# ContextMetrics™ : Semantic and Syntactic Interoperability in Cross-Border Trading Systems

Chito Jovellanos
*forward look, inc.*
*8 Faneuil Hall Marketplace, 3rd Fl*
*Boston, Massachusetts US 02109*
*chito@forwardlook.com*

## Abstract

*This paper describes a method and system for quantifying the variances in the semantics and syntax of electronic transactions exchanged between business counterparties.* ContextMetrics™ *enables (a) dynamic transformations of outbound and inbound transactions needed to effect 'straight-through-processing' (STP); (b) unbiased assessments of counterparty systems' capabilities to support STP; and (c) modeling of operational risks and financial exposures stemming from an enterprise's transactional systems.*

## 1. Introduction

Computer-driven applications utilize transactions (ie, structured collections of discrete data transmitted between sending and receiving systems) to enable end-to-end automation of business workflows and volume-insensitive 'straight-through- processing' (STP).

However, STP has never been satisfactorily attained to-date. For example, the Society for Worldwide Interbank Financial Telecommunication (SWIFT) has, since its founding in the early 70's, fostered a number of standards over the years, such as ISO 7775 and ISO 15022, that describe how various securities and related payment transactions are to be communicated between investment institutions, brokerages and custodial banks. Despite committed and concerted effort by industry participants to comply with both the letter and the intent of the standards, a significant amount of manual reconciliation is still required to process these transactions. The latest industry reports suggest that as much as 36% of securities transactions do not meet the goal of STP resulting in significant trade management and reconciliation costs [1].

The core reasons underlying operational non-performance are the semantic and syntactic differences with respect to how individual companies represent the data elements that populate the industry-compliant transaction formats they generate in the normal course of business. Four decades of experience with message standardization in the global securities industry shows that standards usage steadily evolves away from the idealized norm. Constrained systems capabilities (eg, small mutual funds cannot often afford the data quality requirements for tradeable instrument identifiers such as global securities cross-references) and limited contexts for business workflows (eg, an equity trading firm may not effectively support fixed income derivative business) often result in the naïve application of transaction standards. The variances are further influenced by selective adoption of the standards' features, and the non-stationary nature of business contexts and opportunities (eg, novel derivative products and their supporting workflows).

## 2. Existing Approaches

Commercial implementations of transaction management tools have focused on static data mappings [4, 7]. These result in brittle data management frameworks that require significant manual effort to set-up, operate and maintain. For example, large custodial banks have as many as 1300 trade-related counterparties. Similarly, institutional broker-dealers cater to over 2,000 external OTC derivative customers on average [3]. Manually setting and modifying the underlying data maps for each of these counterparties cannot scale effectively to support this level of operational complexity (eg, 1300 counterparties will require

(N*(N-1))/2 or over 860,000 variants of data mappings to evaluate, develop and maintain; each mapping initiative will require several hundred hours of effort to code, test and integrate).

## 3. The ContextMetrics™ Framework

Our approach to transactional interoperability recognizes that differences in the application of messaging standards will persist, and that the ability to achieve STP is tied intimately to quantifying the gaps that continually arise in the semantics and syntax of transactional data. Doing so allows us to spontaneously define specific data transformations to bridge these gaps. This aspect of our work supports Stonebraker's notion of "integration on demand" [6] rather than "integration in advance" which relies on the premise of *de facto* harmonization through standards adoption.

These variance metrics also enable us to track the quality of the transactions exchanged between counterparties, where quality is defined in the context of each counterparty's ability to send or receive a transaction in support of STP.

Furthermore, by quantifying transaction quality, we can model the operational risks and contingent financial exposures that a firm might be subject to, given the potential for failed or late transactions.

ContextMetrics™ embodies our technique for deriving the semantic and syntactic variances in transactions exchanged between counterparties. This paper describes its prototype implementation ("System") as a service invoked from our client's middleware infrastructure. The client's middleware provided controlled access to transactions, and we ensured that our System cooperated fully with the environmental and behavioral requirements of that infrastructure.

