

CROSS-ORGANIZATIONAL DATA QUALITY AND SEMANTIC INTEGRITY: LEARNING AND REASONING ABOUT DATA SEMANTICS WITH CONTEXT MEDIATION

Allen Moulton, *Research Affiliate, MIT Sloan School of Management*

Stuart E. Madnick, *Professor of IT, MIT Sloan School of Management*

Michael L. Siegel, *Principal Research Scientist, MIT Sloan School of Management*

Introduction

Efficiently integrating new sources of information from outside the enterprise is often critical to success in a world of global competition, interdependency, and rapid market change. Traditional means of assuring data quality and semantic integrity focus on data within the control of the organization (Huang et al., 1999). Without some automated assistance, internal planning and control mechanisms for assuring data quality will have difficulty in responding in a timely manner to changing demands. Within an organization, data can be created, stored, and used by people and computers sharing a common implicit understanding of data semantics. We use the term *context* to refer to this implicit understanding of the relationship between data elements and structures and the real world that the data represents. The *context interchange* problem arises when organizations with different contexts must exchange information (Madnick, 1999).

A context interchange (COIN) mediator is an automated reasoning engine to assist an organization in learning about semantic conflicts between its own receiver context and the contexts of data sources (Goh et al., 1999). Because context definitions are declarative, they need only be prepared once for each source and receiver context (Bressan et al., 2000). Data sources may be relational databases, XML documents, HTML webs wrapped to appear as relations with limited query capability (Firat et al., 2000), or stateless computational procedures. Using declarative context knowledge, a COIN mediator identifies semantic conflicts and designs plans for combining sources with data conversions to meet receiver semantic requirements.

Given a large number of component systems operating in a diversified and dynamic environment, COIN mediation facilitates rapid incorporation of new information sources, dynamic substitution of information sources, extension and evolution of semantics, data representation in the user's context, access to the meaning of data represented, identification and selection of information source alternatives, and adaptation to changes in user and business operations.

Research Motivation

The fixed income securities industry suggests an excellent example of how COIN can help resolve context. Portfolio managers may need to draw upon external sources for data about security characteristics, for market valuation information, and for models and calculations (Moulton, Madnick, et al., 1998). All these sources may need to be combined with internal portfolio holdings data and used by a decision support application system (see Figure 1). Suppose the receiver system requires a security valuation as a "dollar price" expressed as a percentage with fractions in 32nds. If a source offers such a price in the required form, data can be simply sent from source to receiver unmodified. But suppose that the best source for market information offers valuations as "nominal spread" expressed in basis points (100ths of a percent). To meet the receiver's requirements, general industry knowledge and additional data sources must be brought to bear, along with conversion of units and scaling.

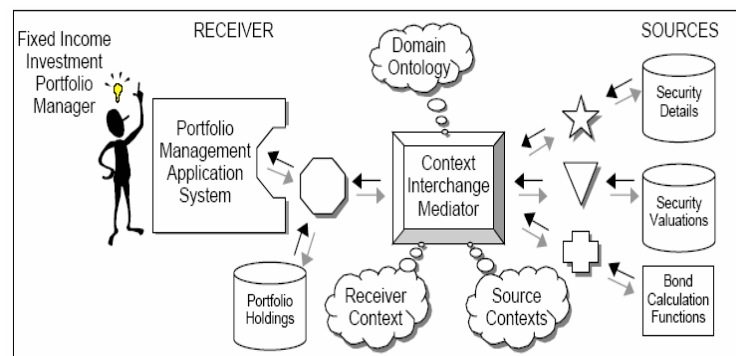


Figure 1: Fixed income securities investment mediation scenario

To resolve the semantic conflict, the mediator must know 1) nominal spread means the difference between yield on a security and a benchmark yield, 2) the on-the-run 10-year T-note yield is an appropriate benchmark provided by another source expressed as a percentage, 3) a bond calculation from another source can convert yield to price given the security's interest rate and other details in factor form, 4) another source provides security details, and 5) methods for converting among percentages with or without 32nds, basis points, and factor form.

Research Strategy

Our research strategy includes investigation of practical problems of semantic integrity and information integration across autonomous sources and receivers, including financial services (Galpaya, 2000), equity securities analysis (Fan, et al., 2000), and fixed income securities investments (Moulton, Madnick, et al., 1998). Based on these industry studies, we have developed prototypes of COIN mediator knowledge representations and reasoning engines. These prototypes and theoretical proofs are used to evaluate approaches to practical problems of semantic interoperability.

Current COIN Mediation Research

Building on earlier work by Goh et al. (1999), we are exploring knowledge representation and reasoning methods to expand the functionality of COIN mediation to include: 1) identifying data representation conflicts and introducing conversions to transform data from source to receiver form, 2) applying domain ontology and context knowledge to map between receiver schema and source schemata, 3) determining when and how to combine sources, feeding data from one source to another with appropriate data representation conversions, 4) deriving missing data by applying domain ontology, context knowledge, or by combining sources. Where possible, we employ a knowledge representation consistent with common system design practices (e.g., UML, E-R, and repositories).

COIN mediation is based on the semantic proposition that interchange of information is possible when sources and receivers share a common subject domain. Sources and receivers are seen as autonomous *implementations* of common subject domain abstractions. Source and receiver system designers make decisions about how to conceptualize abstract constructs and about how to

represent those conceptualizations in data. The mediator uses declarative information about source and receiver contexts, as shown in Figure 1, to devise a plan for integrating sources and data conversions to meet receiver needs.

