Towards a Management Framework for Data Semantic Conflicts: A Financial Applications Perspective

Raphael Yahalom & Stuart E. Madnick

Sloan WP# 3958 CISL WP# 97-05
April 1997

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

Towards a Management Framework for Data Semantic Conflicts: A Financial Applications Perspective

Raphael Yahalom* and Stuart E. Madnick

MIT- Sloan School of Management
{yahalom, smadnick}@mit.edu

April 25th, 1997

Abstract: An application-oriented analysis of data semantic challenges is presented. The analysis serves as a basis for developing a framework for effectively managing such challenges. The paper focuses on financial information providers as representative cases. A model representing relationships between attributes is proposed and is used for presenting and discussing certain semantic conflicts and their business implications. Directions of current work which are based on that model and which aim to advance towards an effective management framework are outlined.

1 Introduction

It is well known that *data semantics* which are associated, often implicitly, with data in a database may be very source-specific, and thus that *data semantics conflicts* may result when users or applications access multiple data sources (e.g. [SL90, LA87, DH84]).

In many cases such database sources are autonomous, indeed potential competitors in some information market, and so may or may not have incentives to adhere, often at significant cost, to externally imposed constraints aiming to improve interoperability.

Thus semantic mediation approaches, such as the *Context Interchange* project ([Goh96, BFG$^+$97b, BFG$^+$97a, GMS94]) which provides a foundation for the current work, are based on the assumption that each source may maintain its semantic unique characteristics and aim to provide dynamic automatic semantic reconciliation capabilities to be invoked as required.

However many semantic conflicts represent inherent application-specific, often complex relationships, and so issues related to their resolution may not be straight-forward or globally agreed upon.

*On leave from the Hebrew University of Jerusalem - School of Business Administration

For example, assume that a user accessed from an information source the *Net-Sales* value for British Petroleum Plc, as part of a comparative analysis she is performing. That source provided the value 56,065,397 as the *Net-Sales* value for British Petroleum for 1995. Obvious questions that come up are ones such as *Is that figure in British Pounds or US Dollars?* and *Which scaling factor is used?*. However more subtle but potentially equally important questions are ones such as *Which definition of Net-sales is used here? Are items such as VAT taxes or Excise taxes included in or excluded from that figure ? Are Other-Income items such as Interest-income included in or excluded from that figure?*

Similar issues need to be considered whenever such a value is used in some application process, and in particular when different values are used within the same process. Assume that our user also obtains from a different source the Net-Sales value of 64,767,000,000 for Mobil Corp for 1995. Without addressing questions such as the ones above, it is not clear that these two values may be consistently used for a meaningful comparison of the relative performance of these two companies.

In this paper we aim to motivate, and advance towards a more flexible data semantic conflict management framework. Within such frameworks complex data-semantics relationships may be captured. Effective reconciliation and explication strategies may be selected based on the characteristics of, and implications to, the corresponding decision-processes.

Although a similar approach is applicable in various application domains, we chose to focus here on the important, interesting, and rich domain of financial databases.

We describe some relevant general characteristics of certain types of financial database environments and present a preliminary model which associates potential semantic conflicts with expected decision process implications. That model is used to present and analyze two example cases: one constructed specifically to demonstrate the points raised in this paper, and another which is based on some of our empirical findings when analyzing a large financial information service provider. The Context Interchange project, and particular extensions to it currently underway, provides a promising foundation on which to design effective semantic conflict management frameworks.

The rest of this paper is organized as follows. In the next section previous relevant literature is surveyed. In section 3 we provide an overview of certain characteristics of financial information sources environments. In section 4 we present a general model which associates between attributes relationships and quality of application decisions. Section 5 summarizes some relevant findings resulting from an analysis of a large financial information service provider. In section 6 we outline some potential implications from an organization decision making perspective, and in section 7 we discuss certain system design aspects and the relationship to the Context Interchange prototype. Finally, in section 8 we present some conclusions and directions of current work.

2 Previous Data Integration Approaches

Over the last few years, a significant body of work has focused on issues related to the integration of multiple potentially-autonomous information sources. Such systems have commonly been referred to as *federated database systems* (e.g. [SL90]), *heterogeneous databases* (e.g. [SYE+90]), or *multidatabase systems* (e.g. [BHP92]).

We focus particularly on the issue of *Data Heterogeneity* across different information sources. In [Goh96], Data Heterogeneity is classified to three categories:

- *Schematic Conflicts* (e.g. [KS91, KLK91]) which include conflicts such as *Data Type conflicts* - different primitive data types across systems; *Labelling conflicts* - mismatches associated with different naming conventions of schema elements across sources (e.g. [KS91, Mot87, DH84]); *Aggregation conflicts* - different clustering of attributes across sources (e.g. [SS77]); and *Generalization conflicts* - different relation structures across sources (e.g. [KS91]).
- *Semantic Conflicts* (e.g. [SSR94, SM91, BFG+97b, BFG+97a]) which include conflicts such as *Naming conflicts* - different ways of naming entity values across sources; *Scaling and Units conflicts* - different measurement conventions across sources; and *Confounding conflicts* - confounding of distinct concepts across sources [Goh96].
- *Intensional Conflicts* which include conflicts such as *Domain conflicts* - differences in the scope of the domain across sources and *Integrity Constraints conflicts* - differences in integrity constraints enforced across sources (e.g. [AKWS95]).

Various approaches and prototype systems for achieving interoperability between multiple information sources have been developed.

In general, it is often useful to categorize these efforts by positioning them on a spectrum of *tight-coupling to loose-coupling* strategies. In principle, strategies which are oriented towards tight-coupling are often based on the premise that conflicts are detected *a priori* and consequently that interoperability may be facilitated by relying on some flavor of predefined views, or *shared schemas*, between information sources. Important approaches and prototype systems which exhibit certain tight-coupling characteristics include for example [LR82, BT85, TBD+87, ASD+91, AK92, CHS91].