Our System performs the following core software operations :

1. Sample data streams of tagged transactions (NB, not necessarily XML-based) created by a computer system (eg, a firm's trade management systems) prior to outbound transmission to another computer system (eg, the counterparty's payment system). Similarly, complementary transactions from the counterparty's systems (eg, trade confirmation systems) are also inspected. This universe of samples reflects a coherent collection of bi-directional messages that support an existing business-to-business workflow. More importantly, the sample provides an unbiased view of "real-world" utilization of transaction standards, and allows us to identify emergent properties arising from actual usage.
2. Measure the frequencies and correlations of [data field names] and [data values] for each of the data elements within and between sampled transactions. For example, within a transaction, ISO 15022 field tag 35 ("Identification of the Financial Instrument"), is highly correlated with field tag 97 ("Safekeeping Account") which identifies the holder of the financial instrument. Between transactions, field tag 35 will be highly correlated with both field tag 36 ("Quantity of Financial Instrument) and field tag 22 ("Credit/Debit Indicator") in counterparty messages which indicate whether a particular quantity of the financial instrument is being added or removed from the holder's account.

Partitioning the sampled transaction into "Topic Domains" prior to analysis increases the accuracy of the System. Each Topic Domain comprises a discrete collection of data elements that delineate a generally understood and logically coherent subject matter area. Examples of Topic Domains are:

- *client information* consisting of company name, account number, SWIFT or IBAN identifier, etc
- *product information* consisting of securities identifier, description, custody location, etc
- *settlement information* consisting of order date and time, price, currency, etc.

For computational efficiency involving large data streams (ie, greater than 100,000 securities transactions sampled per day in real to 'near'-time), a Topic Domain can be managed exclusively on one physical server. Topic Domain servers communicate with others over the backbone network. Since transactions are sampled and analyzed "in-flight", file I/O is minimal and the bulk of caching and computation is performed in memory.

Based on historical transaction profiles, the System compiles an initial measure of variance in the syntax and semantics of transactions sampled from a sending party's system, versus the semantics and syntax of the data elements in the complementary transactions transmitted by the corresponding receivers' system (NB. a formal exposition of the underlying algorithm will be published in a subsequent paper).

Going forward, new transactions are sampled from the data streams (ie, outbound as a sender, or inbound as a receiver). Results are compared in real-time with the variance metrics from historical samples, and concurrently updates the variance metrics.

The variances in the semantics and syntax of sampled transactions are tracked using a Data Interoperability Grid (DIG). The DIG is a network of multi-dimensional numeric arrays that store :

a) [data field name] correlations within a transaction generated by each active party (ie, for the sender of the transaction and for each of the counterparties it transacts with)

b) [data field name] correlations between transactions (ie, the sender and all its counterparties); and

c) [data value] cluster metrics for each transaction data element across all counterparties. For example, ISO 15022 Field 11 ("Currency of Denomination") will cluster around a specific range of values for those counterparties whose trading business involves only developed economies (eg, United States and Great Britain) and not emerging markets (eg, Malaysia and Latvia).

The variance profile of every data element that makes up a particular transaction is accessible via a Universal Data Structure (UDS). The UDS represents a description of one data element (eg, "Settlement Location") as used with each counterparty (eg, "Goldman Sachs London"). The UDS is a package of Uniform Resource Identifiers (URI) that point to the specific locations in the DIG which contain the statistical information needed to quantify the semantic and syntactic differences in usage of that data element. Although there is a large resulting collection of UDS objects, they are inherently simple to use since each and every UDS is consistent in structure and uniformly addressable by an interrogating program.

As new transactions are received and inspected from incoming or outbound data streams, significant deviations in the metric values in excess of a Data Operability Threshold (DOT) flag a semantic and or syntactic shift (eg, that can then trigger systematic data transformations). The DOT is a mathematical region delineated by the various metrics (captured in the DIG) that establishes a stable optimum for data usage where complete semantic and syntactic interoperability exists between sender and receiver transactions. Once the DOT region is perturbed, the interoperability of the transaction is questionable (eg, requiring automatic transformations of the data elements and transaction structures if STP is to persist).

The DOT is application specific (eg, it differs for UK domestic trades versus those transacted cross-border with emerging markets), and conditioned to a large extent by the semantic and syntactic peculiarities inherent in a business workflow.

## 4. Results and Discussion

Preliminary results from our initial deployment of our ContextMetrics™ System for automating transaction mappings showed promising results. We validated improvements in transactional interoperability by tracking STP rates for corporate actions (eg, notifications for stock splits, mergers and other events that affect the capital structure of a company), and trade processing transactions (eg, to effect a change in ownership of stock on receipt of payment). Our results are summarized in Table 1 below.

