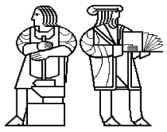

Massachusetts Institute of Technology
Engineering Systems Division

Working Paper Series

ESD-WP-2008-02

SEMANTIC INTEGRATION APPROACH TO EFFICIENT BUSINESS DATA SUPPLY CHAIN: INTEGRATION APPROACH TO INTEROPERABLE XBRL

Hongwei Zhu[1] and Stuart E. Madnick[2]

[1]College of Business and Public Administration
Old Dominion University
hzhu@odu.edu

[2]Sloan School of Management
Massachusetts Institute of Technology
smadnick@mit.edu

January 2008

Semantic Integration Approach to Efficient Business Data Supply Chain: Integration Approach to Interoperable XBRL

Hongwei Zhu
College of Business and Public Administration
Old Dominion University
Norfolk, VA 23529, USA
hzhu@odu.edu

Stuart E. Madnick
Sloan School of Management
MIT Room E53-321
Cambridge, MA 02139, USA
smadnick@mit.edu

Abstract

As an open standard for electronic communication of business and financial data, XBRL has the potential of improving the efficiency of the business data supply chain. A number of jurisdictions have developed different XBRL taxonomies as their data standards. Semantic heterogeneity exists in these taxonomies, the corresponding instances, and the internal systems that store the original data. Consequently, there are still substantial difficulties in creating and using XBRL instances that involve multiple taxonomies. To fully realize the potential benefits of XBRL, we have to develop technologies to reconcile semantic heterogeneity and enable interoperability of various parts of the supply chain. In this paper, we analyze the XBRL standard and use examples of different taxonomies to illustrate the interoperability challenge. We also propose a technical solution that incorporates schema matching and context mediation techniques to improve the efficiency of the production and consumption of XBRL data.

Keywords: XBRL, semantic data integration, context mediation, ontology, schema matching

1. Introduction

Modern organizations generate and consume a large amount of business data. The data is exchanged, integrated, extracted, and transformed as it flows within a data supply chain (e.g., from divisions to corporate, regulators, finical data processing intermediaries, analysts, and investors). To improve the efficiency of the supply chain, XBRL International, a not-for-profit consortium of approximately 550 companies and agencies around the world, has developed a language called the eXtensible Business Reporting Language (XBRL) (Chang and Jarvenpaa 2005). This language makes it possible that data encoded using XBRL by any organization can be processed by any other organization using XBRL software. Thus XBRL has the potential of reducing the cost of exchanging business data within the supply chain. However, as we will see later, there is still substantial semantic heterogeneity XBRL data which hinders the interoperability of various parts of the supply chain. Technology is needed to enable interoperability and help realize the potential benefits of XBRL.

As a language, XBRL can be used to develop data standards (known as *XBRL taxonomies*, analogous to database schemas) to be used by organizations to encode their data as *XBRL instances* (analogous to database records). Several countries (e.g., Canada, China, Korea, U.S., etc.) and industry sectors within a country (e.g., commercial and industrial, banking and savings, insurance, and several other sectors in the U.S.) have developed XBRL taxonomies for financial reporting (often known as external reporting as the taxonomies are intended for producing

financial reports for regulatory filings). XBRL International has also developed a taxonomy, called the General Ledger (GL) taxonomy, to support internal reporting that often requires more detailed information than external reporting.

Publicly traded companies in countries such as China and Korea have begun to report their financials using XBRL. In the U.S., several dozen firms have experimented with XBRL in a voluntary program by the Securities and Exchange Commission (SEC), which will likely require all filings to use XBRL in the near future (Hannon 2006). As the amount of XBRL data continues to grow, there will be an increasing need for effective use of the data.

