Digging Into The Web:

XML, Meta Data and Other Paths to Unstructured Data

By Robert Blumberg and Shaku Atre

This is the fourth in a series of articles discussing various aspects of unstructured data.

While the Web is widely recognized as an amazing source of unstructured data, much of this data is rather difficult to search and navigate. In fact, the Web was organized along a model meant for human consumption, not for optimizing machine searches. HTML (hypertext markup language), the means by which hypertextual information is organized for the Web, is a presentation language. It concerns itself with the appearance of data rather than its underlying structure. This makes the process of extracting content from the Web a daunting task for automatic processors. For this reason, there's a widespread sense that unstructured data on the Web represents a great untapped value.

However, as Laura Ramos, director of research at Giga Information Group, has said, "Although no one disputes the value contained in unstructured information, the work and expense required to add structure to the data or to make it consumable and searchable by workers is significant." Ramos has added, correctly, that because the return on investment (ROI) for unstructured-data projects is difficult to calculate, many organizations are instead putting their limited technology dollars toward other types of applications.

Even though the ROI of mining unstructured data is unproven, given the potential value of this data, organizations should consider embarking on this strategic journey. In today's economy, making unstructured data available to decision-makers at all levels of the organization is required just to remain competitive.

Unstructured data, viewed correctly, is really *semi*-structured data because nearly all information exists with at least some context. Consider, for example, an e-mail message. Yes, the body is unstructured. However, an e-mail also contains contextual information such as the message's subject and date of delivery, as well as other embedded information, including the addresses of both the sender and the receiver.

It is not feasible to identify, convert, analyze and consume all of the unstructured information in an enterprise in one gigantic step. Instead, a more prudent approach is to phase the information areas into the decision-support infrastructure based on available budget and the expected impact of an information area on the overall business intelligence (BI) process. As illustrated in Figure 1, there are four phases in the road map for converting semi-structured information into BI.

The Journey Begins

Many companies are overburdened with information tucked away in hard-to-find places. To stay competitive, these organizations must create a strategy for identifying information that can help them make better-informed decisions faster. The problem is, how can they identify the information when they may not even know the information exists? Or, if they do know it exists, how can they be sure to appreciate the information's importance?

These are some of the challenges that many organizations face while identifying and assessing the importance of a specific data item. Given the common requirements and challenges, several vendors – including Autonomy, Verity, and Convera – now provide built-in dictionary and thesaurus support. Another vendor, Stratify, leverages taxonomy relationships to provide certain thesaurus capabilities. Knowledge-workers using this tool can enter English-language queries to find information targeted to their specific context. Then the system automatically enhances these queries with a set of related terms and concepts. Other features of the Stratify solution include the ability to recognize idioms for more accurate searches.

Functionally, most of these products are similar; thus, they usually differentiate themselves on other factors, such as total cost of ownership (TCO). Where one product may have a relatively long implementation cycle, another may require more operational and administrative training. These vendors also have developed their niches. For example, Autonomy plays on its strength in organizing unstructured data by appealing to organizations that provide content to knowledge-workers. Verity, on the other hand, emphasizes its ability to access information in a variety of formats from disparate data sources; as a result, it appeals to software developers such as SAP and Sybase, which integrate Verity's search technology into their products. Meanwhile, Convera positions its RetrievalWare flagship product as a solution for accessing large volumes of data in disparate formats and media types. Convera caters to the needs of large corporations, including EDS, Microsoft, and the U.S. Navy. Finally, Stratify promotes its ability to manage the total taxonomy life cycle, including the ability to create, define, test, publish and refine taxonomies. It is also working with the CIA to support the classification of large amounts of unstructured data.

Meeting Meta Data

A key role of enterprise content-management products is to describe business data accurately and consistently in meta data so that it can be found with relative ease and speed and then acted upon. This combination of accuracy and consistency can be achieved when each content-management system makes reference to a carefully designed, well-maintained taxonomy. If the taxonomy is not kept up to date, then content management cannot function.

The need to make unstructured data accessible by machines has been understood for years. In fact, it is in this context that the Semantic Web initiative started back when the first proposals to standardize XML were taking place. In 1998, Web pioneer Tim Berners-Lee of the W3C published a seminal work, entitled "Semantic Web Road Map" (http://www.w3.org/DesignIssues/Semantic.html), in which he states that most information on the Web has been designed for humans, not automatic processors. Therefore, Berners-Lee then asks, rather than instructing machines to interpret content not immediately comprehensible, why not instead develop new languages that can express information in a more easily machine-processable format?

