UNSTRUCTURED DATA AND THE ENTERPRISE

by **Paul Williams**

Collaborated with Christine Connors

DATAVERSITY™

DATAVERSITY.NET

TABLE OF CONTENTS

EXECUTIVE SUMMARY

In its most basic definition, unstructured data simply means any form of data that does not easily fit into a relational model or a set of database tables. Unstructured data exists in a variety of formats: books, audio, video, or even a collection of documents. In fact, some of this data may very well contain a measure of structure, such as chapters within a novel or the markup on a HTML Web page, but not a full data model typical of relational databases.

- ANYWHERE FROM 40 TO 80 PERCENT of an enterprise's stored data currently resides in an unstructured format. While most unstructured data is in various text formats, other formats include audio, video, Web pages, and office software data.

- FIRMS ABLE TO SUCCESSFULLY CAPTURE AND MANAGE THEIR UNSTRUCTURED DATA hold a competitive advantage over firms unable to do the same. Over 90 percent of enterprises are currently planning to manage unstructured data or are already doing it.

- METADATA MARKUP, TEXT ANALYTICS, DATA MINING, AND TAXONOMY CREATION are all industry-proven techniques used to capture an enterprise's unstructured data. Included case studies show these techniques in action as part of successful problem-solving projects.

- DIFFICULTY IN INTEGRATION with enterprise systems along with the immaturity of the currently available unstructured data management tools are two major barriers to the successful management of unstructured data.

Many industry pundits claim that 80 to 85 percent of all enterprise data resides in an unstructured format. Some studies counter that statistic, arguing the true percentage is significantly less. DATAVERSITY's own 2012 survey reflects a percentage below the commonly stated 80 percent. No matter the actual percentage compared to structured formats, there is little doubt the amount of unstructured data continues to grow.

Gartner Research, one organization quoting the 80 percent metric for unstructured data, predicts a nearly 800 percent growth in the amount of enterprise data over the next 5 years. The large majority of this data is expected to be unstructured. This data growth is one factor driving corporate investments in Big Data, Cloud Computing, and NoSQL.

THE GROWING INDUSTRY AROUND UNSTRUCTURED DATA

A large industry has grown around the task of deriving valuable business information out of unstructured data. With parsers scraping information out of pages and pages of text, in addition to full systems built around data taxonomy and discovery, many options exist for any enterprise trying to make sense of their unstructured information.

Commercially available Content Management Systems (CMS) use metadata to provide more accurate searching of an organization's document library. In fact, metadata remains a very important tool in the handling of any unstructured data. Business Intelligence system providers also offer mature solutions around making sense of the vast arrays of an enterprise's seemingly unrelated information.

Despite any associated costs, enterprises better able to leverage meaningful Business Intelligence from unstructured data gain a competitive advantage over those companies who cannot.

Once that data is successfully under management, finding the best techniques and applications for leveraging the data remains key in deriving a competitive advantage for any organization.

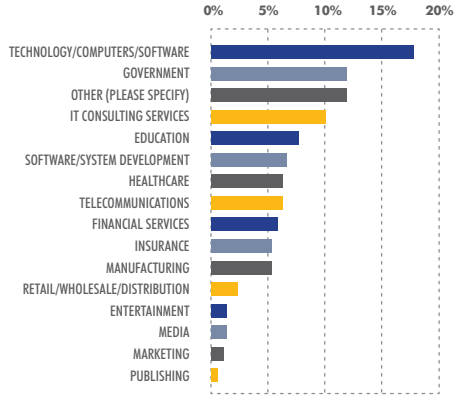
CURRENT USAGE OF UNSTRUCTURED DATA IN THE ENTERPRISE

DATAVERSITY recently sent out a survey to its readership base covering a host of topics related to unstructured data in the enterprise. With nearly 400 respondents, the answers provide a reliable cross section of industry types and organizational sizes.

The survey results provide a unique insight as to how organizations perceive their unstructured data problem, what steps they are taking to derive meaningful business information from that data, what formats the data resides in, and finally some demographic information about their job function, organization, and industry.

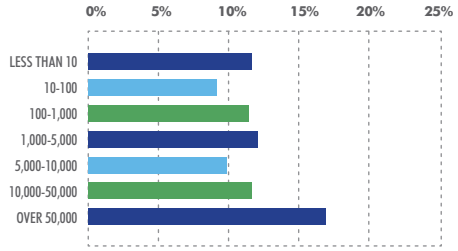
Referencing the organizational demographics of the survey respondents provides a sense of the types and sizes of the companies attempting to manage unstructured data. In short, a wide cross section of organizational sizes and industries are currently managing or considering implementing projects to manage unstructured data. See Graphs One and Two below:

WHAT INDUSTRY ARE YOU IN?