To illustrate semantic heterogeneity in XBRL and the problems it creates, let us consider a simple case where a U.S. analyst needs to compare the *operating profit* of two firms: one in the U.S. and the other in China. For the U.S. firm, the analyst may be able to identify the element for operating profit, *OperatingProfit*. Below is a fact in the XBRL instance of the U.S. firm:

```
<!-- a fact of a U.S. firm -->
<usfr-pte:OperatingProfit contextRef="P1YFY2004" decimals="-6" unitRef="USD"> 9034000000</usfr-pte:OperatingProfit>
```

For the Chinese firm, it is not as easy to identify the equivalent element because it is named *YingYeLiRun*, the Pinyin (i.e., the standard Mandarin Romanization) of operating profit in Chinese. Below is a fact for the concept from the XBRL instance of the Chinese firm:

```
<!-- a fact of a Chinese firm -->
<cnfr-pt:YingYeLiRun contextRef="C_duration_20030101_20031231" decimals="0" unitRef="U_CNY">
943893000</cnfr-pt:YingYeLiRun>
```

This is an example of the typical semantic heterogeneity problem: the same data is named differently in sources having different schemas. At this point, the two numbers still cannot be compared directly because they also have different contexts: they are measured in different currencies (indicated by *unitRef* attribute) and have different accuracies (indicated by *decimals* attribute). We will use the term context to refer to any information useful for interpreting the data, which is broader than what is included in the *context* element of XBRL.

Given the large sizes of various taxonomies (approximately 2,000 concepts in a typical financial reporting taxonomy) and the diverse contexts of instances, we cannot expect every user of XBRL to be familiar with all the concepts in multiple taxonomies, understand their relationships, and perform context reconciliation manually. A tool is needed to alleviate the users from these tedious tasks. Likewise, XBRL instance producers also have difficulties in converting their internal data into XBRL instances, especially when using multiple taxonomies.

This example shows that to facilitate useful analysis of XBRL data, and more generally, to improve the efficiency of XBRL data supply chain, we need to develop effective tools for reconciling the schematic and contextual differences of data.

This paper makes two contributions towards this objective. First it identifies and illustrates semantic heterogeneity that exists in multiple XBRL taxonomies and instances. Second it proposes a solution that integrates schema matching techniques into a context mediation framework to automate the reconciliation of semantic heterogeneity.

2. Preliminaries of XBRL

XBRL as a language is defined in XBRL Specification (XBRL International 2006). It offers syntactic uniformity (e.g., a standard set of data types) desired for automatic data processing. The language can be used to develop XBRL taxonomies in different jurisdictions. These taxonomies are essentially different XBRL data standards. Using the same taxonomy, different firms can still produce instances that have different contexts because contexts and units are specified by individual firms in the instances, not in the taxonomies.

XBRL taxonomies define a set of concepts (e.g., *operating profit*) as XML elements. For each element, its data type, attributes, relationships with other elements, and relationships with other resources (namely labels for human readers and references to authoritative sources). The instances contain the facts tagged using the elements defined in the taxonomies. In addition, each fact is associated with a context, which specifies the entity related to the fact and the time period of the fact. If the fact is of a numeric type, it is also associated with a unit of measure. Below we first provide an overview of the language specification and various taxonomies developed using the specification. Then we provide detailed examples of taxonomies and instances.

2.1 XBRL specification

Using XML Schema and XML Linking (XLink), XBRL Specification defines the business reporting language by specifying a set of data types, XML elements, and the attributes of each element. It contains a set of XML Schema files, accompanied by a narrative document with explanations, examples, and further constraints that cannot be expressed using XML Schema.

In addition to all the built-in types of XML Schema, XBRL also includes types of particular relevance to the business information reporting domain. These types are derived from the built-in types. For example, *monetaryItemType*, *sharesItemType*, *pureItemType* (for value changes in percentages), and *fractionItemType* are all derived from the built-in numeric types.

