The Misguided Silver Bullet: What XML will and will NOT do to help Information Integration

Stuart E. Madnick[1]

ABSTRACT

The eXtensible Markup Language (XML) offers many important benefits and improvements over its predecessor, HTML. But, articles have appeared about XML with exaggerated claims of it being a "Rosetta Stone" with "miraculuous ways" to almost automatically provide information integration. These claims are actually being believed by some executives. It is almost surprising that no one has claimed that XML can cure cancer and provide world peace!

In reality, XML must face many of the same challenges that plagued Electronic Data Interchange (EDI) and database integration efforts of the past. To a large extent, there are both managerial and technical challenges – much related to the difficulties of attaining universally accepted semantically-rich standards. In this paper, these challenges will be discussed with specific emphasis on the issue of dealing with a real-world with multiple "contexts." Some promising research directions, some overlapping with the "semantic web" effort, will be presented.

1. Introduction

The eXtensible Markup Language (XML) offers many important benefits and improvements over its predecessor, HTML. Whereas once XML was merely described as "HTML on steroids," articles have appeared about XML with even more exaggerated claims of it being a "Rosetta Stone"[2] with "a universal way to translate data"[3] and "miraculous ways"[4] to almost automatically provide information integration. Some executives actually believe these claims. It is almost surprising that no one has claimed that XML can cure cancer and provide world peace!

Before proceeding, it must be emphasized that XML does have real benefits and most of the technical community, including the World Wide Web Consortium (W3C at www.w3.org), XML's originators, have taken a much more realistic perspective, recognize XML's limitations (e.g., [10]), and are working on further improvements [1]. The purpose of this article is to look at certain aspects of information integration and understand XML's capabilities and limitations.

In reality, XML must face many of the same challenges that plagued Electronic Data Interchange (EDI) and database integration efforts of the past. To a large extent, there are both managerial and technical challenges – much related to the difficulties of attaining universally accepted semantically rich standards. In this paper, these challenges will be discussed with specific emphasis on the issue of dealing with a real world with multiple "contexts." Some promising research directions, some overlapping with the "semantic web" effort, will be presented.

[1] Stuart E. Madnick, John Norris Maguire Professor of Information Technologies and Professor of Engineering Systems; MIT Sloan School of Management; Cambridge, MA 02139; smadnick@mit.edu.
[2] "Rosetta Stone," *Federal Computer Week*, June 19, 2000; "A 'Rosetta Stone' for the Web?" *Business Week Online*, June 14, 1999.
[3] "XML: The Rosetta Stone for the Web," *Philly-Tech.com*, April 2001 / Vol. 4.
[4] "The Threat of XML," *ComputerWorld, July 9, 2001*

2. Examples of Information Integration Applications and Requirements

"Information integration" is a term used to describe many different activities. For the purposes of this paper, we will focus on a particular set of applications and requirements, often referred to as "information aggregation."

Two particularly popular current examples include "comparison" aggregators and "relationship" or "account" aggregators. Aggregators with *comparison* capabilities are focused on collecting information, especially prices, about specific products from multiple sources, primary online merchants. Shopbots such as for those for purchasing books, music, and electronics are good examples of this capability. These include MySimon (www.mysimon.com), C|net (www.cnet.com), and DealTime (www.dealtime.com). *Relationship* aggregators focus on collecting information related to the individual (or organization) rather than a product. Financial account aggregator technology (e.g., www.yodlee.com) has been adopted by most major financial (e.g., Chase, Citibank) and many non-financial institutions (e.g., CNBC, AOL). These organizations provide their customers with the ability to manage all their financial relationships through a single aggregator. For example, this includes the ability to see all of their account balances, from all sources (e.g., bank accounts, brokerage accounts, credit cards, mortgages), integrated onto a single web page. These comparison and relationship aggregators might operated *intra-organization*ally, collecting information from multiple parts of a given enterprise (e.g., financial information from all company divisions, manufacturing data from different plant locations) or might operate *inter-organizationally,* combining information from multiple enterprises (e.g., price and account balance information from multiple online sites.) A single aggregator may combine both relationship and comparison capabilities for a given application.