On the other hand, in strategies which are oriented towards loose-coupling users are exposed to inter-source conflicts and are thus responsible for the detection and resolution of such conflicts. Important examples of efforts which are based on loose-coupling approaches include [LA87, KL88, Lit92, Wol89].

The TSIMMIS project at Stanford ([GMPQ+95]) focuses on environments in which many of the data sources are unstructured and is based on the premise that many of the challenges of information integration cannot be fully automated and require human intervention.

Strategies such as the Context Interchange (COIN) scheme on which this work is based [Goh96, BFG$^+$97b, BFG$^+$97a, GMS94] can be seen as a hybrid approach, attempting to benefit from particular advantages associated with certain amount of global sharing, but with minimal sacrifice with respect to local flexibility and global scalability.

In particular, within the COIN framework each source as well as each user is associated with a well specified *context* representing certain relevant semantic characteristics. Thus conflicts are detected and resolved dynamically when a user interacts with relevant sources by mediation processes which rely on the corresponding contexts.

In the following sections we examine some particular characteristics and conflicts which are associated with financial information sources and certain application challenges they represent. The COIN approach with certain extensions to it provides a promising framework for addressing many of these challenges.

3 Financial Information Sources

Open global communication networks provide enhanced opportunities for users to access a large number of information services. Increasingly it is becoming cost-effective for users, such as decision-makers in financial domains, to rely on multiple different information sources as support for a decision-making process, leveraging on the relative advantage of each source.

Financial information services contain values of given economic-related attributes of given entities at given points in time. In this paper the entities we focus on are companies, however in principle an analogous discussion applies to other *financial entities* such as financial instruments, real estate assets, or private individuals.

Multiple financial information services often diverge in their contents even in cases in which the scope of target entities (companies) is similar. Such discrepancy is a result of various possible reasons, for example:

1. Different attributes are represented.
2. Attribute values correspond to different points in time.
3. Different attribute value representation formats are used.
4. Certain attributes are initialized from different originating entities that may be associated with various degrees of reliability.
5. Certain attributes reflect subjective assessment by different originating entities.
6. Certain attributes definitions may be adjusted to reflect particular financial practices or local accounting conventions.

7. Certain attributes definitions may be adjusted to enhance the value of the information item and to facilitate comparisons of financial attributes within and across national boundaries and operation domains.
8. Operational errors may corrupt certain values.

Historically each information service evolved to reflect the requirements of particular target applications and particular user communities. In turn, each user and each application routinely interacted only with one or very few information sources.

In the modern networked environments of today, multiple information sources are often easily and inexpensively accessible to users and applications. Such potential access to multiple information sources may lead in certain cases to increased information market efficiency: higher quality of user processes, and increased productivity.

However such an integrated and ad-hoc access to multiple sources may expose users and applications to significant inter-source discrepancies. The challenge of managing such conflicts successfully is fundamental. Not only may such conflicts limit the ability to exploit the full potential of the multi-source accessibility, but in some cases such inconsistencies may result in significant losses.

3.1 Integrated Access to Multiple Sources

Users may gain significant advantages by the ability of financial applications to retrieve and integrate information from multiple information sources:

1. Additional sources may contain additional attributes associated with some target company.
2. Multiple sources may provide access to information on a larger scope of companies, enabling effective searches, comparisons, or aggregative analysis.
3. Multiple sources may enable more effective cross-checks and reconciliation of information, dealing with cases of varying degrees of source reliability and information quality.
4. Some attributes values are based on certain subjective assessments - multiple such assessments may increase the resulting level of confidence.
5. Multiple source accessibility implies potential higher availability of information in the face of component failures.

Such advantages for integrated access does not imply that multiple information sources will necessarily consolidate into uniform centrally-managed global ones:

- Different information services may be associated with independent business players in a competitive market.

- Different information services may be addressing different local business and user needs (e.g. correspond to different national accounting conventions, etc.).
- It may be advantageous to maintain information services' developed brand-names, as information products relies so heavily on reliability and reputation assumptions associated with a provider.
- The cost of data conversion from multiple local data formats and models to a single global one may be prohibitive.
- The cost associated with organizational restructuring to reflect different operations and data maintenance tasks may be significant.
- The requirement to maintain compatibility with a large existing application programs base imposes operational and cost constraints.
- The need to maintain consistency with respect to established user's developed mental conceptual models of the data semantics may constrain various database restructuring efforts.

3.2 Multiple Information Sources

We model each information source as consisting of a set of relations. Each tuple represents an entity in a financial domain. This abstract model does not prevent us from incorporating information sources which are non-relational (e.g. object-oriented ones) or which are semi-structured or unstructured (such as WEB text sources). Rather, it implies that for each non-relational information source, there is an appropriate *wrapper* process which enables to refer to that source as a relational one.

For each relation we consider:

1. The names of the key attributes and of the non-key attributes.
2. The semantics specifications of each attribute, consisting of
   - its semantic definition - what is the precise financial domain based notion that the attribute represents (e.g. Net-Sales excluding or including Excise Taxes?)?
   - its semantic representation - what is the financial domain based representation method used for representing the values of this attribute (e.g. Which currency is used for representing Net-Sales?)?
3. The scope of the relation - which set of entities are represented in that relation.

Consider the following multi-source example environment. This example environment is used in consequent sections of the paper.

<u>Source A</u>

Financial information service containing certain financial attributes on publicly-traded companies above a certain threshold market-capitalization.

(**company**, **date**, net-income, net-sales, employees, ...)

