# Semantic Interoperability in the Securities Industry: Context Interchange Mediation of Semantic Differences in Enumerated Data Types (WEBH)

Allen Moulton, Stuart Madnick, Michael Siegel

CISL WP #02-10

**MIT Sloan School of Management**
**50 Memorial Drive**
**Cambridge, Massachusetts 02142-1347**

# Semantic Interoperability in the Securities Industry: Context Interchange Mediation of Semantic Differences in Enumerated Data Types

Allen Moulton Stuart E. Madnick Michael D. Siegel
*MIT Sloan School of Management*
*amoulton@mit.edu smadnick@mit.edu msiegel@mit.edu*

## Abstract

*Using securities industry examples, the context interchange mediation knowledge architecture is applied to interoperability problems for enumerated data types, such as codes and other symbols used to represent conceptual distinctions. Ongoing efforts in the securities industry to develop new XML-based standards for information interchange are examined. Using components representing similar securities information, drawn from different but complementary securities standards and sources, example problems of information interoperability are examined. We show that transforming data representation into an autonomously specified context model and thence into a general domain ontology allows successful interoperability in several ways depending on how each context is explained to the mediator.*

## 1. Introduction

The context interchange mediation architecture in [12] can be applied to the resolution of semantic differences in enumerated type data that occur frequently in securities industry, and most other commercial, databases. Codes, symbols and other enumerated type data are used to represent a wide variety of properties of securities, transactions, issuers, counter-parties, and virtually any kind of conceptual entity that a data object may represent. To mediate between one system of enumerated codes used by a source and another system of codes used by a receiver, our architecture interposes a subject domain ontology containing conceptual classification schemes that each system of enumerated data codes implements. Using this declarative knowledge, a context interchange mediator designs code conversion tables or procedures to enable a source to provide semantically valid information to a receiver.

With the Internet providing ubiquitous physical data connectivity and XML becoming the standard protocol for structuring web content and transporting structured data across the web, interoperability of diverse and autonomous information services will depend more on semantics than on the technical feasibility of moving data.

XML, in essence, offers the data equivalent of Star Trek's teleportation: a hierarchically structured source object can be serialized, transported over the net, and reconstituted in its original form by the receiver software.

As discussed in [8], access to data is not enough unless the meaning of the data is correctly interpreted and applied by the receiver. We use the term *context* to refer to the implicit understanding of the meaning of data – the relationship between data elements and structures and the real world that the data represents. If receiver and source share a common context, as within an organization, access to data is sufficient for interoperability. If not, then something must be done to either change one or both contexts – or to insert an intermediary conversion of data across contexts.

The traditional approach to obtaining a shared context is the adoption and enforcement of data *standards*, either within an organization or across organizations. More recently, subject domain *ontologies* extend data standards to incorporate other kinds of general knowledge shared by members of a community. The efforts of ISO TC68/SC4 WG10 in the securities industry are directed toward achieving both XML data standards and the rough equivalent of an ontology for the industry.

Data standards and shared subject domain ontologies offer the prospect of interoperability as long as sources and receivers adopt the same standard and understand the meaning of the elements of the standard in the same way. There are, however, a number of problems that stand in the way of such an information utopia. First, standards almost never appear in the singular form, with competing and overlapping standards at a given time and different versions of a given standard evolving over time. Second, a standard must be adopted by all interacting parties and must be enforced to be effective. Third, each organization that adopts a standard for communicating with others must still work out the interaction between the standard and its own internal processes, systems, and data. Different interpretations of a common standard leave an unresolved semantic interoperability gap.

## 2. Evolving securities industry standards

Rapid, reliable, and meaningful interchange of information is a vital element of the modern securities industry. Before a trade is agreed to, information about offerings and requirements is interchanged among traders and salesmen at dealer firms and portfolio managers at investment management firms. After a trade, the details of the sale are exchanged and matched by buyer and seller, agent banks representing each side, and many interested parties. Interchange and integration of information about transactions, markets, and portfolios is also needed by investors, plan sponsors, beneficiaries, regulators, auditors, and many other parties.

