Published in DM Review in September 2004.
Printed from DMReview.com

Two Worlds of Data – Unstructured and Structured

by Geoffrey Weglarz

Summary: *The ability to drive the correct visualizations from the marriage of the unstructured world and the structured world is crucial to BPM.*

Google, one of the premier free-form search engines on the planet, may be getting a little skittish about the Microsoft Longhorn project. Google wants to unleash its search technology on the enterprise and on the desktop, but with Longhorn, Microsoft plans to have that capability built into its operating system and its new file system. Is free-form searching really the next battleground? Or is bringing together two needles in a haystack of information really the holy grail of search technology?

In the new category of enterprise software called business performance management (BPM), bringing together the worlds of structured and unstructured data can add significant value to the enterprise. BPM fosters new levels of corporate accountability, financial rigor and tangible value creation across the distributed global organization. It is driven by the imperative to align internal and external constituencies with business objectives through real-time availability and continuous exchange of financial, transactional and operational information. Effectively implemented, BPM enables enterprises to better shape and influence business outcomes by improving the caliber and speed of decision making. With it, executives can anticipate and respond to shifting market dynamics, intelligently allocate and utilize critical resources and consistently meet management and shareholder expectations.

Data in BPM

People use unstructured data every day. Although they may not be aware, they use it for creating, storing and retrieving reports, e-mails, spreadsheets and other types of documents. Unstructured data consists of any data stored in an unstructured format at an atomic level. That is, in the unstructured content, there is no conceptual definition and no data type definition - in textual documents, a word is simply a word. Some current technologies used for content searches on unstructured data require tagging entities such as names or applying keywords and meta tags. Therefore, human intervention is required to help make the unstructured data machine readable.

People also use structured data every day. Structured data is anything that has an enforced composition to the atomic data types. Structured data is managed by technology that allows for querying and reporting against predetermined data types

and understood relationships.

Two Categories of Unstructured Data

Unstructured data consists of two basic categories:

- Bitmap Objects: Inherently non-language based, such as image, video or audio files.
- Textual Objects: Based on a written or printed language, such as Microsoft Word documents, e-mails or Microsoft Excel spreadsheets.

Both of these object types may be classified as data, but the technology and methodology for harnessing relevant information from bitmap objects is still in its infancy. Most of today's technology addresses textual objects. Enterprise content management (ECM) technologies, for example, can help contain unstructured data. Textual data mining and analysis vendors provide analysis tools for unstructured textual objects, and business intelligence vendors supply solutions for querying and analyzing structured data. However, bringing them together - querying both the unstructured and structured worlds - and then associating these two worlds at relevant points is where the most value is gained and also where the highest level of challenge is presented.

Comparing these categories with structured data raises three distinct challenges:

1. Even if unstructured data is in a format such as a Microsoft Word template, the data is still not consumable from a semantic level without a compatible interface or application.
2. Even with a compatible technology, we cannot necessarily gain insight into the context of the information unless we can actually read it.
3. And lastly, the way we interpret what we read is largely subjective.

"A Picture is Worth..."

One of the challenges when dealing with unstructured data is the written word and the fact that it often does not communicate the exact meaning intended. There is a stark division between the written word and the spoken word. The phonetically written word sacrifices worlds of meaning and perception that were once secured in hieroglyphics and still are in the Chinese ideogram. Alphabets such as this provide gestalt - an understanding of the whole within the picture. The Western alphabet lacks the ability to distinguish context and concept from the symbols.

A recent Wall Street Journal article provides a good example of why it is not necessarily appropriate to assign a qualitative value to unstructured data. The Wall Street Journal performed an analysis of a collection of high schools, both public and private, and calculated the percentage of graduating seniors who were accepted to Ivy League schools. At the top of the list was a private school in Brooklyn, New York, called Saint Ann's. Saint Ann's came in first with a whopping 41 percent of graduates gaining admission to 10 of the nation's most exclusive schools, such as Yale, Harvard, Brown, Duke and Cornell. Saint Ann's even beat the Hopkins School in New Haven, Connecticut, where 51 percent of the students in the senior class this year were National Merit Scholars.[1]

The interesting point about Saint Ann's and its success rate comes from the way teachers assign grades to their students. They don't. While at the school, students get written reports about their achievements and areas that need improvement, but there is never a quantitative number or letter assigned to their work. Upon graduation, students receive personal essays about their work from the school's headmaster. Therefore, Saint Ann's college applicants cannot be placed in a "GPA bucket" with other applicants. Each application from Saint Ann's must be read to acquire a full picture of the student. Students from other high schools may well be disqualified at the gate because of a poor GPA - a number assigned to a student meant to represent the quality of knowledge or learning the student has achieved. Perhaps there is some unstructured information that is not meant to have a number value assigned to it as a predicate of value.

Two Approaches to the Problem

There are two ways to approach the question of using this unstructured data. One is from the semantic construction viewpoint. Here, finding word variants and similar words as in search engine usage is inadequate. To differentiate between the word "balance" as a verb or a noun is a feat of semantics and linguistics categorization. This approach requires taxonomies, ontologies and a semantic layer to build concept and category relationships. The second way, which is achievable, is to bring the unstructured parts of the enterprise into the structured world. If we have already identified the context and semantics of our unstructured data, we can bring this information together with our structured data, bridging the two worlds and ultimately providing greater business insight.