A.company

- *Semantic Definition:* traded company unique identifying name
- *Semantic Representation:* local company naming convention

A.date

- *Semantic Definition:* year identifier for annual data (closing day)
- *Semantic Representation:* DD/MM/YY (e.g. 31/12/95)

A.net-income

- *Semantic Definition:* income after all operating expenses and extraordinary items for that company in the year which is represented by *date*
- *Semantic Representation:* US Dollars , scale-factor of 1000

A.net-sales

- *Semantic Definition:* gross sales and other operating revenue less discounts, returns, and allowances for that company in the year represented by *date*, **excluding** excise taxes
- *Semantic Representation:* US Dollars , scale-factor of 1

A.employees

- *Semantic Definition:* number of full-time employees of that company and all its subsidiaries in the day represented by *date*
- *Semantic Representation:* scale-factor of 1

......

Source B

Financial information service containing certain financial attributes on publicly-traded USA industrial companies.

(**company**, **date**, net-revenues, income, emps, ceo, ...)

B.company

- *Semantic Definition:* traded company unique identifying name
- *Semantic Representation:* local company naming convention

B.date

- *Semantic Definition:* year identifier for annual data (closing day)
- *Semantic Representation:* DD/MM/YY (e.g. 31/12/95)

B.net-revenues

- *Semantic Definition:* gross sales and other operating revenue less discounts, returns, and allowances, **including** excise taxes, for that company in the year represented by *date*
- *Semantic Representation:* US Dollars , scale-factor of 1000

B.income

- *Semantic Definition:* income after all operating expenses and extraordinary items for that company in the year which is represented by *date*
- *Semantic Representation:* US Dollars , scale-factor of 1000

B.emps

- *Semantic Definition:* number of full-time employees of that company on the day represented by *date*, **excluding** the employees of subsidiaries.
- *Semantic Representation:* scale-factor of 1

B.ceo

- *Semantic Definition:* the name of the company CEO on the day represented by *date*
- *Semantic Representation:* local naming convention

......

Source C

Financial information service containing formatted data fields corresponding to SEC 10K report filings.

(**comp**, **date**, net-sales, net-income, investment-income, employees, excise-tax, ...)

C.company

- *Semantic Definition:* traded company unique identifying name
- *Semantic Representation:* local company naming convention

C.date

- *Semantic Definition:* year identifier for annual data (closing day)
- *Semantic Representation:* DD/MM/YY (e.g. 31/12/95)

C.net-sales

- *Semantic Definition:* gross sales and other operating revenue less discounts, returns, and allowances **including** excise taxes, for that company in the year represented by *date*
- *Semantic Representation:* US Dollars , scale-factor of 1000

C.net-income

- *Semantic Definition:* income after all operating expenses and extraordinary items for that company in the year which is represented by *date*
- *Semantic Representation:* US Dollars, scale-factor of 1000

C.investment-income

- *Semantic Definition:* realized gain or loss from the sales of investment securities for that company in the year which is represented by *date.*
- *Semantic Representation:* US Dollars , scale-factor of 1

C.employees

- *Semantic Definition:* Average number of full-time employees of that company in the corresponding year, excluding the employees of subsidiaries.
- *Semantic Representation:* scale-factor of 1

C.excise-tax

- *Semantic Definition:* The amount of excise-taxes received by that company in the year which is represented by *date*
- *Semantic Representation:* US Dollars, scale-factor of 1

......

Source D

Information service containing company ownership information for worldwide companies.

(**company**, **date**, subsidiary-of, emps, total-assets, ...)

D.company

- *Semantic Definition:* traded company unique identifying name
- *Semantic Representation:* local company naming convention

D.date

- *Semantic Definition:* year identifier for annual data (closing day)
- *Semantic Representation:* MM/DD/YY (e.g. 12/31/95)

D.subsidiary-of

- *Semantic Definition:* the name of the parent company if this company is a subsidiary.
- *Semantic Representation:* local company naming convention

D.emps

- *Semantic Definition:* number of full-time employees of that company in that date, excluding the employees of its subsidiaries.
- *Semantic Representation:* scale-factor of 1

D.total-assets

- *Semantic Definition:* the sum of total current assets, long term receivables, investment in unconsolidated subsidiaries, and other investments.
- *Semantic Representation:* Local currency , scale-factor of 1000

......

Thus even in cases in which a company is represented in more than one source due to different semantic specifications of attributes the corresponding values may indeed be different. For example, IBM may be represented in both source A and B, but its number of employees at the last day of 1995 will be 290,215 according to source A and only 225,347 for the same day, according to source B (which excludes subsidiaries). Similarly, Exxon may be represented in both source A and source C, but its Net-Sales values for 1995 will be 121,804,000 in source C and only 107,893,000 for the same year according to source A (which excludes excise-taxes).

In the next section, this example environment is used for demonstrating certain classes of semantic conflicts.

4 Semantic Conflicts and Application Consistency

We present a general model for representing different classes of inconsistencies between different information sources. This model and certain refinements of it will be used as a framework for presenting and discussing various real-world information services inconsistencies, and their potential implications.

Consider any two attributes x and y each from a different information sources. Each attribute is associated with an *attribute name* (AN) at each source (AN_x and AN_y denote respectively the names of attribute x and y at their respective sources). Each attribute is also associated with a *attribute semantics specifications* (AS) at each source (AS_x and AS_y denote respectively the semantic specifications of x and y at their respective sources. Note that the semantic specifications contain both a semantic definition and a semantic representation components).

Thus for any two attributes there are four different AN-AS relationship possibilities. Each such possibility has different inter-source access consistency implications.

The following matrix summarizes these possibilities and their implications. Intuitively, similar attribute names should normally correspond to attributes with the same semantic specifications. Different names associated with attributes with the same semantic specifications may hide the correspondence, potentially limiting certain types of user application inferences. Conversely, similar names in cases in which the underlying attributes are associated with different semantic specifications may potentially mislead user applications , and result in wrong inference conclusions.

	Same AN	**Different AN**
Same AS	Consistent Conclusions (case 1)	Partial Conclusions (case 3)
Different AS	Wrong Conclusions (case 4)	Consistent Conclusions (case 2)

Consider an example application environment. An analyst is performing a background analysis on certain companies in the energy industry, prior to a significant investment decision. In particular the analyst is examining and comparing certain financial performance values for certain American and European companies in that industry. Assume the analyst has access to information sources A, B,C, and D in section 3.