In the past four decades, the securities industry has moved from paper documents and manual processing to increasing degrees of electronic records and automation. The goal is global straight through processing (GSTP):

> "The current cross-border trade-processing environment is characterized by manual procedures, multiple service providers, incompatible databases, lack of standardization, relatively high error rates, excessive costs and a relatively high rate of expensive trade failures....
>
> "The primary objectives...are to accelerate the flow of cross-border trades information, to reduce the number of failed cross-border trades, and to reduce the risks and the costs of cross-border trade settlements.[17]"

The need for standards for financial information interchange has been recognized for many years. The United Nations established ISO TC68 in 1948 as the international standards body for financial services. In 1973, the financial services industry established SWIFT[18] as an industry-owned cooperative to provide a secure network for electronic interchange of financial transactions. Now based in Belgium, SWIFT has grown into a network handling 1.5 billion messages for 7,000 financial institutions in 196 countries involving $6 trillion. Historically, SWIFT has been primarily focused on money transfer transactions and on securities transaction message flow between trade and settlement.

SWIFT is both a network and a standard protocol for messages sent on the network. SWIFT expects to transition from its own X.25 network to Secure IP built on top of the internet. At the same time, SWIFT is redesigning its current ISO 15022 message protocols into new SWIFTML XML standards intend to be:

> "... developed more scientifically and tak[ing] into account the full business domain—including all its players, processes and business interactions. As a result, the new standards fit into the end-to-end business transaction even though the actual messages might focus on only a part of the transaction chain.[21]"

FIX Protocol is a consortium founded by bond industry firms in 1993 with an emphasis on supporting pre-trade electronic information interchange. FIX is developing its own FIXML standard building on its existing "tag=value" protocol for fixed income securities. Another industry-sponsored group, fpML.org, is working on devising XML standards for the complex and constantly evolving derivatives business.

Recently, SWIFT and FIX Protocol have announced that they intend to merge their separate XML efforts into common ISO 15022 XML standards under the auspices of ISO TC68/SC4/WG10[20], was created in 2000 to

> "evolve ISO 15022 to permit migration of the securities industry to a standardized use of XML, guaranteeing interoperability across the industry..."

The primary efforts involve developing a common conceptual business model in UML, reverse engineering industry requirements from existing standards, such as SWIFT, FIX, and fpML, and developing a common repository of transactions and data definitions.

Fixed income securities analytic calculations have also been the subject of standards efforts under the auspices of the Securities Industries Association (SIA) and the Public Securities Association (PSA). These standard securities calculation methods are now offered as software and as a web service by TIPS, Inc.[19]

Ongoing securities industry information standards efforts offer the prospect of an increasing degree of well-organized documentation and codification of knowledge relating to securities transactions. As standards evolve, interchange of information should become smother. As described above, however, semantic problems will remain. These problems are illustrated by this note appearing in a document from the Swiss Exchange SWX:

> "Note that the ISMA-99 rules given in this document are the **SWX interpretation** of the ISMA rules, and cover exceptional cases not fully addressed by those rules – this can mean that the accrued interest calculated by SWX may differ from that calculated in other markets under exceptional circumstances." [16]

## 3. The role of mediators

Our research focuses on the semantic problems that exist in the absence of standards and that may remain even after standards are adopted. We are investigating *context interchange (Type-C)* mediation which takes declarative statements of source and receiver semantics and devises plans for meeting the receiver's requirements without requiring either side to accept (or even know) the other's data representation[2, 5, 8]. We use the term *context* to refer to the implicit understanding of the meaning of data – the relationship between data elements and structures and the real world that the data represents.

Other mediation research takes a different approach. Type-W mediation offers a pre-designed mediator global ontology, into which sources are mapped through wrappers[13]. Type-M mediation begins with source ontologies and merges them into a global ontology or schema offered by the mediator [7]. Both Type-W and type-M mediators decouple the receiver from sources, but each requires the receiver to learn about and adapt to the mediator's global ontology. The mediator provides interoperability while reducing source-receiver combinatorics and preserving source autonomy. Type-C mediation goes one step further, preserving receiver autonomy as well by explicitly representing and reasoning over receiver semantic requirements as well as source specifications.

In [12] we described a three-tier knowledge representation architecture for Type-C mediation using examples from fixed income securities investment. We now examine an important issue briefly alluded to in that paper: semantic differences in enumerated data types – the codes, mnemonics, and symbols used to represent categories, classifications, and other similar properties of entities in databases, web documents, and programs.

The lowest tier is the *data model*, the data structures and data value domains used by the source or the receiver. The data model tier of information needed by the mediator can be obtained from the catalog of relational database sources or from the XML Schema or its equivalent for XML sources. For receiver requirements and computational sources, the data model is taken from specifications or documentation.

Data structures and data values in combination implement the semantics of the conceptual model of the subject matter used in a source or receiver context. Earlier work by Goh et al. [5] examined Type-C mediation for resolving semantic differences representation of Gregorian dates and in units and scaling of data values representing numeric measures. We focus here on an additional problem of mediating semantic differences in enumerated data types.

