Overcoming Ontological Conflicts in Information Integration

Aykut Firat, Stuart Madnick and Benjamin Grosof
MIT Sloan School of Management
{aykut, smadnick, bgrosof}@mit.edu

Presentation for
ICIS 2002
Barcelona, Spain
December 15-18
Death. Taxes. Integration.

• DATEK IT INTEGRATION CHALLENGES AMERITRADE
  “...consolidate data centers, develop a single online-trading Web site and set up unified systems…” (CW)

• DEFENDING THE U.S.
  “..a sea of unconnected islands of information technology..” (EW)

• A FRANCO-GERMAN PARTNERSHIP IN THE COCKPIT
  “EADS: a big challenge to forge an integrated, cross-border group within Europe…” (FT)

• MARKET MAKES IT PRIORITY IN DRUG MERGER
  “Competitive pressures make it priority for (Glaxo Wellcome and Smith-Kline Beecham) to combine their systems ...” (CWorld).
Death. Taxes. Integration.

- DATEK IT INTEGRATION CHALLENGES AMERITRADE
  "...consolidate data centers, develop a single online-trading Web site and set up unified systems..." (CW)

- DEFENDING THE U.S.
  "..a sea of unconnected islands of information technology.." (EW)

- A FRANCO-GERMAN PARTNERSHIP IN THE COCKPIT
  "EADS: a big challenge to forge an integrated, cross-border group within Europe..." (FT)

- MARKET MAKES IT PRIORITY IN DRUG MERGER
  "Competitive pressures make it priority for (Glaxo Wellcome and Smith-..." (CWorld).
Roadmap

- Key Concepts
- Case Study
- Solution Methodology
- Test Example
- Concluding Remarks
Semantic Integration

Key Concepts

Disconnected Sources  Physical Integration  Semantic Integration
Ontology

Key Concepts

- “Specification of a conceptualization”

A snapshot from a financial ontology
Ontological Heterogeneity

**Key Concepts**

- **Bloomberg Ontology**
  - Company
  - Financials
  - Country
  - Earnings
  - Currency
  - **PE Ratio**

- **Market Guide Ontology**
  - Company
  - Financials
  - Country
  - Currency
  - Earnings
  - **PE Ratio**

**Trailing**
- Price
- Earnings in the last 4 quarters

**Ontological Conflict**

**Forward Looking**
- Price
- Earnings in the last 3 quarters + forecasted next quarter
Found Three Different Types of Heterogeneities

- Data Level
- Ontological
- Temporal
Data Level Heterogeneities

*Definition:* **Same** entity **different** representations

### Case Study

**Sales (FIAT)**

<table>
<thead>
<tr>
<th>COMPANY/DATASOURCE</th>
<th>FIAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hoovers</td>
<td>48,741.0 (Dec99)</td>
</tr>
<tr>
<td>Yahoo</td>
<td>N/A</td>
</tr>
<tr>
<td>Market Guide</td>
<td>45,871.5 (Dec99)</td>
</tr>
<tr>
<td>Money Central</td>
<td>49,274.6 (’99)</td>
</tr>
<tr>
<td>Corporate</td>
<td>57,603,000,000</td>
</tr>
<tr>
<td>Information</td>
<td>(’00)</td>
</tr>
<tr>
<td>World Scope</td>
<td>93,719,340,540 (99)</td>
</tr>
<tr>
<td>Disclosure</td>
<td>48,402,000 (99)</td>
</tr>
<tr>
<td>Primark Review</td>
<td>51,264 (99)</td>
</tr>
</tbody>
</table>

**Currency:** Local Scale Factor: 1000

**Currency:** USD Scale Factor: Millions

Diagram:

- **WorldScope:** 93,719,340,540
- **Market Guide:** 45,871.5
- **Sales (FIAT):**
  - Currency: Local Scale Factor: 1000
  - Currency: USD Scale Factor: Millions

**Scale Factor:**
- Local: 1000
- USD: Millions
Data Level Heterogeneities

*Definition*: Same entity different representations

### Case Study

<table>
<thead>
<tr>
<th>Sales (FIAT)</th>
<th>COMPANY/DATASOURCE</th>
<th>FIAT</th>
<th>DAIMLER</th>
<th>CHRYSLER</th>
<th>BENZ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hoovers</td>
<td>48,741.0 (Dec99)</td>
<td>152,446.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yahoo</td>
<td>N/A</td>
<td>131.4B</td>
<td>145,076.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Market Guide</td>
<td>45,871.5 (Dec99)</td>
<td>152.4 Bil</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Money Central</td>
<td>49,274.6 (*'99)</td>
<td>162,384,000,000</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Corporate Information</td>
<td>57,603,000,000 (‘00)</td>
<td>162,384,000,000 (‘00)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>WorldScope</td>
<td>93,719,340,540 (99)</td>
<td>257,743,189 (98)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Disclosure</td>
<td>48,402,000 (99)</td>
<td>131,782,000 (98)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Primark Review</td>
<td>51,264 (99)</td>
<td>71354 (97)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*WorldScope Market Guide Sales Data*

- Currency: Local
- Scale Factor: 1000
- Currency: USD
- Scale Factor: Millions
Ontological Heterogeneities

Definition: Different entity type definitions and/or relationships

\[
\text{Price/PE\_Trailing} = \frac{\text{Price}}{\text{PE\_Forward} - \text{Forecasted Quarterly Sales}(t+1) + \text{Quarterly Sales}(t-3)}
\]
Temporal Heterogeneities

Definition: Entity values or definitions belong to different times, or time intervals.