It is important to note that in such applications, the primary and original purpose of the source sites was not to support information aggregation. The individual online stores posted their product prices for users visiting their site. The individual banks and other financial institutions made customer account balances available online as a service and convenience for their customers. Although in some cases direct data feeds and data exchange arrangements were made between source sites and aggregators, in most cases the data was obtained from the sources using techniques often referred to as "screen scraping" or "web wrapping." These techniques involve the aggregator accessing the source site as if it were a user (e.g., a browser) and then extracting the desired information from the information provided (usually an HTML or XML page).

3. Benefits of XML

The benefits of XML have been described extensively in the literature (the list shown in Figure 1 is adapted from [10]), so only a few key highlights will be discussed here. Probably one of the most important benefits is that XML does help to create structured web pages, compared with HTML.

Feature	HTML	XML
Extensibility	Fixed set of tags	Extensible set of tags
Tag purpose	Tags describe presentation	Tags describe data content
Views	Single presentation	Multiple views of same document (by XSL)
Orientation	Documents	Documents plus semi-structured data
Search	Keyword search only	Keyword plus field-sensitive queries

Figure 1. Comparison of HTML and XML

In Figure 2(a) we see an example of an HTML page that might be returned when requesting price information, in this case for a Palm Pilot V, from an online store. The HTML tags are used to provide formatting information, such as margin sizes, font size, and such. The actual price information might be simple text, as shown in Figure 2(a), or embellished with HTML tags defining table delimiters and different font types, sizes, and/or colors for the different information (e.g., "Regular Price" in different color from "Our Price"). A considerable amount of programming effort would be required to extract the price information from such a page in order to produce the desired comparison aggregation of listing the corresponding prices for Palm Pilot V's from multiple stores – especially since it is likely that different formats will be used by different stores. Tools to support and simply this effort, sometimes called "web wrappers," have been developed [3].

(a) HTML
```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 3.2 Final//EN">
<html>
<head> . . .
<BODY topmargin=18 leftmargin=6 bgcolor="#ffffff" link="#0000ee" VLINK="#551A8B"
ALINK="#ff0000">
<pre><font size=2>
                    Regular  Our
                    Price    Price
Palm Pilot V        329.00   236.00    In stock
</font></pre>
<table cellpadding=0 cellspacing=0 border=0>
<tr><td align=left valign=middle width=455 nowrap height=20>
<tr><td align=left valign=top nowrap width=455>
<font size=1 face="helvetica,arial"> . . .
</BODY>
</HTML>
```

(b) XML
```
<?xml version="1.0" ?>
<Document>
 <Product info>
   <Product> Palm Pilot V <\Product>
   <Regular price> 329.00 <\Regular price>
   <Our price> 236.00 <\Our price>
   <InStock> yes <\InStock>
 <\Product info>
<\Document>
```

Figure 2. Product information using HTML and XML

In Figure 2(b), we see a possible XML page providing the same information. The actual formatting of the page for presentation purposes is controlled by other facilities, such as eXtensible Stylesheet Language (XLS) – and different online stores might use different stylesheets to produce very different appearances of their data. For purposes of data extraction and aggregation of the price information, the XML tags make the programming much easier. Furthermore, various XML parsers are readily available to facilitate this process. Assuming all online stores used the same XML tags, once the extraction procedure is established for one store, it should be fairly trivial to apply it to many more stores. Thus, XML does, indeed, greatly facilitate such information integration.

4. XML Tends to Assume One Context

For purposes of information exchange and integration, to a large extent, XML relies upon a generally agreed understanding of the XML "tags". That is, it assumes or hopes that these tags are all understood in the same way by all of those sharing the information. In cases where there is initial disagreement, it is assumed that a "best way" will prevail or if there is not a way to agree upon "best" than an acceptable compromise can be found. This is a primary basis behind the concept of standards.

