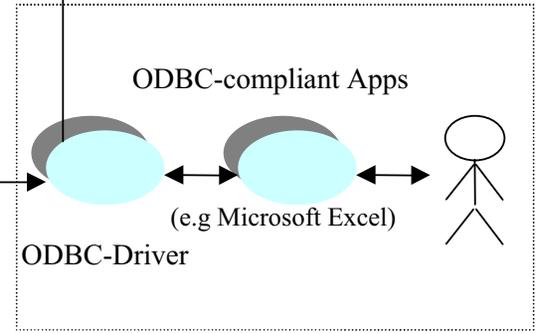
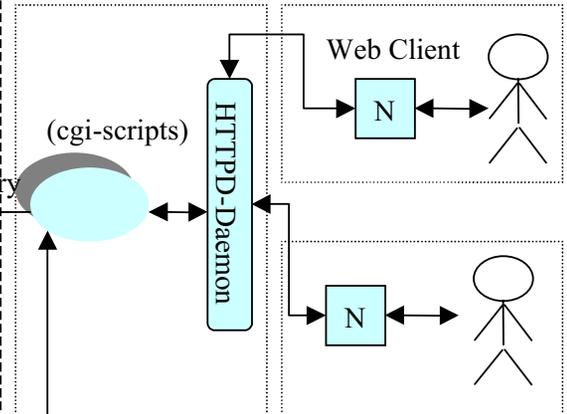
SERVER PROCESSES