2.2. XBRL taxonomies – Schema and Linkbases

Using XBRL, any country or industry can develop its own taxonomies and have them formally recognized. A taxonomy consists of a *taxonomy schema* and a set of *linkbases*. An organization often develops its own taxonomies by adopting and extending recognized taxonomies (e.g., Microsoft developed its taxonomies for its SEC filings by adopting the US GAAP Commercial and Industrial taxonomies and adding approximately 500 elements of its own.

Although all taxonomies are developed using the same XBRL language, the elements they define may differ among different taxonomies, so do the contexts and the units in the instances. These differences pose significant challenges to the producers (who may need to produce instances conforming to different taxonomies) and the consumers (who often need to integrate instances created using different taxonomies) of XBRL data.

A taxonomy schema defines the reporting concepts as XML elements. Each element is given a name and a type, and may have a set of required or optional attributes. Below are the definitions of *operating profit* in the taxonomy schemas used to create the facts in the example:

```
<!-- element definition in U.S. taxonomy -->
<element name="OperatingProfit" id="usfr-pte_OperatingProfit" type="xbrli:monetaryItemType"
substitutionGroup="xbrli:item" nillable="true" xbrli:balance="credit"
xbrli:periodType="duration"/>

<!-- element definition in Chinese taxonomy-->
<element name="YingYeLiRun" type="xbrli:monetaryItemType" substitutionGroup="xbrli:item"
nillable="true" id="clcid-pt_YingYeLiRun" xbrli:periodType="duration"  xbrli:balance="credit"/>
```

The two definitions are different in *name* and *id* attributes. The data type is *monetaryItemType* (defined in XBRL Specification). The element value is valid over a period of time ("duration" for *periodType* attribute) as opposed to at an instantaneous moment (in which case "instant" will be assigned to *periodType* attribute). The element will appear on the balance sheet as a "credit" (indicated by *balance* attribute). In addition to contents that have simple types (e.g., *monetaryItemType*), an element may have a content model of a complex type, e.g., a sequence of simple-typed elements.

XBRL has five types of linkbases: definition, calculation, presentation, label, and reference. A definition linkbase specifies the conceptual prelateships between elements, mainly the generalization-specialization relationship often found in OO, extended ER, and ontology modeling. For example, *profit* is a more general concept than *operating profit*. Although this is conceptually usefully, most taxonomies in the U.S. do not include a definition linkbase.

A calculation linkbase defines the numeric relationships between elements. For example, in the U.S. financial reporting taxonomy, *operating profit* is related to *gross profit* and *operating expenses* according the following formula:

OperatingPrfoit = GrossProfit – OperatingExpenses

The right hand side of the above calculation can be considered as a weighted sum of the elements (i.e., first term has a weight 1 and the second has a weight -1). In contrast, the corresponding element of operating profit in the Chinese taxonomy is related to five elements (sum of two profit related elements less three expense related elements). Calculation links essentially define constraints between elements. However, when the value of *use* attribute is "optional", the constraint will not be enforced.

A presentation linkbase specifies the hierarchical grouping (mainly the parent-child relationship) and the order of the elements when they are presented in a report for viewing purposes.

A label linkbase provides the human-readable documentation for the elements defined in the taxonomy schema. An element may have multiple labels in different languages (indicated by the *lang* attribute), or to be used for different documentation purposes and in different scenarios, which are indicated by the label's *role* attribute. The value of the *role* attribute can be one of the nearly two dozen values defined by the XBRL language. For example, the value `http://www.xbrl.org/2003/role/label` (in the ensuing discussion we will omit the substring up to the last slash in the URI) indicates that the label is the standard label for the element; the value `terseLabel` indicates the label is a short label for the element; the value `postiveLabel` indicates that this label will be used when the element has a positive value; the value `documentation` indicates that the label provides explanation of the meaning and other documentation of the element. In the U.S. taxonomy, OperatingProfit has several labels, e.g.,

"Gross profit less operating expenses" as the documentation label and "Operating Income/(Loss)" as the standard label. For YingYeLiRun, the Chinese taxonomy has one label in English, "Iii. Operating Profit ('-' for Losses)". Taxonomies from non-English speaking jurisdictions usually provide English labels.