This standardization assumption can be a critical problem. For example, consider Figure 3; is it a picture of an old lady or a young lady? The point here is that some will see it one way, some will see it the other way, and most be able to see both images – but only one at a time. [If you are unable to see both, email me and I will send clues for seeing each.] This is the situation that we

Figure 3. Old woman or young woman?

often face in real life. There is no "best" answer and different people will continue to see things in different ways. Merely saying that everyone should see it the same way, does not change the reality that multiple different legitimate, and often essential, views exist. Examples of this situation will be presented in this paper.

5. XML Challenges

There are many challenges that XML must overcome in order to provide effective information integration.

5.1 Multiple Standards

One of the key issues in using XML for information interchange or integration is the need for consistent and standardized tags. There have been various efforts to create such standards in various industries. Disputes over whether catalogs should use the tag "price" or "cost" must be settled. Once agreed upon, these standardized tags can then become an effective way to exchange information. Unfortunately, what is so great about XML standards – is that there are so many! For example, the Director of Electronic Trading at Credit Suisse First Boston and Chairman of a Financial Services XML Working Group was quoted as saying "there are more than a dozen XML protocols – for Financial Trading applications alone."[5]

To a certain extent, it is possible to map between the tags used in different XML standards using eXtensible Stylesheet Language – Translations (XSLT) facilities which would help convert "price" to "cost", or vice versa. This can work well for fairly simple cases. But XSLT is primarily designed for pair-wise translations, if there were n different sets of tags being used, there might need to be n^2 translations – which could be difficult if n were a large number and/or tag names change periodically.

5.2 Rich Semantics

Let us assume that somehow all of these conflicting standards efforts converge and we all agree upon a tag label "price" to be used in describing catalog information. But how precisely defined is the notion of "price"? Is it in dollars ($) or pounds (£)? Even if it is "Dollars" is it US dollars, Canadian dollars, Hong Kong dollars, or Singapore dollars? If we are dealing with international commerce, these become important issues. There are various pragmatic ways to provide more semantic information with the XML tags such as by having an auxiliary tag "currency" which specifies what currency the price should be interpreted as. There are at least two problems with this approach: (1) it is possible that the group developing the standard may have neglected to consider the need for this "currency" modifier tag, specially if it was a group primarily dealing with domestic commerce and (2) "currency" is only one of many semantic issues that must be addressed – will all of them had been considered? For example, does the "price" include sales tax? (A key issue in the United States) Does it include the value added tax (VAT) (an issue common in much of Europe)? Does it include shipping charges? Etc. etc. etc.

One of the major challenges that Electronic Data Interchange (EDI) faced was coming up to a sufficient level of detail to address all such cases. Because there can be so many nuances (and we are not only considering the single issue of "price") these efforts can go on for years – often without reaching agreement. Even once agreements are sometimes reached, the end result is often of such complexity that it becomes a major burden and difficulty to actually deploy. Once XML gets

---

[5] "The Threat of XML," *ComputerWorld, July 9, 2001*

beyond simple cases, it must also confront the same challenges. To illustrate how complex this can become, consider figure 4, which represents actual data retrieved from the web.

Although the actual documents were encoded using HTML at the time, they all used the heading "p/e ratio" to describe data displayed as shown. Presumably, a similar XML tag would be used in the future. But, you should notice that the four web sources, on the same day at approximately the same time, had radically different values for "p/e ratio". How can these happen? Which one is correct? The possibly surprising answer

Source	P/E Ratio	Dividend
ABC	11.6	0.29
Bloomberg	5.57	8.127
DBC	19.19	0.899
MarketGuide	7.46	0.47

Figure 4. Key Financials for Daimler-Benz

is: they are all correct! The issue is, what do you really mean by "p/e ratio". Some of these sites even provide a glossary which gives a definition of such terms and they are very concise in saying something like "p/e ratio" is "the price divided by the earnings". As it turns out, this does not really help us to explain the differences. The answer lies in the multiple interpretations and uses of the term "p/e ratio" in financial circles. It is for the entire year for some but for others it is only for the last quarter. Even when it is for a full year, is it the last four quarters? the last calendar year? the last fiscal year? or the last three historical quarters and the estimated current quarter (a popular usage)?