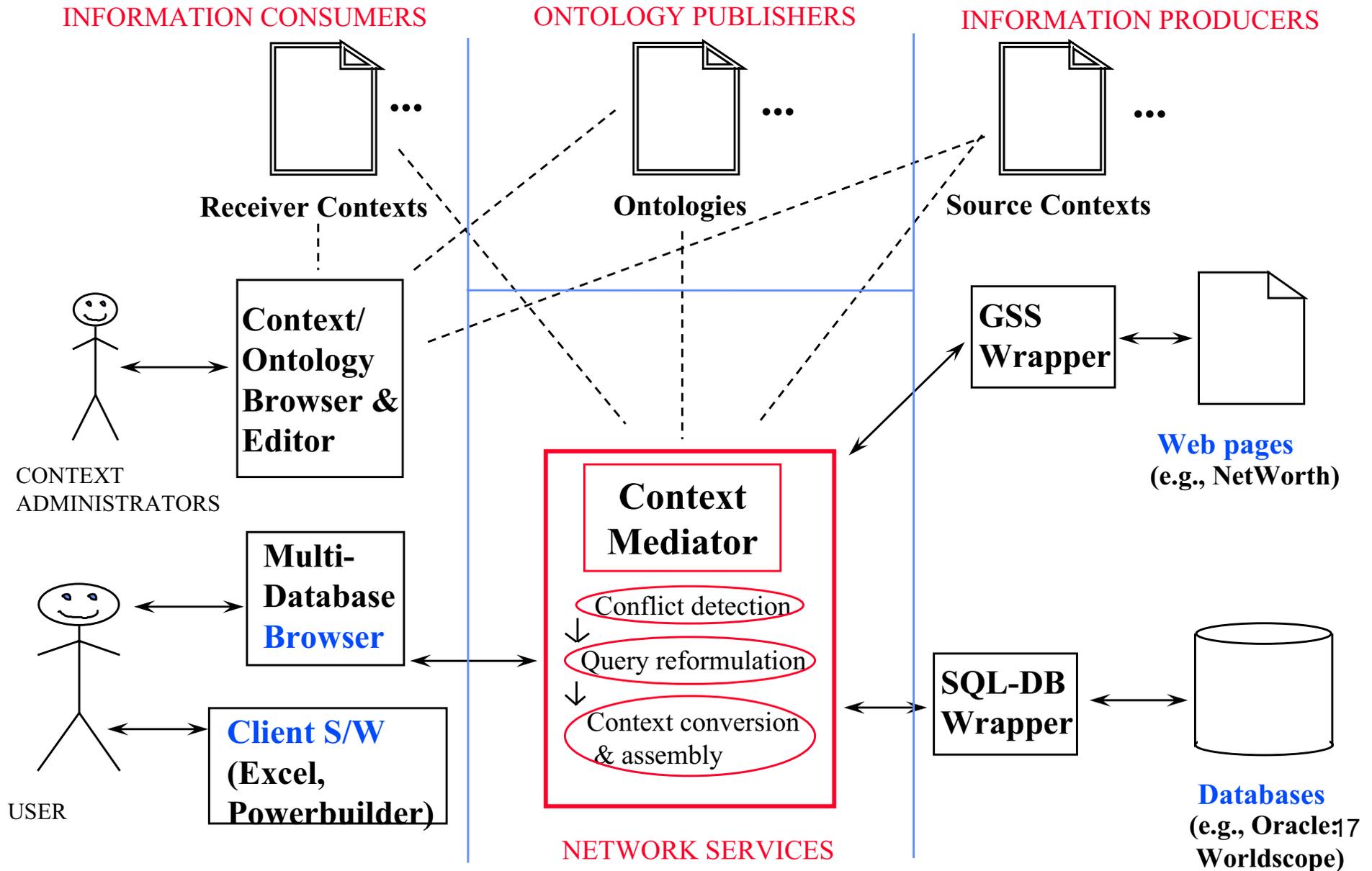
MEDIATOR PROCESSES



CLIENT PROCESSES



Prototype: COIN on World Wide Web



System Demonstration

Single Source Queries with Mediation

Q6. Scenario: Using Context Interchange, the financial analyst can look at the Disclosure data using Datastream Context.

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.

Demonstration – context2.mit.edu

GCMS Demo - Tasc06 SQL

Management: [[ontology](#) | [contexts](#) | [conversions](#) | [sources](#) | [elevations](#) | [integrity constraints](#)]
Currently Effective Internal Reps: [[ontology](#) | [context](#) | [conversion functions](#) | [sources](#) | [elevation](#)]

[[Start From Datalog](#) | [Enable Timing Log](#)]

Queries

- [New01](#)
- [New02](#)
- [New03](#)
- [New04](#)
- [New05](#)
- [New06](#)
- [New07](#)
- [New08](#)
- [Tasc01](#)
- [Tasc02](#)
- [Tasc03](#)
- [Tasc04](#)
- [Tasc05](#)
- [Tasc06*](#)
- [Tasc07](#)
- [Tasc08](#)
- [Tasc09](#)
- [Tasc10](#)
- [Tasc11](#)

Description select DiscAF.LATEST_ANNUAL_DATA, DiscAF.NET_INCOME from DiscAF where DiscAF.COMPANY_NAME = 'DAIMLER-BENZ';

SQL

```
select DiscAF.LATEST_ANNUAL_DATA, DiscAF.NET_INCOME  
from DiscAF  
where DiscAF.COMPANY_NAME = 'DAIMLER-BENZ';
```

Context

Datastream ▾

Stage

- Naive Datalog SQL Translation
 Context Sensitive Datalog Execution
 Conflict Detection
 Mediation

Conflict Detection and Mediation

Conflict Detection

Mediation

Submit

Reset

Result

SemanticType	Column	Source	Modifier	Modifier value in source context	Modifier value in target context	Conversion Function
companyName	Name	DiscAF(Name, FYEnd, Shares, Income, Sales, Assets, Incorp)	format	c_ds : ds_name	c_dt : dt_name	name_map(V4, V3, V2, V1)
companyFinancials	Income	DiscAF(Name, FYEnd, Shares, Income, Sales, Assets, Incorp)	scaleFactor	c_ds : 1	c_dt : 1000	V5 is V4 / V3, V2 is V1 * V5
date	FYEnd	DiscAF(Name, FYEnd, Shares, Income, Sales, Assets, Incorp)	dateFmt	c_ds : American Style /	c_dt : European Style -	datexform(V4, V3, V2, V1)

Mediation

Submit

Reset

Mediated Query in Datalog

Result

```
answer('V9', 'V8') :-
  'DiscAF'('V7', 'V6', 'V5', 'V4', 'V3', 'V2', 'V1'),
  datexform('V6', "American Style /", 'V9', "European Style -"),
  'V8' is 'V4' * 0.001,
  'Name_map_Dt_Ds'("DAIMLER-BENZ", 'V7').
```

← Date convert
 ← Scale factor convert
 ← Name convert

Mediated SQL Query & Result

Mediated SQL Query

```
select datexform.date2, discaf.net_income*0.001
from (select company_name, latest_annual_data, current_shares_outstanding, net_income, net_sales, total_as
      from discaf) discaf,
      (select date1, 'American Style /', date2, 'European Style -'
      from datexform
      where format1='American Style /'
      and format2='European Style -') datexform,
      (select 'DAIMLER-BENZ', ds_names
      from name_map_dt_ds
      where dt_names='DAIMLER-BENZ') name_map_dt_ds
where discaf.company_name = name_map_dt_ds.ds_names
and discaf.latest_annual_data = datexform.date1
```