Clearly, the introduction of extensible markup language (XML, http://www.w3.org/XML/) has provided a convenient notation for organizing Web content. When using XML, applications need not analyze how data is presented. That's because the intrinsic data structure is described by means of meta data tools such as DTDs (document type definitions) and XML schema documents. Also, data transformations from one context to another can be achieved easily by using declarative tools such as XSLT (XML stylesheet language for transformations, http://www.w3.org/Style/XSL/).

One issue addressed by XML is the fact that different content-management providers use different techniques for storing and managing their meta data. Meta data repositories are supposed to store descriptive information on business and operational data that is common across an organization or industry segment. However, dissimilar management techniques make it difficult for knowledge-workers to access this meta data in a consistent manner. To resolve this issue, the industry is moving toward making XML the de facto standard for creating meta data.

XML is simply an open standard for describing data. As defined by the W3C, XML can be used for defining data elements on both Web pages and business-to-business documents. XML uses a tag structure similar to HTML, but with a difference: Where HTML defines *how* elements are displayed, XML defines *what* those elements contain.

Another difference: While HTML producers must use predefined tags, XML developers can define their own unique tags. As a result, virtually any data item may be identified. This, in turn, lets a Web page function somewhat like a database record. In this way, XML provides a common method for identifying data. This is especially useful in business-to-business transactions.

XML has been further developed by several vendors, mainly to provide intelligent defining of these tags and common adherence to their usage. Two examples are cXML (commercial XML) from Ariba and CBL (common business library) from Commerce One; they're among the first XML vocabularies for business data. Another, DSML, is a set of XML tags that defines the items in a directory.

XML tags are defined in an XML schema, which defines content type as well as name. XML tags can also be described in the original SGML DTD format because XML is actually a subset of the SGML language. In addition, several Web sites provide repositories for publishing and reviewing XML schemas. These XML repositories are simply Web sites that serve as central storehouses for publishing and reviewing XML schemas.

In fact, the driver behind convergence is XML's ability to provide a common way of labeling digital information. XML does this by tagging the data and setting the context. It also simplifies the process of data exchange between computers and humans. That said, there is some debate over the underlying systems for storing and managing XML-based meta data. The current thinking is roughly split between XML-enabled relational databases and native XML databases. "XML databases will do well in certain applications, such as content-management data stores, and XML-enabled databases will dominate in more data-centric XML applications," says Grant Laing, a senior analyst at Intellor Group, Inc.

However, without the right tools, the use of XML can be inefficient. Because manually created descriptions are subject to their creator's understanding and interpretation of the data, they can become inconsistent. This results in an increased number of inconsistent descriptions for a single data item. When coupled with inadequate relationships, it can result in longer search cycles that produce inaccurate, out-of-context responses.

What's more, companies, institu-

Figure 1: Strategic Steps for Consuming Unstructured Data for Business Intelligence

tions and individuals can create collections of proprietary XML schemas. This makes query executions of XML documents from different XML schemas a difficult task. Although XML schema mappings can be created between schemas, it quickly becomes overly complex to develop vocabularies for many schemas and for many knowledge-workers. In fact, the task is beyond the capabilities offered by XML schema merges and mappings.

Therefore, bringing consistency to meta data and making it easier for knowledge-workers to use is crucial. Software vendors are trying to resolve these common problems by combining content-management and taxonomy-creation capabilities in their products. These products, once equipped with customizable rules, can automatically insert proper XML tags and add appropriate context to categorize and manage meta data for efficient searches and relevant results.

In Search of the Common Meta Data Repository

Yet the work of making unstructured data easy to use is far from finished, and using off-the-shelf packages to create XML-based meta data is only part of the solution. If a truly informed decision-support process is the ultimate goal, then both structured and unstructured data must be integrated into a single context-sensitive set of information. For example, information on a specific customer could reside in customer relationship management (CRM) and enterprise resource planning (ERP) systems, as well as in legal-contract images and e-mail correspondence between customers and the sales and service staff. Some of this data could be stored in relational databases, while the rest could remain in e-mail, word processor and image files. An accurate view of a customer's status would be possible only when all this information can be shown from an integrated view.

However, the tasks of integrating data from disparate sources and integrating different applications are not trivial. In many cases, they may not even be possible. To achieve uniformity, a common, integrated meta data repository is required. This would provide a consistent view of the data to disparate applications and humans alike. Various organizations and committees have recognized this fact and are developing standards to this end. For example, the Object Management Group (OMG) has introduced several meta data-related standards, including XMI (XML Metadata Interchange), MOF (Meta Object Facility) and CWM (Common Warehouse Metamodel).