Case 1: The analyst wishes to compare the net-income in 1996 values for two companies $comp_1$ and $comp_2$. Source A's scope includes $comp_1$ and source C's scope includes $comp_2$. This case represent the "Same AN - Same AS" scenario: *A.net-income* and *C.net-income* have the same attribute name and the same semantic specification. The analyst may obtain consistent comparison results by formulating a query that simply retrieves the corresponding values.

Case 2: The analyst now wishes to compare the investment-income in 1996 values for the two companies $comp_1$ and $comp_2$. Source A does not include an attribute corresponding to *C.investment-income* so that the analyst may consistently conclude that it does nor have access to the required value for $comp_1$. In particular, the analyst will not attempt to compare the *C.investment-income* with an attribute such as *A.net-income* whose different *AN* suggests a different associated *AS*.

Case 3: The analyst now wishes to compare the net-sales in 1996 values for two companies $comp_3$ and $comp_4$. Source B's scope includes $comp_3$ and source C's scope includes $comp_4$. This case represent the "Different AN - Same AS" scenario: *B.net-revenues* and *C.net-sales* have the same semantic specification. However, the analyst's query mechanism may fail to provide the appropriate comparison due to the naming differences, potentially leading to a situation of a missed opportunity : required data which is obtainable but not provided.

Case 4: The analyst wishes to compare the net-sales in 1996 values for two companies $comp_5$ and $comp_6$. These companies are included in the scope of source A and source C, respectively.

The relevant attributes have the same *AN* *A.net-sales* and *C.net-sales.* However these attributes have a different *AS*. In particular, *A.net-sales* excludes excise taxes and *C.net-sales* does not. The magnitude of such a discrepancy may in some cases be significant, leading to a potentially wrong comparison conclusions by the analyst. This represent an example of the "Same AN - Different AS" scenario.

Finally the analyst wishes to compare the total number of employees of $comp_5$ (source A) with that of $comp_4$ (source C). Again, this represent an example of the "Same AN - Different AS" scenario. *A.employees* and *C.employees* have the same *AN* but a different *AS*. Such a discrepancy related to the number of employees in subsidiaries may have a significant impact on the comparison results and may in turn lead to inconsistent conclusions.

4.1 Computation-Path Relationships

As will be demonstrated below, it is often useful to further classify the notion of *Different AS* to two sub-categories.

Such a classification requires a new notion - a *computation-path relationship.*

Definition: *Let an environment contain information sources* $\mathcal{S}_{\mathcal{M}}$, ... , $\mathcal{S}_{\mathcal{N}}$. *Consider any two attributes* AN_x *and* AN_y *with different semantic definitions* AS_x *and* AS_y.

Attribute AN_y *is said to be* **computation-path related** *to attribute* AN_x *in that environment,* *if there exist a computation sequence, involving only data values contained in $\mathcal{S}_{\mathcal{M}}$, ... , $\mathcal{S}_{\mathcal{N}}$, which enables the conversion of any value of AN_x (which has a semantic specifications AS_x) to a single corresponding value associated with a semantic specification AS_y.*

For example, the attribute *A.net-sales* is *computation-path related* to attribute *C.net-sales* in our example environment $\mathcal{S}_{\mathcal{A}}$, $\mathcal{S}_{\mathcal{B}}$, $\mathcal{S}_{\mathcal{C}}$, $\mathcal{S}_{\mathcal{D}}$. Any value of *C.net-sales* may be converted to a single value associated with the semantic specification of attribute *A.net-sales* by obtaining from source C the corresponding value of *C.excise-tax* and subtracting it.

Similarly, the attribute *A.employees* is *computation-path related* to attribute *B.emps* in environment $\mathcal{S}_{\mathcal{A}}$, $\mathcal{S}_{\mathcal{B}}$, $\mathcal{S}_{\mathcal{C}}$, $\mathcal{S}_{\mathcal{D}}$. Any value of *B.emps* may be converted to a single value associated with the semantic specification of attribute *A.employees* by obtaining from source D all the corresponding values of employees in the subsidiaries and adding to the value of *B.emps* (assuming that the semantic representations of the company name attributes in these two sources are identical or are mutually convertible).

On the other hand, the attribute *C.employees* is **not** *computation-path related* to attribute *B.emps* in environment $\mathcal{S}_{\mathcal{A}}$, $\mathcal{S}_{\mathcal{B}}$, $\mathcal{S}_{\mathcal{C}}$, $\mathcal{S}_{\mathcal{D}}$. None of the information sources in that environment contain the value of the *average* number of employees in the subsidiaries in that period, which is needed for that conversion computation.

Our classification matrix of the previous section may now be extended as follows.

	Same AN	**Different AN**
Same AS	Consistent Conclusions	Partial Conclusions
Different AS with computation path	(Potentially) Consistent Conclusions	Partial Conclusions
Different AS without computation path	Wrong Conclusions	Consistent Conclusions

In particular, in our example scenario an analyst comparing the net-sales values of two companies by accessing *A.net-sales* and *C.net-sales* may be in a position to obtain consistent results, if the computation-path conversion is performed.

The analyst comparing the number of employees values by accessing *A.employees* and *B.emps* (different *AN*s and different but computation-path related *AS*s) may be prevented from reaching a consistent comparison conclusion, even though such a consistent comparison is obtainable.

5 Semantic Conflicts in Existing Information Services

Various types of semantics conflicts exist between financial information services which are currently accessible via the Internet or via other networks.

We present here some of our analysis findings focusing on a particular information service provider, Primark, which provides a few leading financial information sources. Although Primark serves as a useful case study to highlight various potential inconsistencies and related challenges, such inconsistencies are by no means unique to the Primark environment. We have detected various similar inconsistencies when examining other financial information sources.