5.3 Evolution of Semantics

Even if one could agree upon a standard set of tags and even reach agreement upon the precise meaning of each tag, there is also the problem of evolution. For example, in Europe most countries are going through the conversion from using local currency (such as French Francs) to using Euros as a common currency throughout Europe. Let us say that this transition takes place in France on January 1, 2002. Thus, documents to be exchanged that originated prior to that date, will be assuming local currency whereas documents created after that date will be assuming "Euros". Furthermore, not all organizations may make this change at the same time. This is a somewhat simple case that could be resolved if there was a "currency" tag modifier being used – but in a more general case, such as with "p/e ratio", the nature of such a change may not be incorporated in the preexisting tag set.

5.4 Multiple Purposes

Probably the fundamental challenge and difficulty with attaining standards in general, is often there are multiple purposes to be served for information exchange and these different purposes necessitate different interpretations of the information. For example, consider figure 5, which is a list of organization names.

What is the relationship among these names? As it turns out, these are all names that are in some way related to each other and International Business Machines Corporation (the name at the top of the list). These names include abbreviations (such as IBM), divisions (such as IBM Microelectronics division), wholly owned subsidiaries (such as IBM Global Financing), partially

International Business Machines Corp.
IBM
IBM Microelectronics Division
IBM Global Services
IBM Global Financing
IBM Global Network
IBM de Colombia, S.A
Lotus Development Corporation
Software Artistry, Inc.
Dominion Semiconductor Company
MiCRUS
Computing-Tabulating-Recording Co.

Figure 5. List of organization names

owned subsidiaries (such as IBM de Colombia, S.A.), companies that were acquired by IBM (such as Lotus Development Corporation), companies that were acquired and then later sold by IBM (such as SoftwareArtistry, Inc.), companies in which IBM has a minority joint venture interest (such as the Dominion Semiconductor company), and companies that IBM has a majority joint venture interest in (MiCRUS). This list even includes IBM's original name, Computing-Tabulating-Recording Company!

 What is the point of this discussion of these names? Well, let us consider a rather simple question: "How many employees does IBM have?" In our recent study of integration activities in a major insurance company, this was an example of an important question asked in setting premium rates for business owner protection insurance [13]. In considering the entities listed in figure 5, which ones should be included in this count? Also, how can one be sure that one is not also double counting? In fact, the answer to the question itself depends upon the purpose of the question. The important and subtle issue is, when are two entities to be considered as part of the same entity. This issue is sometimes referred to as the Corporate Household or Corporate Family Structure. As figure 6 illustrates, there are many different purposes for which such corporate structures are used – in some cases two entities should be combined and in other cases the two entities should not be combined.

Corporate household /family structure purposes -
- Financial: Risk (credit - bankrupcy, country)
- Accounting: Account consolidation
- Marketing (multiple divisions & subsidiaries):
- Customers & Supplier consolidation (economy of scale)
- Customer Relationship Management (CRM)
- Managerial: Regional and/or Product separations
- Legal: Liability (insurance)
- Relationship: Consultant Conflict of interest & competition
- Ad hoc/temp structures
- and these are dynamic, changing over time