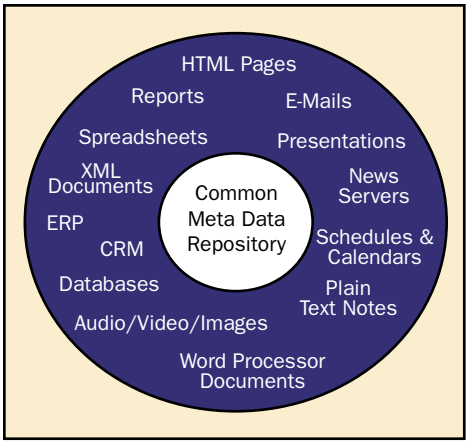

Figure 2: Common Meta Data Repository

As shown in Figure 2, common and consistent meta data repositories are critical to obtaining an integrated view of both structured and semi-structured data. Common, descriptive meta data facilitates the following processes:

- Transformations and relations mapping between disparate data resources.
- Uniform internal schemas in disparate vendor applications, enhancing interoperability.
- Consistent visualization of information across disparate data marts, operational data stores and data warehouses.

RDF: A Framework for Resource Descriptions

One approach taken to engage a more powerful formalism is to use a language to define universal Web assertions. An assertion is the basic way we can state a fact about something. In the semantic Web, assertions are defined using the resource description framework (RDF, http://www.w3.org/RDF/). RDF is the language for representing information about Web resources by means of assertions.

RDF is able to process statements about something in a machine-processable way. The RDF standard itself is strikingly simple. The basic model starts with the concept of *assertions.* An assertion is made up of three parts: a subject, a predicate and an object. In the *subject,* we identify the thing the assertion is about. The *predicate* specifies the property of the subject. The *object* expresses the value of the property.

However, being able to state something about the subject of a statement is not enough if we lack the means to uniquely identify the subjects, predicates and objects in the statements. As part of the Web standards, a generalized identifier is defined called the universal resource identifier (URI, http://www.isi.edu/in-notes/rfc2396.txt). Within RDF, an extension to the URI specification, known as a URI reference, is often used. By employing URI references, the RDF can not only describe practically anything, but it can also state relationships between such things.

How does one use the RDF standard? One application to leverage RDF is the representation of meta data. Here, we include digital libraries catalog (e.g., the Dublin Core, http://dublincore.org/) and the Platform for Privacy Preferences Project (P3P, http://www.w3.org/P3P/Overview.html). In this context, we talk about using the RDF to represent data. Predicate queries are issued on the defined assertions to find all pieces of information that match the query predicates. In other application categories, we may use RDF to help the enterprise solve application-integration problems. Each application specification can be converted into RDF, and queries can then be run over any selection of this data.

A Matter of Semantics

Agreeing on a common notation for describing assertions and unique resource identifiers still does not cover the whole picture. In fact, different data sets may have different identifiers for the same basic concepts. It is critical,

then, that we discover concept similarities whenever and wherever they occur.

To help solve this problem, the semantic Web defines a collection of information called *ontologies.* In this domain, ontologies specify descriptions for classes, or relationships that exist among things or their properties. Because we need automated tools to process ontologies, ontologies are usually expressed in a logic-based language. Ontologies can be critical pieces of information for applications that search the Web. The search program can retrieve only those pages that match a precise concept rather than specific keywords. Web portals could also benefit greatly from ontologies, which would help them to return information on a common topic from different databases.

Yet, according to semantic Web advocates, the real power of this approach will come about when automated programs, also known as software agents, can collect Web content automatically from diverse sources. The semantic Web can enable these agents to automatically integrate information from disparate resources using concepts rather than keywords. Agents should become increasingly effective as their numbers increase and they begin to more closely interact with each other.

Still, according to Tim Berners-Lee, the semantic Web can be much more than a simple tool for conducting individual tasks. Instead, he says, a properly designed semantic Web can actually further the development of human knowledge.

While the challenges and the objectives set for the semantic Web are daunting, this vision could fundamentally change the way we access information. Yet, it also runs the risk of being sidetracked by its own ambitions. For one, the task of adding semantics to the Web may turn out to be prohibitively expensive. For another, people may have become so accustomed to HTML and search engines that they will resist changing to an unfamiliar paradigm, even if it is more powerful. Perhaps a better way would be to take a more gradual approach, one that involves providing Web semantic support to restricted industrial or institutional sectors. So says Giovanni Guardalben, VP of R&D at XML vendor HiT Software Inc. in San Jose, California. "So far, the semantic Web is still very much in the hands of researchers and committees," he says. "If and when business will identify value in the Web semantic approach, we will see rapid adoption of the standards."