Graph #1

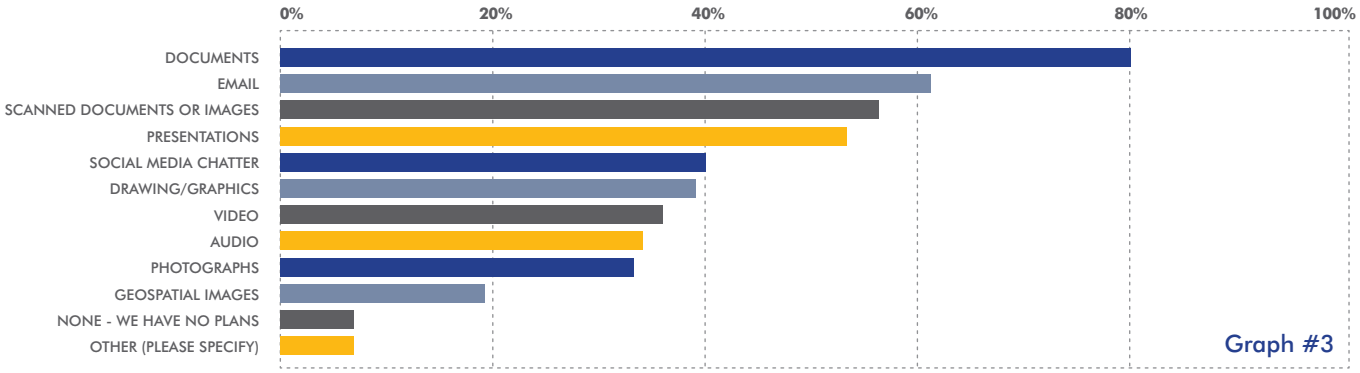
NUMBER OF EMPLOYEES IN YOUR COMPANY?



Graph #2

Documents are the unstructured data type most commonly under management or considered for management, followed closely by emails, presentations, and scanned documents. Various media formats (images, audio, and video) and social media chatter are also important. See Graph Three below:

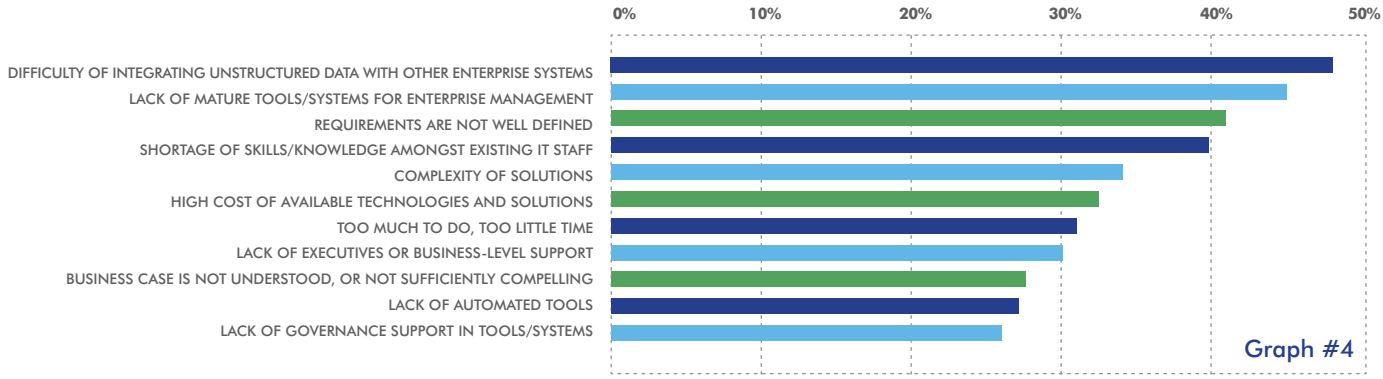
WHAT TYPES OF UNSTRUCTURED DATA DOES YOUR ORGANIZATION CURRENTLY MANAGE, OR IS CONSIDERING FOR MANAGEMENT?



Most organizations hope to improve business efficiencies and reduce costs at their organization through the successful management of unstructured data. Additional potential benefits from unstructured data management include the improvement of communication, deriving sales leads, meeting compliance requirements, improving data governance, improving enterprise search, parsing and analyzing social media channels, and Business Intelligence system integration.