Figure 6.   Examples of multiple purposes

In fact, figure 6 does not illustrate all the subtlety and complexity that may be involved. For example, a few years ago when Russia announced that it would suspend payments on its financial obligations, all of the major financial organizations were immediately concerned to know how much money they had at risk. (Although not a daily occurrence, similar problems do occur quite regularly, most recently involving financial instability in Argentina, previously Brazil, Asia a few years ago, and the threaten financial instability of Long Term Capital Management Corp, are other examples). In considering the case of the Russian risk, there are many complexities. The obvious cases are where the financial institution had directly loaned money to a Russian company or government organization. But, in fact, there are subtleties here. If, for example, the Russian company is a wholly owned subsidiary of a US Company, then the US company is obligated to be responsible for the debts of its subsidiary. Thus, this particular Russian company would not be included in the risk assessment, especially if the consideration is regarding bankruptcy. Likewise there can be considerations of liability in case of an industrial accident; there are complex laws that affect which entities will be held responsible.

Also as noted before, the corporate structures change over time, thereby also changing the context over time. Thus, at one point Lotus Development Corporation was a separate corporation. When doing a historical comparison of growth or decline in "number of employees" should the current Lotus employees be counted in a total as of today? Should the Lotus employees in 1980, when it was a separate corporation, be added with the IBM employees of 1980 to make a meaningful comparison?

6. Role of Context

In our work at MIT we have considered the importance of Context in interpreting information. Most information, no matter how carefully defined, usually has incomplete specifications. Some information is "assumed" by the users of that information. This is not a problem if the information and the people who use it are closely coupled together and share a common "context". For example, if an MIT student asked me what course I was teaching this year and I say "15.561" – a meaningful exchange of information has taken place. To an outsider who is not familiar with the MIT numbering conventions and meaning of the various courses, this information would not be particularly meaningful. The challenge we face is that the Internet and World Wide Web have made it possible to electronically gather documents from around the world, but the context is often left behind.

To illustrate the significance of this issue, consider the vignettes displayed in Figure 7(a) and Figure 7(b). In the case of Figure 7(a), the emissaries of the Austrian and Russian emperors thought that they had agreed on the battle being "October 20th". What they had not agreed upon was which October 20th! To illustrate that this kind of semantic misunderstandings do not only resided hundred of years in the past, consider Figure 7(b) where a similar mishap also had dramatic consequences for the Mars Orbiter satellite.

(a) The 1805 Overture

In 1805, the Austrian and Russian Emperors agreed to join forces against Napoleon. The Russians promised that their forces would be in the field in Bavaria by **Oct. 20**.

The Austrian staff planned its campaign based on that date in the **Gregorian calendar**. Russia, however, still used the ancient **Julian calendar**, which lagged 10 days behind.

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

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

(b) The 1999 Overture
Unit-of-Measure mixup tied to loss of $125 Million Mars Orbiter

"NASA's Mars Climate Orbiter was lost because engineers did not make a simple conversion from English units to metric, an embarrassing lapse that sent the $125 million craft off course … The navigators [JPL] **assumed metric units** of force per second, or newtons. In fact, the numbers **were in pounds** of force per second as supplied by Lockheed Martin [the contractor]."

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

Figure 7. Examples of consequences of misunderstood context

---

7. Context Interchange Architecture

The shortcomings of XML alone for information integration and meaningful interpretation have been recognized by others. The work by the W3C on the Resource Description Framework (RDF) and the "Semantic Web" [1] are intended to address some of these issues.

The COntext INterchange (COIN) project at MIT is also addressing these needs through a mediation approach for semantic integration of disparate (heterogeneous and distributed) information sources [5,6,7,11,12]. The overall COIN project includes not only the mediation infrastructure and services, but also wrapping technology and middleware services for accessing the source information and facilitating the integration of the mediated results into end-users applications. The wrappers are physical and logical gateways providing a uniform access to the disparate sources over the network [3].

The set of Context Mediation Services comprises a Context Mediator, a Query Optimizer, and a Query Executioner. The Context Mediator is in charge of the identification and resolution of potential semantic conflicts induced by a query. This automatic detection and reconciliation of conflicts present in different information sources is made possible by general knowledge of the underlying application domain, as well as informational content and implicit assumptions associated to the receivers and sources. These bodies of declarative knowledge are represented in the form of a domain model, a set of elevation axioms, and a set of context theories respectively.