In a relational database, an enumerated datatype may appear as a attribute value domains or as a foreign key integrity constraint with an associated table of allowed values. For ontologies using DAML+OIL [15] a "OneOf" daml:collection construct that can be used to specify an integrity an enumerated value constraint.

For XML sources, section 4.3.5 of the XML Schema part 2 specification of datatypes [14] provides for an enumeration constraint, e.g:

```
<simpleType name="daytype">
  <restriction base='integer'>
    <enumeration value='1'/>
    <enumeration value='7'/>
    <enumeration value='8'/>
    <enumeration value='10'/>
    <enumeration value='13'/>
  </restriction>
</simpleType>
```

The enumeration tag in XML Schema defines an integrity constraint that can be used to validate a document. But, other than documentation attached as an annotation, there is no way to tell what the enumerated values represent. Relational database attribute value integrity constraints play a similar role, validating data and giving a list of possible values, but not defining the meaning of those values. In many cases, enumerated type values are not explicitly implemented at all and must be obtained from documentation or empirically observing data instances.

The second tier of our architecture, the *context model*, adds semantics to the data model of a source or receiver. Each enumerated value is mapped one-to-one to a semantic construct in the context model, which in turn may be translated into a conceptual construct drawn from the *domain ontology* in the third tier of the architecture.

We require that the context specification of each source and receiver be declarative and contain no information about any other context. We also require that the domain ontology contain no data values that appear in any source or receiver.

## 4. Example securities mediation cases

A few examples from the fixed income securities industry will illustrate how this multi-tier knowledge architecture facilitates the reasoning needed for mediation in a manner analogous to the way human analysts perform the same task. We will explore a simple case involving methods of calculating accrued interest for fixed income securities. In brief, when an interest bearing security is bought or sold, market convention requires the buyer to pay the seller a portion of the next interest payment, which will be received by the new owner in full on its due date. If the applicable convention is known, the buyer and seller can each apply the rule to calculate the exact amount of accrued interest and thereby know in advance the amount of money to be exchanged at settlement.

Assume we want to use the TIPS Ficalc.com standard bond calculator to calculate accrued interest or discount on a trade in a fixed income security. TIPS describes the FiCalc.com as:

> "a browser-based user interface to the Standard Securities Calculations Server (SSC Server). The SSC Server, built by TIPS, Inc., makes available to internet and intranet applications all of the power and functionality of the Standard Securities Calculations Software Library."[19]

The first problem we consider is how a mediator might figure out that a value of "A001" in the field called "22F::MICO" in an ISO 15022 transaction message should be converted to a value of "10" for an attribute called "dt" in Ficalc. The second problem is how the mediator can determine that a "dt" value of "1" should be used when the the "Security Type" element has a value of "NOTE" in an XML auction record from the US Bureau of the Public Debt web site.

Interest accrual conventions are often called “day count” rules, since the usual method is to apply a rule for counting “accrued days” since the last interest payment. We are not concerned with how the accrual calculation is done here, but with identifying the convention to be used. From general industry knowledge our ontology might contain this partial list of accrual conventions:

| Table 1. Ontology: accrual conventions | |
|---|---|
| mAccr | description |
| ♣101 | 30/360 |
| ♣102 | 30/365 |
| ♣103 | 30/Actual |
| ♣104 | Actual/360 |
| ♣105 | Actual/365 |
| ♣106 | Actual/Actual |
| ♣107 | 30E/360 – old Eurobond basis |
| ♣108 | Actual/M |
| ♣109 | Actual/365L |
| ♣110 | Japanese/365 |

Note that for expository purposes we show a distinct ID called ‘mAccr’ for each convention.

The receiver context is the FiCalc web bond calculator. Table 2 shows the allowable values for the day type attribute “dt” along with a description taken from the documentation. The left hand three columns (Table 2A) show the symbols used for ‘dt’ along with the description from the documentation and an arbitrary unique context construct ID ‘fDay.’ The right hand columns (Table 2B) show the mapping from the context construct to an ontology construct.

| Table 2. FiCalc day type codes | | | | | |
|---|---|---|---|---|---|
| A-context semantic constructs | | | | B-ontology map | |
| dt | description | fDay | | fDay | mAccr |
| 1 | Actual/Actual | §201 | | §201 | ♣106 |
| 7 | Actual/365 | §202 | | §202 | ♣105 |
| 8 | Actual/360 | §209 | | §209 | ♣104 |
| 10 | 30/360 | §210 | | §210 | ♣101 |
| 13 | Jpn/365 | §215 | | §215 | ♣110 |

Note that, since the receiver is a program, its requirements must be exactly met without the sort of adaptive learning that a human user might do

### 4.1 Direct data value conversion

The simplest case is where both the source and the receiver use enumerated codes for a common conceptual category domain. If both use the same codes, then the mediator should conclude that the data may be passed straight through. If source and receiver use different codes, the mediator should introduce a conversion relation to map one to the other.