Adjust scale factor

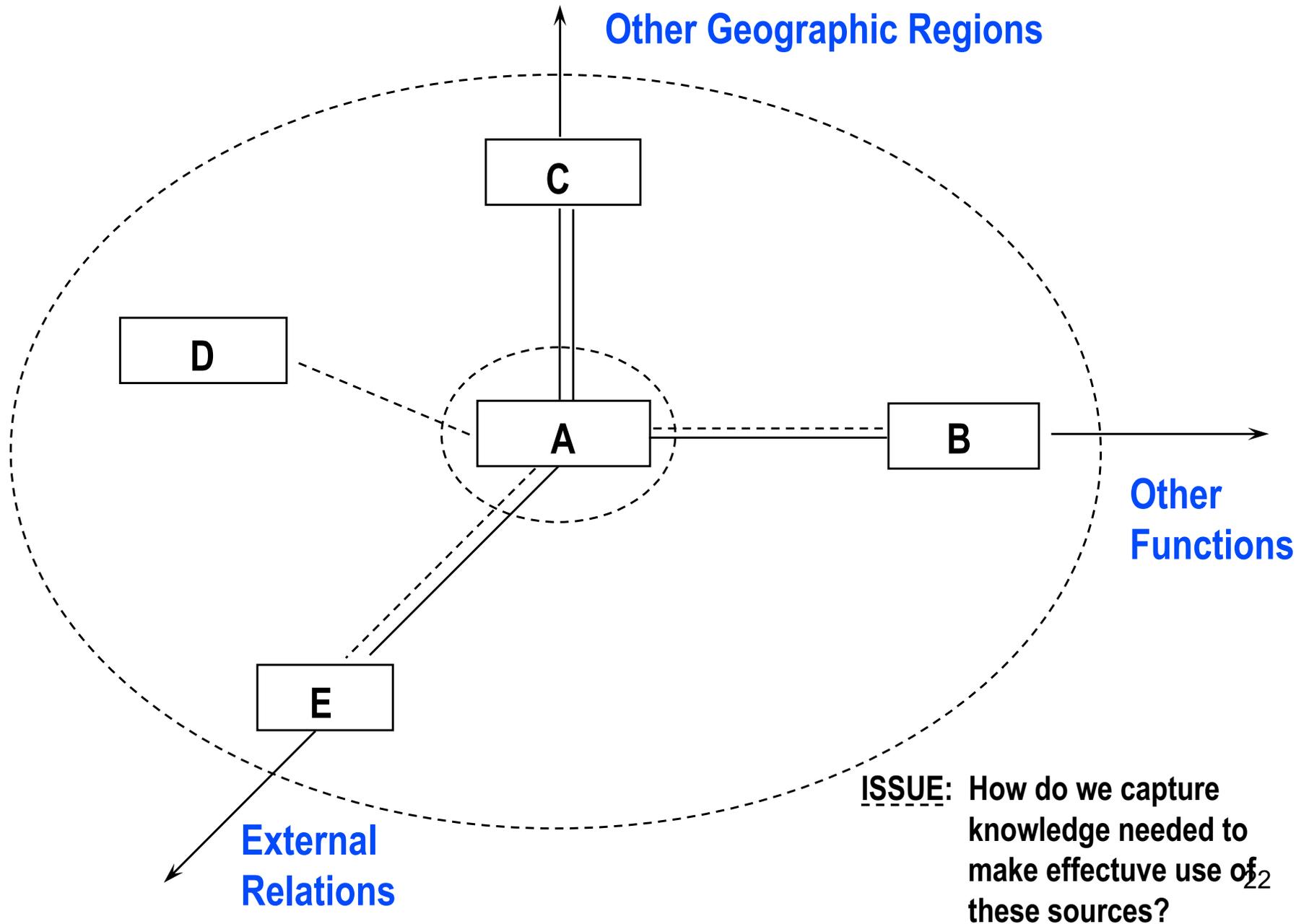
Date format conversion

Name conversion

Final results – from Disclosure but in Datastream context

Result	DiscAF.LATEST_ANNUAL_DATA	DiscAF.NET_INCOME
	31-12-93	615000

Expanded scope of interactions



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.

Source: David Chandler, *The Campaigns of Napoleon*, New York: MacMillan 1966, pg. 390.

Summary

- Tremendous opportunity to gather information
- Need to overcome many context challenges
- Context-type “metadata” plays a critical role
- W3C Resource Description Framework (RDF) is moving in that direction ...
- Some joint experiments at MIT Sloan:
 - Merrill-Lynch: Trader assist & Global risk
 - Price-Waterhouse: B-to-B e-commerce
 - BSCH/Suruga/Fleet: Universal Financial Aggregator
 - Primark: Financial information integration
 - MITRE: Context Mediation prototype
 - AF: Logistics and repair