The result of the mediation is a mediated query. To retrieve the data from the disparate information sources, the mediated query is then transformed into a query execution plan, which is optimized, taking into account the topology of the network of sources and their capabilities. The plan is then executed to retrieve the data from the various sources; results are composed as a message, and sent to the receiver.

The Context Interchange (COIN) approach allows queries to the sources to be mediated, i.e., semantic conflicts to be identified and solved by a context mediator through comparison of contexts associated with the sources and receivers concerned by the queries. It only requires the minimum adoption of a common Domain Model, which defines the domain of discourse of the application.
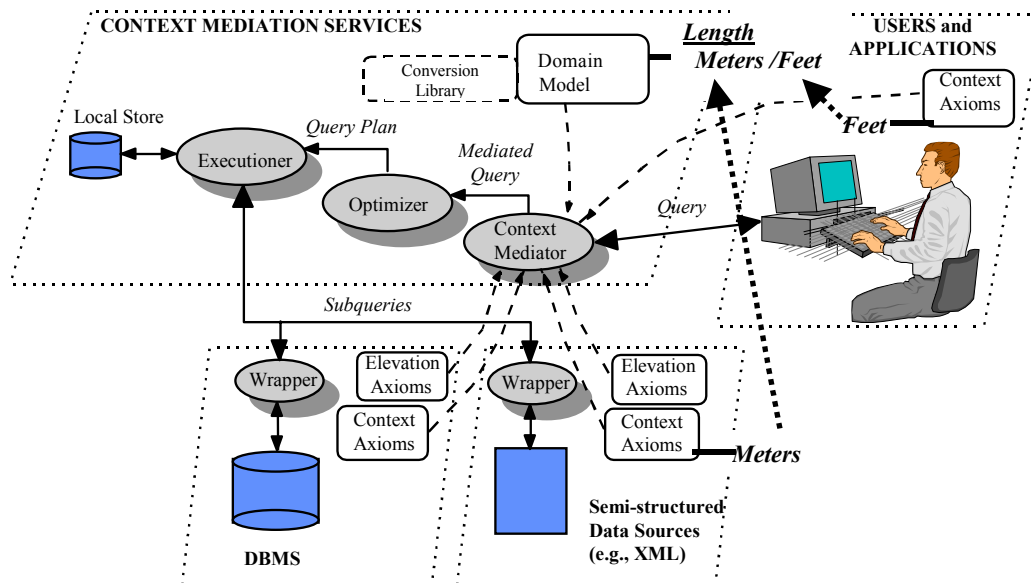

Figure 8. The Architecture of the Context Interchange System

The knowledge needed for integration is formally modeled in a COIN framework [4] as depicted in Figure 8. The COIN framework is a mathematical structure offering a sound foundation for the realization of the Context Interchange strategy. The COIN framework comprises a data model and a language, called COINL, of the Frame-Logic (F-Logic) family [2,9]. The framework is used to define the different elements needed to implement the strategy in a given application:

- The Domain Model is a collection of rich types (semantic types) defining the domain of discourse for the integration strategy (e.g., "Length");

- Elevation Axioms for each source identify the semantic objects (instances of semantic types) corresponding to source data elements and define integrity constraints specifying general properties of the sources;

- <u>Context Definitions</u> define the different interpretations of the semantic objects in the different sources or from a receiver's point of view (e.g., "Length" might be expressed in "Feet" or "Meters").

Finally, there is a conversion library which provides conversion functions for each modifier to define the resolution of potential conflicts. The conversion functions can be defined in COINL or can use external services or external procedures. The relevant conversion functions are gathered and composed during mediation to resolve the conflicts. No global or exhaustive pair-wise definition of the conflict resolution procedures is needed.