Sarbanes-Oxley

A good example of bridging unstructured data to BPM concerns what is happening around the Sarbanes-Oxley Act. Companies are implementing solutions that must bring together the unstructured and structured worlds. Consider, for example, an analyst working on consolidating the Property, Plant & Equipment (PP&E) account for a company. The analyst will notice if the account appears out of tune with the rest of the schedule as it will be highlighted in red. After noting the error, the analyst should be able to bring up a context menu for that account and gain access to a list of relevant items. The error may be a link to the GAAP definition of PP&E, the internal accounts that roll up to that master account, an audit trail of all the ledger entries or even information about the control environment. This is one way to help the enterprise gain access to associated and relevant information that exists - both internally and externally. This further illustrates the strength of structured financial information brought together with unstructured content.

Extensible business reporting language (XBRL) is an attempt to standardize structured and unstructured data. However, it is still difficult to relate the financial schedules with footnotes and the Notes to Consolidated Financial Statements, except to physically read them. XBRL helps solve this problem by tagging discrete elements of unstructured regulatory filings. For example, an analyst could pull all of the contingent liability sections across the current quarter's filings for an industry and perform a side-by-side comparison. Unfortunately, the adoption of XBRL is limited. Until regulators such as the SEC accept and provide analysis mechanisms for XBRL, that type of analysis responsibility will remain a function of human

searching and reading.

Mergers and Acquisitions

Another example is mergers and acquisitions (M&A) analysis conducted through a structured tool. Running numbers and creating models and forecasts is useful, but it isn't enough. The greatest value is gained from related stories and the relationship of those stories to the M&A target as part of the analysis. It is critical to combine the structured number-crunching with the important but less tactile content from unstructured sources. While value can be gained from mining the unstructured world itself, the greatest value is obtained by marrying the unstructured and structured worlds. It is this combination that drives the most significant value to BPM.

Sales Force Automation

One other example of bringing unstructured data to the structured world in the context of BPM is in sales force automation. A sales manager may have a dashboard that tracks his sales pipeline. All of the sales representatives reporting to the manager may roll up to one aggregate forecast. One component or metric in this forecast may be the number of calls made to a prospective client. The dashboard is configured to show a green light for eight or more calls to a client, yellow for four to eight calls and red for less than four. These may be good metrics, but the key is what lies beneath the numbers and thresholds.

If past performance has shown that more calls to a client will result in a sale, then it makes sense to move forward with that metric. But what if it doesn't? What if three calls from a certain sales rep is a good sign? What if the fewer calls that rep makes, the better the chance of a sale? Now, the sales manager can read all of the call detail and get the needed information. What is really needed is an automated analysis of that unstructured information that can produce a symbol, color or some type of visualization to uncover and communicate its meaning. For instance, if the unstructured information contains objections from a client, that information will only become evident when reading the report. Wouldn't it be helpful to have the completed analysis and a link pointing to documents or other data that can answer the objections automatically?

To gain true insight from the data, it is necessary to consider not the number of calls, but the calls themselves. Necessary data includes the details of each call, the tone of the call, the length of the call and the participants on the call. Was it a voicemail, a secretary or an "on a conference call, can you call me back" call? Or, more important, was it a follow-up call detailing a proof of concept in progress that lasted for more than an hour? Obviously, this last type of call contains more value than the three prior calls. However, obtaining the deeper insight means referencing the underlying meaning of the metric.

Solutions and Paths to Success

Enterprise content management systems are now gaining wider adoption, and this provides access to unstructured data and the meta data on top of it. However, it is not possible to look upon that data and grasp it as a whole. We must read the text to gain understanding. Additionally, what we gain from reading is still through our

own lens - the picture each of us conjures may be different.

This is precisely why intelligent systems with semantic layers and taxonomies can connect to the unstructured world. As we define what the documents are and what relationships exist between the unstructured and structured worlds of data, we can bring a unified view, a clear definition and greater gestalt or understanding of business drivers - the heart of BPM.

To put it more concretely, one of the tenets of BPM is the ability to gain insight through measuring results and managing performance. In the May issue of *DM Review*, Dan Sullivan of the Ballston Group commented on some approaches to textual data mining. One point he made is that the choice of a tool for textual mining should include some type of clustering algorithms and visualization tools, such as thematic maps. The ability to drive the correct visualizations from the marriage of the unstructured world and the structured world is crucial. This is altogether the correct approach for starting the textual data mining process on top of enterprise content management systems. The next step is relating those thematic maps to the atomic or the aggregate structured data. The result? We are given the ability to improve our overall business performance by leveraging insight that will drive better business decisions such as which product to build to improve profitability, which customers to target based on insight or modeling a proposed acquisition.

Reference:
1. Bernstein, Elizabeth. "The Price of Admission." Wall Street Journal. April 2, 2004. Page W1.

Geoffrey Weglarz has worked in the information industries for 15 years. He has a background in software development, relational database technologies, multidimensional database technologies and linguistics. He can be reached at geoffrey_weglarz@hyperion.com.