Similar to the label linkbase, a reference linkbase provides further explanations to the elements by linking them to authoritative references (e.g., SEC regulations or certain accounting standards) that define the meaning of the elements. For example, it can link a certain concept to page *x* paragraph *y* of chapter *z* in the publication identified by a URL. We shall note that the reference linkbase only provides the pointers to the references, it does not attempt to encode the details of the reference using any kind of knowledge representation language.

2.3 Instances – Facts and Contexts

XBRL instances contain the facts as well as the descriptions of their contexts using the *context* and *unit* elements. The XBRL Specification requires that an element should be associated with a *context* element using the *contextRef* attribute. If the element is of a numeric type (or a type derived from a numeric type), it must be associated with a *unit* element using the *unitRef* attribute.

3. Semantic Heterogeneity in XBRL

The extensibility of XML is a double-edged sword. It gives organizations the flexibility of extending taxonomies to suit their reporting and disclosure needs, but such extensions can introduce incompatibility among the taxonomies used by different organizations. The downside of "extensibility" is well recognized by XBRL stakeholders (Cohen 2004; Debreceny *et al.* 2005). Due to the diverse organizational needs and various other factors (Williams *et al.* 2006), we can not expect to have a single universal XBRL taxonomy and a uniform context for all instances. Thus, there will be semantic heterogeneity in XBRL. In this section, we identify and exemplify various kinds of XBRL semantic heterogeneity.

3.1 Schematic Heterogeneity

There are two kinds of schematic heterogeneity: (1) element naming, and (2) element relationship topology. The different element names for the same *operating profit* concept are an example of naming heterogeneity.

The relationships of the matched elements with other elements can be different in different taxonomies. The topological differences of calculation links related to operating profit concept in the two taxonomies is shown as labeled graphs in Figure 1.

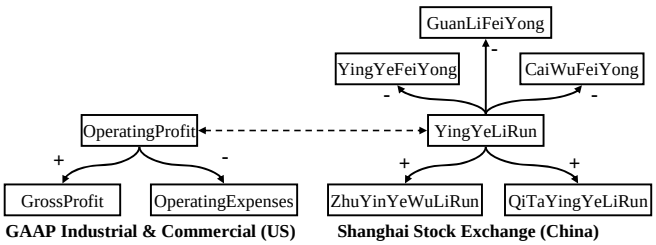

Figure 1. Calculation links in two taxonomies for element corresponding to *operating profit* concept. *OperatingProfit* corresponds to *YingYeLiRun*, indicated by dashed line.

Rich equational relationships among elements can be exploited by matching algorithms, e.g., after identifying the two elements corresponding to *operating profit* concept (dashed line in Figure 1), and given the structures of the two graphs, it can be inferred that *GrossProfit* corresponds to the sum of *ZhuYingYeLiRun* and *QiTaYingYeLiRun*.

3.2 Contextual Heterogeneity

In addition to the *context* element, the context of a fact may include the unit of measure, precision, scale factor (e.g., thousands or millions), etc. XBRL provides several other constructs for taxonomy users to describe the context of a fact in the instances. The *unit* element allows for the description of the unit of measure of a fact. XBRL does not have a construct for scale factor, but it offers a *decimals* attribute to indicate up to which decimal place the value is accurate. It also offers a *precision* attribute to indicate the significant figures of the value.

The values for describing a context may be from a constrained domain or an arbitrary domain. For example, the time period element of context is from temporal domain with data types derived from XML Schema built-in types related to date and time. The currency of monetary types (e.g., profit, earnings per share) is usually represented by the three letter codes defined by ISO 4271 standard. The interpretation of the values from most constrained domains is widely known and can be implemented in software to correctly process the data (e.g., temporal reasoning and currency conversion). In contrast, when the values are from an arbitrary domain (e.g., *sqft*, *ft_2*, or *myAreaUnit* for *square foot*), it will be difficult to detect context differences and implement software to reconcile them.