Both the query to be mediated and the COINL program are combined into a definite logic program (a set of Horn clauses) where the translation of the query is a goal. The mediation is performed by an abductive procedure which infers from the query and the COINL programs a reformulation of the initial query in the terms of the component sources. The abductive procedure makes use of the integrity constraints in a constraint propagation phase which has the effect of a semantic query optimization. For instance, logically inconsistent rewritten queries are rejected, rewritten queries containing redundant information are simplified, and rewritten queries are augmented with auxiliary information. The procedure itself is inspired by the Abductive Logic Programming framework [8] and can be qualified as an abduction procedure. One of the main advantages of the abductive logic programming framework is the simplicity in which it can be used to formally combine and to implement features of query processing, semantic query optimization and constraint programming.

8. Conclusion

We are in the midst of exciting times – the opportunities to access and integrate diverse information sources over the web are incredible but the challenges are considerable. XML can greatly facilitate this process by providing more syntactic structure to web documents. But the effective use of semantic metadata and context knowledge processing is needed to enable us to overcome the challenges described in this paper and more fully realize the opportunities. A particularly interesting aspect of the context mediation approach described is the use of context metadata to describe the <u>expectations of the receiver</u> as well as the <u>semantics assumed by the sources</u>.

<u>REFERENCES</u>
[1] Berners-Lee, T., Hendler, J., Lassila, O. (2001). "The Sematic Web," *Scientific American*, May 2001.
[2] Dobbie, G. and Topor, R. (1995). "On the declarative and procedural semantics of deductive object-oriented systems," *Journal of Intelligent Information Systems*, 4:193-219.
[3] Firat, A., S. Madnick, and Siegel, S (2000). "The Caméléon Web Wrapper Engine,", *Proceedings of the VLDB2000 Workshop on Technologies for E-Services,* September 2000.
[4] Goh, C. (1996). *Representing and Reasoning about Semantic Conflicts* In *Heterogeneous Information System*, PhD Thesis, MIT, June 1996.
[5] Goh, C.H., Bressan, N. Levina, S., Madnick, S.E., A. Shah, and Siegel, M.D. (2000). "Context Knowledge Representation and Reasoning in the Context *Applied Intelligence: The International Journal of Artificial*

Intelligence, Neutral Networks, and Complex Problem-Solving Technologies, Volume 12, Number 2, September 2000, pp. 165-179.

[6] Goh, C.H., Bressan, S., Madnick, S.E., and Siegel, M.D. (1999). "Context Interchange: New Features and Formalisms for the Intelligent Integration of Information," *ACM Transactions on Office Information Systems*, July 1999.

[7] Goh, C.H., Madnick, S.E., and Siegel, M.D. (1994). "Context interchange: overcoming the challenges of large-scale interoperable database systems in a dynamic environment," In *Proc of the Third International Conference on Information and Knowledge Management*, pages 337-346, Gaithersburg, MD.

[8] Kakas, A.C., Kowalski, R.A., and Toni, F. (1993). "Abductive logic programming," *Journal of Logic and Computation*, 2(6):719-770.

[9] Kifer, M., Lausen, G., and Wu, J. (1995). "Logical foundations of object-oriented and frame-based languages," *JACM*, 4:741-843.

[10] Seligman, L. and Rosenthal, A. "XML's impact on databases and data sharing," *Computer*, June 2001, p. 59-67.

[11] Siegel, M. and Madnick, S. (1991). "Context Interchange: Sharing the Meaning of Data," *SIGMOD RECORD*, Vol. 20, No. 4, December p. 77-78.

[12] Siegel, M. and Madnick, S. (1991). "A metadata approach to solving semantic conflicts," In *Proc of the 17th International Conference on Very Large Data Bases*, pages 133-145.

[13] Zeng, D. (2001). *Analysis 0f XML and COIN as Solutions for Data Heterogeneity in Insurance System Integration*, MS Thesis, MIT, June 2001.