Primark (*http://www.primark.com/*) is a global provider of information services to the financial, investments and media industries. Following a series of significant acquisitions, Primark has assembled a number of subsidiaries with well-known brand names in the areas of financial information services. We focused in particular on three financial databases associated with Primark subsidiaries:

1. **Disclosure** collects and distributes financial data on approximately 16,000 US companies and 13,000 international companies.
2. **Worldscope** creates standardized comparable financial information on approximately 11,900 companies in 45 countries.
3. **Datastream** provides comprehensive, historical coverage on over 140,000 securities, financial instruments, and companies from markets worldwide, supporting users in evaluating investment opportunities and corporate performances, comparing international markets, and managing portfolios.

As a major and significant player within the global information services market, Primark represents a rich and important example environment in which to explore the opportunities and challenges associated with the management of multiple information sources.

In order to demonstrate the Primark information services capabilities, the *Primark Investment Research Center* has set up a Web page from which links are available to lists of top 25 US companies and top 25 International Companies in the categories of Net-Sales, Net-Income, Total-Assets , Number of Employees, and Five-Year Growth in Earnings per Share (the page's URL is *http://www.pirc.com/top_companies/*).

Each list was generated based on a different Primark source: the Top US list is based on the Disclosure SEC source (Disc) and the Top International list is based on the Worldscope source (WS).

This Web Demo represents an interesting real application which integrates two separate financial information sources.

Net-Sales

The semantic definition of Net-Sales in WS is the net-revenues excluding *Interest income* and *Other income*, and excluding *Excise taxes* , whereas according to the semantic definition of Disc all of these values are included.

Consequently, there are various discrepancies between the two lists, for example:

- Exxon's value for the same date (12/31/95) is $121,804,000,000 in Disc and only $107,993,000,000 in WS (A difference of 13.811 Billion USD - the amount received as excise-taxes in that year).
- Mobil's value is $73,413,000,000 in Disc and only $64,767,000,000 in WS. According to its US list figure Mobil should be ranked 15th in the International list where it is in fact ranked as 20th.
- Philip Morris' value is $66,071,000,000 in Disc and only $53,139,000,000 in WS. According to its US list figure Philip Morris should have been ranked as 19th in the world. However in fact it is not included at all in the corresponding International top 25 list.

Total Employees

The semantic definitions of Total Employees in WS and Disc differ significantly with respect to inclusion of , among others, non-permanent employees and employees of subsidiaries (in Disc non-permanent employees and employees of subsidiaries are included whereas in WS they are not) and in some instances with respect to whether the value represent the average value over the period or rather an as-of-date value, as demonstrated below.

Consequently, there are various discrepancies between the two lists, for example:

- IBM's value for the same date (12/31/95) is 290,215 in Disc and only 225,347 in WS. According to its US list figure IBM should be ranked 12th in the International list where it is in fact ranked as 19th.
- Kelly Services Inc. ranks 3rd in the US list with a Disc value of 660,600 employees. That value should have placed Kelly Services Inc. 3rd in the International list as well. However it is not ranked at all in the top 25 International list. It turns out that the corresponding

Total Employees in WS is only 5600 (less than 1% of the value in Disc!). As Kelly Services is a player in the employment market, differences in the semantic definition of the notion of an employee of Kelly Services make a substantial difference on the total figure.

- Similarly, Olsten Corp. , another player in the employment market, has a value of 568,800 total employees in Disc (4th in the top 25 US list) and only 8,800 in WS (less than 2%).
- General Motors Corp. is ranked first in both lists but with different Total-Employees values (745,000 in Disc and only 709,000 in WS). The WS value is the average number of employees in 1995 and the the Disc value is the number of employees at the last day of 1995.

Net-Income, Total-Assets, and 5-Year Growth in EPS

Various types of semantic definition differences have also resulted in multiple obvious inconsistencies between the other three pairs of top-25 US and International rankings provided in that WEB site, For example:

- In the top 25 US by Net-Income list (Disc source) Venezuelan Petrolium Inc is ranked number 12 with a value of $3,103,000,000. With that value it should have ranked 15th in the corresponding top 25 International list (WS source) however it does not appear in that list at all.
- Seagram Company Ltd has a Net-Income value of $3,406,000,000 at 12/31/95 according to the Disc source and $3,381,192,000 at the same date according to the WS source.
- The Federal National Mortgage Associations is ranked 17th in the top 25 International companies by Total-Assets (WS source) with a value of $ 315,992,000,000 and is the highest US company on that list. However in the corresponding top 25 US companies by Total-Assets (Disc source) it does not appear at all.
- Rhone-Poulenc Rorer Inc., Amgen Inc., and Weirton Steel Corp. are the three US companies listed in the WS-based top 25 International companies by 5-Year Growth in Earnings per Share (ranked 9th, 11th, and 12th with values of 201.71, 163.90, and 163.07, respectively). However none of these 3 companies appear in the corresponding Disc-based top 25 US companies list at all!
- Similarly, the top company in the Disc-based top 25 US Companies by 5-Year Growth in Earnings per Share, Partnerre Ltd, according to its growth value in that list (474.4) should have been ranked 5th in the corresponding WS-based top 25 International list. However it does not appear on that list at all.

Some of the above attributes with different semantic specifications are actually *computation-path related*, with respect to the Primark information sources. That is, it is possible to convert a value in one semantic specification to the other, using only data which is obtainable from Primark sources.

Other attributes which are semantically inconsistent can become *computation-path related* by a minor extension to the set of available information sources.

For example, an attribute such *Other-Income* which is available in the Disc source can be used in the computation of a Net-Sales according to its WS semantic specifications from a corresponding Disc Net-Sales value.

On the other , the value of *Excise-Taxes* which is also required for that computation may only be available from an external source, such as one providing a formatted representation of all SEC 10K filings values.