Case Study

Data Level

For Exxon:


Ontological
Our Challenge

How to represent and reason with semantic heterogeneities?

Why challenging?

- Machine to machine communication: AI
- Declarative, scalable, extendible solution required
- Modeling: combination of art and science
- Includes NP-Complete problems (e.g. Source Selection)
Approach

Context Interchange (COIN): Data Level

• Research undertaken by Sloan IT Ph.D. Cheng-Goh
• Logical Framework: COIN Data Model + Abductive Logic Programming
• Loosely coupled approach to semantic integration

Extended COIN: Ontological

• Extended Data Model + Symbolic Equation Solving Techniques
• Based on Constraint Logic Programming
• Source Selection
• Ontology Merging
COIN Architecture

Solution Methodology

Scale factor: 1000
Currency: USD...

Context Axioms

Currency: USD

Query

e.g Sales of FIAT

Scale factor: 1000
Currency: Local...

Context Axioms

DataStream

Semistructured Data Sources (e.g. XML)

Context Axioms

DBMS

Context Axioms

Market Guide

Optimizer/Executioner

Conversion Library e.g. Currency Converter

Context Mediator

Mediated Query

Ontology

Currency: ITL

COIN Architecture
ECOIN Architecture

\[
\text{Price/PE}_{\text{Trailing}} = \text{Price/PE}_{\text{Forward}} - \text{Forecasted\_Quarterly\_Sales}(t+1) + \text{Quarterly\_Sales}(t-3)
\]

Solution Methodology

- **Context Axioms**
  - PE Ratio: Trailing
  - Scale factor: 1000
  - Currency: USD

- **Context Axioms**
  - PE Ratio: Forward
  - Scale factor: 1000
  - Currency: Local

- **Context Axioms**
  - PE Ratio: Forward
  - Scale factor: 1000
  - Currency: Local

- **Context Axioms**
  - PERatio of FIAT

**Market Guide**
- DBMS
- Context Axioms

**DataStream**
- Semistructured Data Sources
  - (e.g. XML)
- Context Axioms

**Optimizer/Executioner**
- Constraint Engine
- Symbolic Equation Solver
- Equation Library
- Conversion Library
  - e.g. Currency Converter

**Mediator**
- Mediated Query

**Context**
- Ontology
- Query
  - e.g. PERatio of FIAT

**Conversion Library**
- e.g. Currency Converter

**Solution Methodology**

- PE Ratio: Trailing
  - Scale factor: 1000
  - Currency: USD

- PE Ratio: Forward
  - Scale factor: 1000
  - Currency: Local

- PE Ratio: Forward
  - Scale factor: 1000
  - Currency: Local
Key Link: ECOIN & Semantic Web

Solution Methodology

Semantic Web Architecture (Tim Berners-Lee, W3C)

- Trust
- Proof
- Logic
- Rules
- Ontology

ECOIN

Tagged data: XML + Namespaces + XML Schema

- RDF+ RDF Schema
- Ontology
- Rules
- Logic
- Proof
- Trust

Digital Signatures

Unicode

URIs
E-Business Application

**Context Mediator**

**Price Equations**

**Query**
Prices of Products Cheaper in eToys compared to Kid’s World

**Test Example**

<table>
<thead>
<tr>
<th>Product Code</th>
<th>eToys</th>
<th>Kid’s World</th>
</tr>
</thead>
<tbody>
<tr>
<td>pokemon</td>
<td>17</td>
<td>13.3</td>
</tr>
<tr>
<td>starwars</td>
<td>45</td>
<td>30.1</td>
</tr>
</tbody>
</table>

Results

<table>
<thead>
<tr>
<th>Product Code</th>
<th>eToys</th>
<th>Kid’s World</th>
</tr>
</thead>
<tbody>
<tr>
<td>123456</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>234567</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Concluding Remarks

- We developed ECOIN
  - Symbolic Equation Solving Techniques
  - Proof of the concept prototype
- First system to deal with both data level and ontological heterogeneities
- Temporal in our agenda next
- Contribute to the global semantic interoperability efforts
Problems addressed

Many Experts in Physical and Tightly-Coupled Integration

- Content Aggregation: Yodlee,..
- Workflow Integration: IBM, Microsoft,..
- Enterprise Application Integration: SAP, WebMethods, Siebel
- B2B: Ariba, CommerceOne,…
- Portals: Vignette, Broadvision,..

Our focus is on Semantic and Loosely-Coupled Integration