The semantic Web has the potential to change the way we interact with the Web – and with each other. Currently, work is underway to realize tools and techniques that will help make this vision a reality.

In the area of RDF development, various meta data initiatives are currently underway. The Dublin Core Metadata Initiative, for example, is a forum that aims at developing interoperable, online meta data standards for different business models. Another effort, the OCLC Connection Initiative, is addressing the concerns of libraries. The Open Directory Project is creating the most comprehensive directory on the Web. Additionally, TAP aims to make the Web a giant distributed database.

For ontologies, a W3C working group is currently defining the Owl Web Ontology Language (http://www.w3.org/TR/owl-ref/). Owl is a semantic markup language for publishing and sharing ontologies on the Web. Its purpose is to provide applications with a tool capable of understanding the content of information rather than just the human-readable presentation.

XML Abbreviations and Their Definitions

DAML: The DARPA Agent Markup Language is an XML language designed to facilitate the concept of the semantic Web. (http://www.daml.org/)

DTD: A document type definition is a set of markup declarations that provide a grammar for a class of documents. The document type declaration can either point to an external document or contain the markup declarations internally. (http://www.w3.org/TR/REC-xml#dt-markupdecl)

OCLC: The Online Computer Library Center is a nonprofit organization serving libraries all over the world. It serves people by providing access to library resources at reduced costs. (http://www.oclc.org/home/)

OIL: The ontology inference language is a proposal for the representation and inference layer for ontologies. It is compatible with the RDF schema language. (http://www.ontoknowledge.org/oil/)

TAP: A system for connecting data fragments from disparate XML/SOAP-based Web services into a single unified knowlege base. (http://tap.stanford.edu/)

XSLT: The XML stylesheet language for transformations is designed for transforming XML documents into other XML documents. It is not intended to be a general multipurpose transformation language for XML, but rather to be used in the context of the general XSL stylesheet language. (http://www.w3.org/TR/xslt)

BI: Analyzing Structured and Unstructured Data

How is the semantic Web going to help BI? Until now, analyzing unstructured data for intelligent answers has not been a focus of BI vendors. Instead, most of their abilities have been limited to mining and analyzing data that is structured. Unstructured data, when these vendors think of it at all, has been an afterthought. By contrast, content-management and taxonomy vendors provide advanced capabilities for searching, cataloging and managing both structured and unstructured data. However, they normally focus on building corporate portals and search engines for their OEM customers. As a result, most of their content-management products lack analytics knowledge and shy away from BI opportunities.

Why should predictive modeling be limited to business applications riding over relational databases? It's only a matter of time before vendors in these

two market segments feel the need for each other's capabilities and start to merge. This should result in the creation of comprehensive product suites for collecting, analyzing and consuming both structured and unstructured data.

To be sure, this will not be easy. As Dipendra Malhotra, director of development at DataMirror in Markham, Ontario, puts it, "To achieve this goal, a paradigm shift is required in the way we handle structured and unstructured data today."

To provide optimal business intelligence, current BI products must evolve to the point where they can efficiently analyze both structured and unstructured data. A prerequisite to this evolution will be the ability of unstructured-data search engines to provide answers to questions, rather than simply providing a list of references. For example, the query "What was Barry Bond's batting average for the 2002-2003 season?" should result in the answer ".328" rather than a list of links to PDF files and HTML pages that one must then sift through to get the answer. Once we reach this milestone, we will place more emphasis on the information itself than on either its source or its format. Terms such as text mining, digital mining and data mining will be replaced by the far more useful *information mining.*

Robert Blumberg is president of Blumberg Consulting, Inc. He has broad experience both as a computer software executive and a creator of leading-edge Internet technology, products and solutions. Previously, he was president of Fresher Information, a DBMS vendor specializing in unstructured-data management. Before that, Blumberg founded Live Picture, where he held various executive positions. Blumberg has been a featured speaker at many industry events and management forums in the U.S. and elsewhere. He may be reached at rblumberg@inpub.com.

Shaku Atre is president of Atre Group, Inc., a consulting and training firm in Santa Cruz, California, that helps clients with business intelligence, data warehousing and DBMS projects. She is a former partner of PriceWaterhouseCoopers and has held a variety of technical and management positions at IBM. Atre is the author of five books, including ***Data Base: Structured Techniques for Design, Performance and Management,*** *and* ***Distributed Databases, Cooperative Processing & Networking.*** *She is most recently coauthor with Larissa T. Moss of a new book,* ***Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications*** *(Addison Wesley, 2003). Atre can be reached at shaku@atre.com.*