**Table 1. Comparative summary of STP Rates achieved using a commercial mapping product vs our System. Evaluations were performed on a parallel production environment. STP Rates defined as the percentage of transactions that did not require manual handling or correction. Sample Size shows average daily number of transactions and daily peaks (in parentheses). Results based on a 60 business day period (Apr - Jun 2001). Data based on production transactions (ISO 15022) provided by a global investment institution interacting with their regional broker-dealers and custodial banks. The "training data" set consisted of the samples from the first five days, and were excluded from the "result set" which comprised the remaining 55 days.**

| | STP Rate : Standard Automation | STP Rate Using Our System | Sample Size |
|---|---|---|---|
| **Corporate Actions** | 64% | 97% | 1,214 (3,146) |
| **Securities Settlement** | 86% | 98% | 40,293 (180,680) |

Compared to existing data mapping technologies, usage of our System resulted in significant STP

gains of approximately 33% for corporate actions and 12% for settlement instructions.

We observed that throughput rates for corporate actions (which are semantically far richer than settlement instructions) required significant machine resources. Seven Topic Domain servers were used, versus three for settlement instructions. Each corporate action transaction required 1.46 seconds on average to process, while each settlement instruction only took 0.21 seconds on average. Along the same vein, initializing the System for corporate actions required 33.5 hours based on the training set, while settlement instructions only required 11.2 hours in total. As such, optimizations to the underlying algorithm are areas of ongoing development work with ContextMetrics™ .

We also observed random but significant DOT value shifts. Inspection of the source data indicated that these aberrations were due to nonsense errors arising from manually input records.

ContextMetrics™ has also been applied to ranking the quality of transactions generated by a firm's counterparties. Intuitively, there is a strong correlation between data interoperability and transaction quality, and this was borne out by our observations. Notably, there was no consistent counterparty that was ranked as the "best STP performer" across the board. Rather, transaction quality was a function of the securities market that the counterparties participated in, reflecting the breadth and maturity of the transactional data pool and the relative stability of the underlying workflows.

The ability to generate rich and complex data structures was no guarantee of transaction quality either. Interoperability required that the sophistication of the transactions exchanged between counterparties matched their mutual operational capabilities. In many instances, we observed sub-custodians and broker-dealers with sophisticated systems infrastructures who fared poorly in the interoperability rankings with select investment institutions. They simply overwhelmed those institution's systems with data that the institution could not readily process. In particular, conveying information related to securities positions (eg, on loan; encumbered; traded; settled) was particularly difficult to manage in an STP context for many of the smaller institutions.

Longer-term, ContextMetrics™ will also be used in a financial analytic package for quantifying and hedging operational risk [2], and estimating the contingent financial exposures for firms engaged in securities trading. In brief, the probability of transaction 'fails' (based on the semantic and syntactic gaps in the transaction flows for a given firm) will be used to model impacts on the firm's cash flows and other working ratios. This use of ContextMetrics™ represents one of the nascent applications of data engineering to the regulatory and fiscal imperatives of operational risk management.

## 5. References


[1] Bryant A. 2002. Presentation at SWIFT SIBOS Conference, Geneva, 30 September 2002.

[2] Cruz M. 2002. Modeling, Measuring and Hedging Operational Risk. Wiley. 330 p.

[3] International Swaps and Derivatives Association, Inc. 2003. ISDA 2003 Operations Benchmarking Survey. 34 p. http://www.isda.org

[4] Medjahed et al. 2003. Business-to-Business Interactions: Issues and enabling technologies. VLDB Journal 12: 59-58.

[5] Rahm E and Bernstein PA. 2001. A survey of approaches to automatic schema matching. The VLDB Journal 10:334-350.

[6] Stonebraker M. 2002. Too Much Middleware. SIGMOD Record 31:1 March 2002.

[7] Yan LL, Miller RJ, Haas LM and Fagin R. 2001. Data-Driven Understanding and Refinement of Schema Mappings. Proc ACM SIGMOD Conf 2001: 485-496.

[8] Hagen D. 2003. IT Risk Assessment in the Context of Basel II. Presentation at the Operational Risk Forum (13 May 2003).