Incidentally, as part of our consistency analysis of the information sources, we have detected a number of cases in which the values of an attribute *within* a single source were not consistent with the corresponding , explicit or implicit, semantic specifications of that attribute. Such a conclusion may be reached in some cases by comparing the data source values with the text representation of the original SEC reports filed by the respective companies. For example the value of Total-Employees in WS represented in some cases the *average* number of employees at the corresponding year, and in other cases the number of employees *at* the last day of the year.

In addition to semantic definition differences between corresponding attributes in the different sources there are many cases of semantic representation differences. For example, many of the companies which appear in both the WS and Disc top 25 lists do not have an identical name representation.

Also, in addition to the multiple cases of "Same AN - Different AS" occurrences in the Primark sources, such as the one highlighted by the inconsistencies between the top 25 lists, there are a variety of occurrences of " Different AN - Same AS" or "Different AN - Computation-Path related AS" which, as discussed above, may result in partial conclusions and missed opportunities.

For example, the attribute *Op-Income* in WS corresponds to the attribute *Optg-Income* in Disc, both representing the value of Operating-Income. The attribute *Earned-for-Ordinary* in the Primark source of Datastream corresponds to the attribute *Net-Income* in Disc, but in order to compute a corresponding Net-Income value, the value of another attribute *Extraordinary-Items*, also available in the Datastream source, needs to be added to an Earned-for-Ordinary value.

In summary, a user who has accessed the Primark Web Demo page on November 24th, 1996 would have encountered the following number of different mutual inconsistencies, resulting from "same AN - different AS" situations, in the top 25 companies lists linked to that page :

- Top 25 International/US Companies by *Net-Sales* - **5 inconsistencies**
- Top 25 International/US Companies by *Net-Income* - **5 inconsistencies**
- Top 25 International/US Companies by *Total-Assets* - **1 inconsistency**
- Top 25 International/US Companies by *Total-Employees* - **10 inconsistencies**
- Top 25 International/US Companies by *5 Year Growth in Earnings Per Share* - **23 inconsistencies**

6 Business Decisions Implications

Different discrepancies associated with inconsistencies vary in their magnitude. In the Primark Web-demo example, such discrepancies varied from minor differences between corresponding values, up to a difference of a few orders of magnitude.

Furthermore, the sensitivity of the applications and the corresponding decision processes to data inconsistencies varies. In some cases, in particular in many financial applications, some controlled levels of approximations are inherently part of the decision making process.

However, such approximations or discrepancies need to be controlled and considered explicitly by the decision making process in order to ensure sufficiently high-quality decisions. Various control approaches are possible, each is associated with certain operational trade-offs.

6.1 Semantic Conflicts and Potential Loss

In order to evaluate various alternative approaches for controlling data semantic conflicts it is in some cases important to analyze the *potential loss*, from the user's business perspective, that may result from consequent *inconsistent* or *partial* (*missed opportunities*) conclusions.

Such economic models and analysis methods are currently being developed and are beyond the scope of this paper. Here, we only sketch out certain preliminary observations and aim to provide basic intuition regarding the spectrum of business implications of data semantic conflicts.

A *business decision* is a process which accepts inputs from various processes (e.g. data sources) and generates based on these inputs and its internal state and logic certain actions which cause some state change. Each state is associated with a certain economic value.

Thus with respect to a particular business decision, the *potential loss associated with a semantic conflict* may be defined as the difference between the value of the state generated by that decision with the semantically conflicting inputs, and the highest value of a state that may have been generated by that business decision without that semantic conflict.

For example, assume a decision to invest in a particular company was based, among other things, on a comparison that relied on semantically inconsistent data regarding another company. The potential loss associated with that semantic conflict with respect to this decision is the difference between the value of the investment made and the optimal one that would have been made, given the same circumstances and inputs, but with a semantically consistent comparison.

A potential loss is similarly defined in a case in which the sub-optimal investment decision was taken without access to a potentially useful available data item, which was not accessed due to a semantic conflict (missed opportunity scenario).

6.2 Decision Process Characteristics

A wide variety of business decision process classes may benefit from data originating from multiple sources. In particular, the types of users involved, the sensitivity of the decision process to semantically conflicting data, and the potential loss that may typically be associated with each decision, vary significantly.

Consider decision process classes corresponding to the three general organizational levels:

Operational Level Decisions: These are normally routine low level processes such as consistency checks and information filtering. As most of the multi-source access is routine and pre-defined, ad-hoc semantic-conflicts are likely to be infrequent.

However, there is typically no sufficient professional sophistication at this level to address semantic conflicts when they do arise, for example when new or non-standard sources need to be accessed.

Professional Level Decisions: These normally involve sophisticated users such as analysts and may often benefit from access to a wide variety of information sources. Many of these decisions may be associated with very high economic values and so in some cases with significant to semantic conflicts related potential loss.

Typically the users are expected to be well familiar with their standard data sources, and so semantic conficts are effectively addressed, but such users are likely to be much less familiar with semantic conflicts associated with fall-back data sources (in case of temporary source availability problems) or with new and ad-hoc sources.

Strategic Level Decisions: These normally involve high level managers and focus on longer term periods and wider economic horizons. Decisions at this level typically rely on a very wide variety of information sources, both formal and informal, so the sensitivity of a particular decision to a particular semantic conflict would normally be very low.

> On the other hand, because of the horizon of users at that level they are more likely to rely on a wide-variety of ad-hoc data sources, and be much less familiar with corresponding subtleties such as precise local semantic definitions.

Thus, approaches for explicating and resolving semantic conflicts need to be sufficiently flexible to support a wide variety of classes of users, processes, and potential loss characteristics. In particular, the cost (in terms of conceptual complexity and execution performance) of a semantic conflict management approach should be fine-tunable to reflect potential loss and business decision characteristics.