We want to calculate accrued interest for a transaction from a SWIFT ISO 15022 message, which represents the accrual convention in a field called “22F::MICO’ that allows the values shown in column ‘xMico’ of Table 3A.

| Table 3. ISO 15022 22F::MICO codes | | | | | |
|---|---|---|---|---|---|
| A-context semantic constructs | | | | B-ontology map | |
| xMico | description | cMico | | sMico | mAccr |
| A001 | 30/360 | §301 | | §301 | ♣101 |
| A002 | 30/365 | §302 | | §302 | ♣102 |
| A003 | 30/Actual | §303 | | §303 | ♣103 |
| A004 | Actual/360 | §304 | | §304 | ♣104 |
| A005 | Actual/365 | §305 | | §305 | ♣105 |
| A006 | Actual/Actual | §303 | | §303 | ♣106 |
| A007 | 30E/360 | §307 | | §307 | ♣107 |
| A008 | Actual/M | §308 | | §308 | ♣108 |
| A009 | Actual/365L | §309 | | §309 | ♣109 |

By taking the inner join of the two contexts using the common mAccr ontology ID, the mediator can conclude that only four of the xMico codes in Table 3A can be converted to FiCalc codes in Table 2A. The mediator can also write the conversion rule for deriving the needed value:

xMico=A001 → cMico=§301 [by Table 3A]
→ mAccr=♣101 [by Table 3B]
→ fDay=§210 [by Table 2B]
→ dt=10 [by Table 2A]

Eliminating the ID’s this becomes:

xMico =A001 → dt=10

The mediator could generate a conversion table to map source values directly to the required receiver values. This table would then be used in the mediated query written by the mediator to meet the receiver’s needs. By using the conversion table and the rewritten query, the mediator itself would be unnecessary at run time.

### 4.2 Indirect conversion with source side inference

Assume now that we want to get security information from the US Bureau of the Public Debt web[22], which does not include an explicitly stated interest accrual convention. Instead, the accrual convention must be inferred from the security type code (column ‘secType’ in Table 4A). One way of doing this inference is to use the source context model to map each security type context construct into the appropriate ontology accrual convention construct (as shown in Table 4B):

| Table 4. Bureau of the Public Debt - source inference | | | | | |
|---|---|---|---|---|---|
| A-context semantic constructs | | | | B-ontology map | |
| secType | description | bStyp | | bStyp | mAccr |
| BILL | US T-Bill | §401 | | §401 | ♣104 |
| NOTE | US T-Note | §402 | | §402 | ♣106 |
| BOND | US T-Bond | §403 | | §403 | ♣106 |

Given the source context model in Tables 4A and 4B and the receiver context model in Table 2B and 2A, the mediator can write the conversion rule:

secType=NOTE → bStyp=§402 [by Table 4A]
→ mAccr=♣106 [by Table 4B]
→ fDay=§201 → dt=1 [by Tables 2B and 2A]

Here the mediator converts a source context value into an alternative pair of receiver context values, which can be substituted for the accrual convention code in FiCalc's derivation of the accrued interest value. Knowing that such as substitution is possible depends on other context and ontology knowledge outlined in [12]. It should also be noted that, while we have illustrated the mappings with tables here, any alternative representation of mapping relations will suffice.

## 5. Conclusion

We have shown how our knowledge architecture for Type-C mediation can be applied to the conversion of enumerated type data, such as the codes used in securities industry standards. This problem fits within the larger scope of transforming the information conceptualization and data implementation of one context into another. Other work addresses the problem of identifying and correlating the conceptual roles of source and receiver data attributes, as well as specifying conceptual relations among attributes associated with one entity or related entities (as in the relation of security type and accrual convention used above).

We are continuing to examine techniques for specifying contexts and ontologies using well established methodologies of business systems analysis and database design [*viz.* 3, 4]. The evolving WG10 securities industry standards will partially solve the interoperability problem, and should also provide detailed substantive information models for building ontologies and context models for additional information interchange requirements [10,11].