3.3 Ontological Heterogeneity

Ontological heterogeneity refers to the case where the elements in different taxonomies that appear to refer to the same concept actually have subtle differences. Consider the earlier example: we have assumed that *OperatingProfit* and *YingYeLiRun* refer to the same concept. But their values are derived using different accounting rules: GAAP of U.S. for *OperatingProfit*, Enterprise Accounting Regulation of the Ministry of Finance (China) for *YingYeLiRun*. We have not analyzed the two sets of accounting rules, but it is conceivable that there may be differences in terms of the items included or excluded in operating profit. XBRL has reference links to point to accounting standards that define the concept, but it does not provide a mechanism for representing the details of the accounting standards. As a result, ontological heterogeneity is difficult to identify and reconcile. We will ignore ontological heterogeneity in the ensuing discussion.

4. Solution: Schema Matching and Context Mediation

We propose a solution that integrates techniques of semantic schema matching (Rahm and Bernstein 2001) and the Context Interchange (COIN) framework (Goh *et al.* 1999, Zhu and Madnick 2007). Schema matchers can semi-automatically identify matching elements to resolve schematic heterogeneity, while the COIN framework uses a mediator to automatically rewrite queries to reconcile context differences. The integrated solution, which we call the XBRL Interoperability Engine, will implement a mediation architecture (Figure 2).

The XBRL processors (near the bottom in the figure) process the taxonomies and instances. For each instance, it produces an ontology corresponding to the taxonomies used by the instance.

The ontology includes all concepts in the taxonomies; relationships and context descriptions are also included in the ontology.

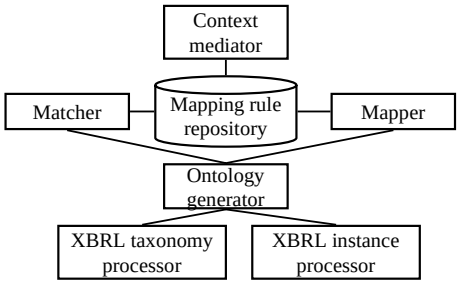

Figure 2. Solution architecture of the XBRL Interoperability Engine.

The matcher component is responsible for identifying one-to-one matching of concepts. It uses several matching algorithms, which can be used jointly, to identify the best matches. The similarity of elements will be evaluated using syntactic (e.g., Levenshtein edit distance) and semantic (e.g, senses and synonyms from WordNet (Miller 1995)) measures of elements names and their documentation. The structure of the element relationship graph will be used to improve the precision and recall of the matchers. The results are stored in the mapping rule repository.

The mapper is responsible of finding the complex relationships between concepts. It will mainly reason about the calculation links of matched elements to produce equational relationships between concepts. The results will also be stored in the mapping rule repository. In addition, the administrator can add other useful rules to the repository.

The context mediator component is a query rewriter that takes a user query expressed against a taxonomy and rewrites it against all the other involved taxonomies by consulting the rule repository. It also introduces context conversions to reconcile context differences in the data.

This Engine will help both XBRL data consumers and producers. The matcher can be used to further facilitate keyword-based queries by consumers. The matcher and mapper can establish relationships between elements in existing systems and those in XBRL taxonomies, and the context mediator can convert data values to create instances in different contexts.

5. Discussion

Significant amount of work has been done in semantic data integration. Although there have been solutions in the form of research prototypes (Bernstein *et al.* 2004; Rahm *et al.* 2004), most of them have not been evaluated using large sized schemas (or ontologies) and with datasets having a variety of contexts (e.g., in Bernstein et al. (2004), only 600 elements were in the test schemas and they did not consider context differences). The variety of XBRL taxonomies and instances offer a realistic test-bed for evaluating existing and emerging semantic integration solutions.