For example consider a case in which a user is performing a comparison of the Net-Sales performance of two companies. The Net-Sales figure for one company exists in a source in which the Interest-Income value is included and the Net-Sales figure for the other exists in a different source in which the Interest-Income is excluded. Various issues such as the following may need to be addressed:

- Is such a semantic definition difference significant to the user's decision process?
- Can a certain approximation procedure be used?
- If the values are mutually computationally-path related, which value should be converted to the other's semantic definition?
- If the value of Other-Income is available from a number of sources , which one should be selected?
- If due to a component failure a source containing the value for Other-Income is temporarily inaccessible, what should be done?
- Under which circumstances and in what form, should information regarding such issues be presented to the user?

Consequently a flexible general framework is required within which:

1. The semantics of data items in information sources is explicitly represented and is accessible to users and decision processes.
2. Flexible tools are available for users and decision processes to address semantic conflicts challenges in a customized way which reflects their local requirements.

7 A Data Semantics Management Scheme

A data semantics management scheme which is based on extensions to Context Interchange (COIN) foundations offers a promising approach for addressing the challenges associated with semantic conflicts across multiple financial information sources. It enables to capture relevant semantic details and provides support for selective user-controlled conflict management operations. However certain important extensions to the COIN approach are required for addressing some of the above challenges.

The approach underlying the COIN prototype system [Goh96, BFG$^+$97b, BFG$^+$97a, GMS94] is based on the notion of a *domain model.* A domain model contains information about semantic-types and certain relationships, such as sub-type and functional-dependency relationships.

Information sources, which may be fully-structured, semi-structured, or unstructured, are wrapped and are consequently integrated into the system as relational-like sources (potentially with some capabilities constraints). *Elevation axioms* may be specified to map each relevant attribute in each source to a semantic-type in the domain-model.

Each source and each target application environment is associated with a semantic *context,* specified in terms of the appropriate corresponding semantic-types. The process of interaction between an application and appropriate sources includes dynamic conversions between the different contexts, as required. The efficient execution is supported by a logic-based *abduction engine.*

Thus various semantics representation conflicts, such as naming, scaling and measurement units conflicts, can be addressed effectively by the COIN framework. For example, in various currently implemented prototype applications (cf. [Goh96]) a semantic type called *Money-amount* was defined in the domain model, with appropriate *currency* and *scaling* Methods. Consequently the semantic representation of any accessed attribute which represents monetary value would be examined (based on the context definition) and appropriate currency conversions or scale-factor conversions performed if required. Similarly for semantic types representing company names, conversions between different naming conventions are performed.

However various important extensions to the COIN framework are required in order to effectively address additional types of challenges described in the previous sections. These extensions fall into two main categories:

1. **Semantically richer domain models:** Development of extensive financial domain models containing a comprehensive set of different semantic types each corresponding to a relevant financial notion, associated with a unique well-defined semantic definition. Incorporation within such extended domain models of information about computation-path relationships between the semantic types.

2. **Extended user capabilities for accessing and manipulating semantic information:** Development of rich graphic-based interfaces to such extended domain models. Development of specification methods for expressing semantic conflict resolution policies

and constraints.

Such extensions may lead to effective bindings between any value and a semantic domain-model-based specification of that value, and in particular may lead to the following application consequences:

- Users and applications will be able to detect cases in which similarly named attributes in different information sources are associated with a different domain-model-based semantic specifications. For example, the attribute of Net-Sales in WS would be associated with a different domain model semantic-type than the attribute of Net-Sales in Disc, reflecting the the fact that one includes items such as Excise-Tax and the other does not.
- Users and applications will be able to detect cases in which differently named attributes in different information sources are associated with the same domain-model-based semantic specification. For example, the attribute *Op-Income* in WS and the attribute *Optg-Income* in Disc may both be associated with the same domain-model semantic type.
- Users and applications will be able to detect cases in which attributes in different information sources are computation-path related, and to specify appropriate application environment-specific alternatives, such as:
  - Which semantic-type and so which attribute should be chosen as the most appropriate, given particular application requirements.
  - Under particular circumstances, whether to compute a new value from a corresponding one.
  - Under particular circumstances, which computation path to prefer if there are multiple ones possible.
  - Under particular circumstances, which alternative fall-back operations to perform in cases that certain data items are unavailable.
  - For a given attribute and under particular circumstances, levels of value approximations which are acceptable with respect to the current application.

Extensions to the COIN model and to the underlying system prototype, for facilitating such capabilities, are currently under development.

8 Conclusions

Various potential financial applications advantages may be derived from accessing multiple information sources so customers will increasingly require such capabilities.

Semantic conflicts represent a real challenge in such environments. The Primark Web-demo inconsistencies constitute a representative motivating example, in particular if we recall that the corresponding information sources used in that demo, although autonomous, are part of the same global corporation.

Different classes of conflicts may lead to different types of inconsistencies with varying degrees of application significance. Generic models and corresponding tools need to be provided so that users may manage application-specific semantic-conflicts according to their particular requirements.

The Context Interchange scheme provides a promising foundation for such a semantic-conflict management framework. The Context Interchange project efforts are currently proceeding in a few directions, including:

- Incorporating a wide variety of additional information sources using extended wrapping technology.
- Richer graphic-based user interfaces to the domain models and to system modules.
- Improving the expressiveness and efficiency of reasoning mechanisms.
- Examining in detail particular application environments in order to further fine-tune certain features of the approach and to empirically validate its feasibility and potential benefits.

Such a framework should be extremely beneficial for addressing the semantic conflicts challenges presented here, in environments of multiple information sources such as that of Primark.

References

[AK92] Yigal Arens and Craig A. Knoblock. Planning and reformulating queries for semantically-modeled multidatabase systems. In *Proceedings of the 1st International Conference on Information and Knowledge Management*, pages 92–101, 1992.