Our work complements research such as Bernstein's on generic model management [1] on one side and Mong-Li Lee's work on using conceptual modeling methods to rigorously design data published on the web[9].

The key to Type-C mediation is to see the problem as analogous to engineering design of physical systems, where models and theories drawn from physical science are applied by transforming the target system into the frame of reference of the model and back again. Our context models do the same between data models and domain ontologies. Neither the data nor the ontology need be a view of the other.

Ongoing research includes the specification of logic rules and reasoning algorithm to traverse the analytic process from receiver data model requirements, through receiver context models and subject domain ontologies, thence to potential source context models and data models, devising plans for meeting the receiver's needs from available source data combined with generated conversion relations.

## References

[1] P. A. Bernstein. "Generic Model Management: A Database Infrastructure for Schema Manipulation," CoopIS 2001: 1-6.
[2] S. Bressan, C. H. Goh, N. Levina, S. E. Madnick, A. Shah and M. D. Siegel. "Context Knowledge Representation and Reasoning in the Context Interchange System," Applied Intelligence (13:2), Sept. 2000, pp. 165-179.
[3] Stephen Cranefield and Martin K. Purvis. "UML as an Ontology Modelling Language," Proc. Workshop on Intelligent Information Integration, IJCAI 1999.
[4] Ramez Elmasri, Shamkant B. Navathe. Fundamentals of Database Systems, 3rd Edition. Addison-Wesley, 2000.
[5] C. H. Goh S. Bressan., S. E. Madnick and M. D. Siegel "Context Interchange: New Features and Formalisms for the Intelligent Integration of Information," ACM Trans. on Office Information Systems, July 1999, pp 270-293.
[6] B. N. Grosof, Y. Labrou, and H. Y. Chan. "A Declarative Approach to Business Rules in Contracts: Courteous Logic Programs in XML,". Proc. 1st ACM Conf. Electronic Commerce (EC-99), 1999.
[7] F. Hakimpour and A. Geppert. "Resolving Semantic Heterogeneity in Schema Integration: an Ontology-based Approach," Proc. Int'l Conf. on Formal Ontology in Information Systems, Ogunquit, ME, Oct. 2001, pp. 297 - 308.
[8] S. E. Madnick. "Metadata Jones and the Tower of Babel: The Challenge of Large-Scale Heterogeneity," Proc. IEEE Meta-Data Conf., April 1999.
[9] Mong-Li Lee, Sin Yeung Lee, Tok Wang Ling, Gillian Dobbie, Leonid A. Kalinichenko. "Designing Semistructured Databases: A Conceptual Approach," DEXA 2001, pp. 12-21.
[10] A. Moulton, S. Bressan, S. E. Madnick and M. D. Siegel. "An Active Conceptual Model for Fixed Income Securities Analysis for Multiple Financial Institutions," Proc. ER 1998.
[11] A. Moulton, S. E. Madnick and M. D. Siegel. "Context Mediation on Wall Street," Proc. CoopIS 1998, pp. 271-279.
[12] A. Moulton, S. E. Madnick and M. D. Siegel. "Knowledge Representation Architecture for Context Interchange Mediation: Fixed Income Securities Investment Examples," DEXA Workshop WEBH 2001, pp. 50-54.
[13] G. Wiederhold. "Mediators in the architecture of future information systems," IEEE Computer 25, 3 (Mar 1992), 38–49.
[14] W3C Consortium. "XML Schema Part 2: Datatypes" sect. 4.3.5. http://www.w3.org/TR/xmlschema-2/#rf-enumeration
[15] DAML+OIL Reference Description for Ontology Markup Language. http://www.daml.org/2000/12/reference.html
[16] Swiss Exchange SWX. "Accrued Interest & Yield Calculations and Determination of Holiday Calendars," Nov. 2001. http://www.swx.com/products/eebeuro_rev.pdf, p.5.
[17] Global Straight Through Processing Association. http://www.gstpa.org/gstpa/ngstpap.nsf/GSTPAExecutiveSummary1
[18] SWIFT. http://www.swift.com
[19] TIPS. "Web's Fixed Income Calculator," http://www.ficalc.com/calc.tips
[20] SWIFT. "ISO Working Group 10," http://www.swift.com/index.cfm?item_id=6610
[21] SWIFT. "SWIFTStandards XML project," http://www.swift.com/index.cfm?item_id=41646
[22] Bureau of the Public Debt http://www.publicdebt.treas.gov