The knowledge used for mediation consists of: 1) a *domain ontology* containing abstract subject matter conceptualizations that would be known to experienced practitioners and system designers in the industry, and 2) *data models* for each source and receiver with the kind of information programmers would use to access data, 3) *context models* for each source and receiver that explain how each source or receiver data model implements the abstract concepts from a domain ontology.

The framework of a subject domain ontology is a structural conceptual model with classes of abstract objects, attributes of objects, and relationships. Semantic types with modifiers capture alternative data representations (Goh, 1999). Enumerated conceptual categories represent object property distinctions that may be implemented differently by each source and receiver. Rules capture functional relationships among conceptual model object attributes that can be determined from general domain knowledge. Default and contingent rules are used for deriving attribute relationships from partial information, following the same reasoning that industry participants would use.

Data models for relational database sources come from schema and catalog information. For XML sources, data models may be obtained from an XML-schema or DTD or reverse engineered from documents themselves. For HTML sources, data models are provided by web wrapping. For computational procedural sources, arguments and return values are treated as relational attributes in a data model that is augmented with functional dependency and input-output combination constraints.

Context models for each source and receiver explain how each data model implements the general concepts in the domain ontology. Classes from the domain ontology structural conceptual model may be used directly or augmented with context-specific extensions. Context-specific functional or equivalence relationships tie elements of the conceptual model to elements of the data model. For

coding schemes, enumerated attribute domains in a context are mapped to conceptual categories from the domain ontology. Semantic types logically encapsulate data attributes and associate context-specific modifier values to identify the particular data representation used by a source or receiver.

A domain ontology is not a global schema. Rather, it is an abstract representation of the subject matter that each data model implements in its own way. Neither sources nor receivers need to accept the domain ontology as the “right way” of representing information about the subject matter at hand, avoiding some of the practical user acceptance problems noted by Moulton, Bressan et al (1998). By allowing each context model to extend the domain ontology and to explain how context-specific concepts map to general domain ontology concepts, mediation is facilitated without imposing the rigidity of view-based systems. The mediator uses domain ontologies and context models internally within its reasoning process without exposing them to sources or receiver. The mediator accepts relational queries against a receiver data model and writes a query plan that uses only source data models, in effect designing a customized view.

Conclusions

Context interchange mediation brings automated methods to the important task of assuring that data exchanged across organizations can meet the data quality and semantic integrity requirements of the receiver – and do so without requiring the source organizations to accommodate the needs of the receiver, or the receiver to adjust to either sources or the mediator.

For more information on this topic, please visit the COIN website at: <http://context2.mit.edu/coin/> or the MIT Center for eBusiness website: <http://ebusiness.mit.edu/>.

CENTER FOR EBUSINESS MISSION

Founded in 1999, the Center for eBusiness is the largest research center in the history of the MIT Sloan School. Our research is supported by the National Science Foundation and corporate sponsors. We fund more than 45 faculty and more than 60 research projects. Our mission is to be the leading academic source of innovation in management theory and practice for eBusiness.

Examples of Current Focused Research Projects:

- Theory T: Trust-Based Marketing
- Implications of e-Commerce for New Services and Structure of Logistics Systems
- How Do Intangible Assets Affect the Productivity of Computerization Efforts?
- Wireless and Mobile Commerce Opportunities for Payments Services
- Two-Tier Support Business Models
- The Impact of the Internet on the Future of the Financial Services Industry
- Pricing Products and Services in the High-Tech Industry

The Center for eBusiness has recently entered into Phase II, adjusting its agenda to focus more explicitly on business value, while at the same time including technologies beyond the Internet in its purview. The early period of exploration and experimentation is coming to an end and there is now the opportunity, and the necessity, to focus more explicitly on using digital technologies to deliver measurable business value. Amidst all this change, the business fundamentals of investment, revenues, expenses, profits, and satisfying customers have only grown more important. At the same time, a broader, inter-related set of technologies is at our disposal. While the Internet has been an important catalyst, related digital technologies are often at least as relevant.

We are co-located with MIT Sloan's Center for Information Systems Research initiative and the Center for Coordination Science to facilitate collaboration. We also collaborate with the Media Lab and the Program on Internet and Telecoms Convergence.



The Center for eBusiness gratefully acknowledges the support and contributions of its Sponsors.

CENTER FOR EBUSINESS SPONSORSFounding Sponsors

BT
 General Motors
 Hewlett-Packard
 Intel
 MasterCard International
 PricewaterhouseCoopers
 Suruga Bank
 UPS

Research Sponsors

CSK
 France Telecom
 Nortel Networks
 Qwest Communications

Member Sponsors

Amazon
 Bank of Tokyo-Mitsubishi
 Cisco
 Citigroup
 GEA
 Publicis Technology
 SAS

CONTACT INFORMATION

Center for eBusiness at MIT
 MIT Sloan School of Management
 3 Cambridge Center, NE20-336
 Cambridge, MA 02142
 Telephone: 617/253-7054
 Facsimile: 617/452-3231
<http://ebusiness.mit.edu/>

Glen L. Urban, Chairman
 Erik Brynjolfsson, Director
 David Verrill, Executive Director
 Steve Buckley, Communications & IT
 Robynne DeCaprio, Program Administration
 Associate