[AKWS95] Shailesh Agarwal, Arthur M. Keller, Gio Wiederhold, and Krishna Saraswat. Flexible relation: An approach for integrating data from multiple, possibly inconsistent databases. In *Proc. IEEE Intl Conf on Data Engineering*, Taipei, Taiwan, March 1995.

[ASD$^+$91] R. Ahmed, P. De Smedt, W. Du, W. Kent, M. A. Ketabchi, W. A. Litwin, A. Raffi, and M.-C. Shan. The Pegasus heterogeneous multidatabase system. *IEEE Computer*, 24(12):19–27, 1991.

[BFG$^+$97a] S. Bressan, K. Fynn, C. Goh, S. Madnick, T. Pena, and M. Siegel. Overview of a prolog implementation of the context interchange mediator. In *Proc. of the International Conference on Practical Applications of Prolog*, 1997.

[BFG+97b] S. Bressan, K. Fynn, C.H. Goh, M. Jakobisiak, K. Hussein, H. Kon, T. Lee, S.E. Madnick, T. Peno, J. Qu, A.C.Y. Shum, and M.D. Siegel. The context interchange mediator prototype. In *Proc. of ACM SIGMOD97 Conference*, 1997.

[BHP92] M.W. Bright, A.R. Hurson, and S.H. Pakzad. A taxonomy and current issues in multidatabase systems. *IEEE Computer*, 25(3):50–60, 1992.

[BT85] Y. J. Breitbart and L. R. Tieman. ADDS: Heterogeneous distributed database system. In F. Schreiber and W. Litwin, editors, *Distributed Data Sharing Systems*, pages 7–24. North Holland Publishing Co., 1985.

[CHS91] Christine Collet, Michael N. Huhns, and Wei-Min Shen. Resource integration using a large knowledge base in Carnot. *IEEE Computer*, 24(12):55–63, Dec 1991.

[DH84] Umeshwar Dayal and Hai-Yann Hwang. View definition and generalization for database integration in a multidatabase system. *IEEE Software Engineering*, 10(6):628–645, 1984.

[GMPQ+95] H. Garcia-Molina, Y. Papakonstantinou, D. Quass, A. Rajaraman, Y. Sagiv, J. Ullman, and J. Widom. The TSIMMIS approach to mediation: data models and languages. In *Proc NGITS (Next Generation Information Technologies and Systems)*, Naharia, Israel, June 27–29 1995. To appear.

[GMS94] Cheng Hian Goh, Stuart E. Madnick, and Michael D. Siegel. Context interchange: overcoming the challenges of large-scale interoperable database systems in a dynamic environment. In *Proceedings of the Third International Conference on Information and Knowledge Management*, pages 337–346, Gaithersburg, MD, Nov 29–Dec 1 1994.

[Goh96] Cheng Hian Goh. Representing and reasoning about semantic conflicts in heterogeneos information systems. *Ph.D. Thesis, M.I.T. - Sloan School of Management*, December. 1996.

[KL88] Eva Kuhn and Thomas Ludwig. VIP-MDBMS: A logic multidatabase system. In *Proc Int'l Symp. on Databases in Parallel and Distributed Systems*, 1988.

[KLK91] Ravi Krishnamurthy, Witold Litwin, and William Kent. Language features for interoperability of databases with schematic discrepancies. In *Proceedings of the ACM SIGMOD Conference*, pages 40–49, 1991.

[KS91] Won Kim and Jungyun Seo. Classifying schematic and data heterogeneity in multidatabase systems. *IEEE Computer*, 24(12):12–18, 1991.

[LA87] Witold Litwin and Abdelaziz Abdellatif. An overview of the multi-database manipulation language MDSL. *Proceedings of the IEEE*, 75(5):621–632, 1987.

[Lit92] W. Litwin. O*SQL: A language for object oriented multidatabase interoperability. In David K Hsiao, Erich J. Neuhold, and Ron Sacks-Davis, editors, *Proceedings of the IFIP WG2.6 Database Semantics Conference on Interoperable Database Systems (DS-5)*, pages 119–138, Lorne, Victoria, Australis, Nov 1992. North-Holland.

[LR82] T. Landers and R.L. Rosenberg. An overview of Multibase. In *Proceedings 2nd International Symposium for Distributed Databases*, pages 153–183, 1982.

[Mot87] Amihai Motro. Superviews: virtual integration of multiple databases. *IEEE Software Engineering*, 13(7):785–798, 1987. view integration.

[SL90] A.P. Sheth and J.A. Larson. Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys*, 22(3):183–236, 1990.

[SM91] Michael Siegel and Stuart E. Madnick. A metadata approach to resolving semantic conflicts. In *Proc. of the 17th International Conference on Very Large Data Bases*, 1991.

[SS77] J.M. Smith and D.C.P. Smith. Database abstractions: aggregation and generalization. *ACM Transactions on Database Systems*, 2(2):105–133, 1977.

[SSR94] Edward Sciore, Michael Siegel, and Arnie Rosenthal. Using semantic values to facilitate interoperability among heterogeneous information systems. *ACM Transactions on Database Systems*, 19(2):254–290, June 1994.

[SYE$^+$90] P. Scheuermann, C. Yu, A. Elmagarmid, H. Garcia-Molina, F. Manola, D. McLeod, A. Rosenthal, and M. Templeton. Report on the workshop on heterogeneous database systems. *ACM SIGMOD RECORD*, 19(4):23–31, Dec 1990. Held at Northwestern University, Evanston, Illinois, Dec 11–13, 1989. Sponsored by NSF.

[TBD$^+$87] M. Templeton, D. Brill, S. K. Dao, E. Lund, P. Ward, A. L. P. Chen, and R. MacGregor. Mermaid — a front end to distributed heterogeneous databases. *Proceedings of the IEEE*, 75(5):695–708, 1987.

[Wol89] Antoni Wolski. LINDA: A system for loosely integrated databases. In *Proceedings of the Fifth International Conference on Data Engineering*, pages 66–73, 1989.