Since the introduction of XML, the W3C has introduced several other standards, namely RDF and OWL, to provide more expressive languages for describing semantics and enabling automated reasoning on the Semantic Web. There has been research that aims to leverage

Semantic Web technologies to process XBRL. An approach of translating XBRL into description logic is reported in (Declerck and Krieger 2006). Using an investment funds XBRL taxonomy as an example, a method of converting an XBRL taxonomy into an OWL ontology is presented in (Lara *et al.* 2006). While converting taxonomies into a description logic based language is an important step, further work should leverage the expressivity of the language to solve problems, e.g., reconciling semantic heterogeneity. We believe the proposed solution here offers a good demonstration of how these new technologies can be integrated to solve a practical problem.

Future research will implement the proposed the solution. The implementation will be evaluated using existing taxonomies and the growing amount of XBRL instances that are becoming available globally.

References

1. Bernstein, P.A., Melnik, S., Quix, C., and Petropoulos, M. "Industrial-Strength Schema Matching," *ACM SIGMOD Record* (33:4), 2004, pp 38-43.
2. Chang, C., and Jarvenpaa, S. "Pace of Information Systems Standards Development and Implementation: The Case of XBRL," *Electronic Markets* (15:4), 2005, pp 365-377.
3. Cohen, E.E. "Compromise or Customize: XBRL's Paradoxical Power," *Canadian Accounting Perspectives* (3:2), 2004, pp 187-206.
4. Debreceny, R.S., Chandra, A., Cheh, J.J., Guithues-Amrhein, D., Hannon, N.J., Hutchison, P.D., Janvrin, D., Jones, R.A., Lamberton, B., Lymer, A., Mascha, M., Nehmer, R., Roohani, S., Srivastava, R.P., Trabelsi, S., Tribunella, T., Trites, G., and Vasarhelyi, M.A. "Financial Reporting in XBRL on the SEC's EDGAR System: A Critique and Evaluation," *Journal of Information Systems* (29:2), 2005, pp 191-210.
5. Declerck, T., and Krieger, H.-U. "Translating XBRL Into Description Logic. An Approach Using Protege, Sesame & OWL," 9th International Conference on Business Information Systems, Klagenfurt, Austria, 2006.
6. Goh, C.H., Bressan, S., Madnick, S., and Siegel, M. "Context Interchange: New Features and Formalisms for the Intelligent Integration of Information," *ACM TOIS* (17:3), 1999, pp 270-293.
7. Hannon, N.J. "Why the SEC Is Bullish on XBRL," *Strategic Finance* (January), 2006, pp 59-61.
8. Lara, R., Cantador, I., and Castells, P. "XBRL taxonomies and OWL ontologies for investment funds," First International Workshop on Ontologizing Industrial Standards, Tucson, Arizona, USA, 2006, pp. 271-228.
9. Miller, G.A. "WordNet: A Lexical Database for English," *Communications of the ACM* (38:11), 1995, pp 39-41.
10. Rahm, E., and Bernstein, P.A. "A Survey of Approaches to Automatic Schema Matching," *VLDB Journal* (10:4), 2001, pp 334-350.
11. Rahm, E., Do, H.-H., and Maßmann, S. "Matching Large XML Schemas," *ACM SIGMOD Record* (33:4), 2004, pp 26-31.
12. Williams, S.P., Scifleet, P.A., and Hardy, C.A. "Online Business Reporting: An Information Management Perspective," *International Journal of Information Management* (26:2), 2006, pp 91-101.
13. XBRL International "Extensible Business Reporting Language (XBRL) 2.1," XBRL International.
14. Zhu, H., and Madnick, S.E. "Scalable Interoperability Through the Use of COIN Lightweight Ontology," in: *Ontologies-Based Databases and Information Systems (LNCS 4623)*, M. Collard (ed.), Springer, 2007, pp. 37-50.