Semantic Information Management

David G. Schwartz
Bar-Ilan University, Israel

Zvi Schreiber
Unicorn Solutions Inc., USA

INTRODUCTION: THE ENTERPRISE DATA PROBLEM

The need to manage enterprise data has been coming into increasingly sharp focus for some time. Years ago, data sat in silos attached to specific applications. Then came the network, with data becoming available across applications, departments, subsidiaries, and enterprises. Throughout these developments, one underlying problem has remained unsolved: Data resides in thousands of incompatible formats and cannot be systematically managed, integrated, unified, or cleansed.

To make matters worse, this incompatibility is not limited to the use of different data technologies (e.g., flat file, COBOL, IMS, Relational, XML, etc.) or to the multiple different “flavors” of each technology, such as the different relational databases (Oracle, DB2, SQL Server, Sybase, etc.) in existence. The most challenging incompatibility arises from semantic differences. Each data asset is set up with its own worldview and vocabulary—known as its schema. This incompatibility exists even if both assets use the same technology.

For example, one database has a table called “client,” intending this to include channel partners, and subdivides customers into individuals and institutions; the other data asset refers to the same concept as a “patron” (although not including channel partners) and subdivides “patrons” into individuals, corporations, government, and nonprofit groups. To make matters worse, “patron” excludes international clients, despite the fact that this is not explicitly mentioned in any documentation and the original developer retired five years ago.

In a larger enterprise, this problem may be multiplied by thousands of data structures located in hundreds of incompatible databases and message formats. And the problem is growing; enterprises continue to acquire subsidiaries, reengineer, and integrate with partners. Moreover, developers are continuing to write new applications and to create new databases based on requests from business users without worrying about overall data management issues.

Therefore, it is imperative to find an efficient way to manage multiple applications and data sources. This requires a shift from an application-centric view of IT to a data- and information-centric view. And it requires a focus on the business meaning—or semantics—of the data. Focusing on a singular business meaning results in the effective use of a single language throughout the enterprise—the common business language approach.

BACKGROUND

Impact of the Data Problem

The enterprise data problem has a strong measurable impact on a company’s bottom line (Schreiber, 2003a).

First, the fragmented data environment inevitably leads to business information quality problems, causing businesses to mishandle information, customer relationships, and internal operations.

Second, the data problem creates a situation which all but prevents the agility that is critical to a modern enterprise responding to a constantly changing environment. Whether application deployment, business process reengineering, or mergers and acquisitions are involved, IT is unable to respond to a dynamic environment due to the fragmented and delicate nature of data assets and hard-coded scripts keeping assets communicating with one another. This is a significant barrier to moving towards the real-time—or zero-latency—enterprise (Khosla & Pal, 2002).

Finally, IT remains unnecessarily inefficient so long as it lacks a strategic approach to data management. In the meantime, IT deals with the frustrating and costly challenge of administering databases (some of which are redundant), mapping each database multiple times and writing and manually maintaining point-to-point translation scripts.

Data needs to be systematically managed if it is to have long-term value to the enterprise. Specifically, data management must start by elevating data into information by explicitly capturing the meaning and context of the data. This process is known as data semantics (Sheth, 1995).
An Overview of Existing Approaches

A partial solution is a metadata repository containing information about each data asset. Repositories act as catalogs and include technical information about the assets, their structure, how they are used, and who is responsible for them (DeMarco, 2000). But a passive catalog does not provide a formal understanding of data or automated support for translating and cleansing the data.

A tactical solution to data quality (Strong, Lee, & Wang, 1997a, 1997b; Wand & Wang, 1996) involves data profiling and data cleansing. While these are important tools within the overall approach to data management, enterprise information quality will never be achieved one database at a time. The real problem lies in ensuring that there is an agreed-upon understanding of each data asset and of its relationship to other data assets. Data assets must be designed, created, maintained, integrated, and decommissioned with attention to the wider quality aims. This is accomplished by understanding data as part of an overall information architecture (Schreiber, 2003b) and by understanding and validating the data with respect to a single agreed-upon business worldview and a series of business rules.

Data modeling typically supports better database design. But data models are usually entity-relationship diagrams (Chen, 2002) with limited business depth, lacking generalization/specialization capabilities (also known as inheritance, subtyping, or polymorphism), and not capturing business rules. They are also technically suitable only for modeling relational databases one at a time, while IT increasingly uses XML, especially for messaging. Furthermore, data models are typically tightly coupled to a specific relational database and are not generally used to promote a common understanding of multiple data assets.

SEMANTICS: TOWARDS A COMMON BUSINESS LANGUAGE

Companies will always struggle with a large number of physically different data formats. While a common data format will likely never be achieved, the key to efficiently managing data is to establish a common understanding. This is the idea of semantics, bridging nomenclature and terminological inconsistencies to comprehend underlying business meaning in a unified manner. Semantics can be achieved by formally capturing the meaning of data. This is accomplished by relating physical data schemas to concepts in an agreed-upon model of the business. Wand and Wang (1996) present some of the foundations for building improved data quality on an ontological basis.

This central model of the business is called an information model (Sheth & Kalinichenko, 1992). The information model does not reflect any specific data model but rather reflects the agreed-upon business view, business vocabulary, and business rules which will provide a common basis for understanding data.

Semantics builds upon traditional informal metadata (Bernstein & Bergstaesser, 1999) and captures the formal meaning of data in agreed-upon business terms. For example, the information model might capture the official business concepts of a “customer” and the more specific concepts of a “business customer” and an “individual customer.” A semantic mapping will then relate physical data schemas to this information model. For instance, a semantic mapping might capture the fact that an information model has a universally accepted concept called “individual customer” that is called “client” by a relational database table, “customer” by an XML schema, and “CUST3” by a COBOL copybook. Thus the semantic mapping formally captures the meaning of the data by reference to the agreed-upon business terminology.