The main barriers to improving unstructured data at the enterprise reflect the growing maturity of the practice. Two major barriers to the successful management of unstructured data are difficulty in integration with enterprise systems along with the immaturity of the currently available unstructured data management tools. Cost, lack of executive support, and a lack of unstructured data knowledge among IT staff are also major barriers. See Graph Four on the following page.

WHAT ARE THE MAIN BARRIERS TO EFFECTIVELY IMPROVING THE MANAGEMENT OF UNSTRUCTURED DATA WITHIN YOUR COMPANY?



System integration with existing structured data appears to be the key technology challenge of the unstructured data management process across most industries (see Graphic Five below). Determining project requirements is another key challenge, especially in the insurance and entertainment industries, as well as improving semantic consistency between multiple systems. Other relevant challenges include security, taxonomy development, scalability, and search result integration across multiple platforms.

WHAT ARE THE KEY TECHNOLOGY CHALLENGES YOU HAVE ENCOUNTERED WHEN ATTEMPTING TO BUILD SYSTEMS OR APPLICATIONS FOR MANAGING UNSTRUCTURED DATA?

Industry	Government	Healthcare	Technology / Computers / Software	Telecommunications	Education	Insurance	Retail / Wholesale / Distribution	Manufacturing	Financial Services	Marketing	Entertainment	Media	Publishing	IT Consulting Services	Software / System Development	Industry Other
Determining Project Requirements	55.9%	47.1%	32.1%	26.7%	33.3%	55.6%	16.7%	33.3%	37.5%	25.0%	50.0%	75.0%	100.0%	35.7%	25.0%	47.4%
Methods for integrating with existing structured data	58.8%	64.7%	56.6%	26.7%	41.7%	50.0%	33.3%	73.3%	62.5%	100.0%	100.0%	75.0%	50.0%	39.3%	20.0%	47.7%
Natural language understanding	11.8%	29.4%	17.0%	13.3%	20.8%	27.8%	33.3%	13.3%	25.0%	0.0%	25.0%	25.0%	50.0%	32.1%	40.0%	18.4%
Entity and/or concept extraction and analysis	29.4%	23.5%	35.8%	13.3%	20.8%	22.2%	33.3%	0.0%	31.3%	0.0%	50.0%	25.0%	50.0%	25.0%	30.0%	26.3%
Filtering / summarizing relevant information	23.5%	23.5%	35.8%	40.0%	33.3%	27.8%	33.3%	60.0%	25.0%	50.0%	25.0%	0.0%	50.0%	21.4%	40.0%	34.2%
Semantic consistency across multiple systems	47.1%	23.5%	41.5%	33.3%	37.5%	44.4%	0.0%	26.7%	37.5%	0.0%	75.0%	75.0%	50.0%	42.9%	40.0%	36.8%
High storage requirements	17.6%	17.6%	28.3%	26.7%	25.0%	33.3%	66.7%	26.7%	25.0%	25.0%	25.0%	25.0%	50.0%	10.7%	35.0%	28.9%
Poor system performance at large scale	11.8%	23.5%	26.4%	20.0%	16.7%	22.2%	16.7%	20.0%	25.0%	0.0%	25.0%	0.0%	100.0%	21.4%	20.0%	15.8%
System modeling and design	17.6%	29.4%	18.9%	13.3%	20.8%	33.3%	16.7%	40.0%	37.5%	25.0%	75.0%	25.0%	50.0%	14.3%	30.0%	21.1%
Digital Rights management	20.6%	5.9%	9.4%	6.7%	12.5%	5.6%	0.0%	6.7%	0.0%	0.0%	25.0%	0.0%	50.0%	7.1%	5.0%	5.3%
Taxonomy development	35.3%	23.5%	24.5%	0.0%	25.0%	11.1%	0.0%	40.0%	37.5%	0.0%	50.0%	50.0%	50.0%	25.0%	25.0%	23.7%
Integrating search results across multiple platforms	23.5%	23.5%	26.4%	20.0%	16.7%	27.8%	0.0%	26.7%	56.3%	25.0%	75.0%	0.0%	50.0%	21.4%	25.0%	47.4%
Security	32.4%	35.3%	17.0%	53.3%	37.5%	16.7%	16.7%	26.7%	31.3%	0.0%	25.0%	50.0%	50.0%	28.6%	15.0%	28.9%
Information Quality problems	32.4%	17.6%	18.9%	40.0%	20.8%	44.4%	0.0%	20.0%	31.3%	25.0%	0.0%	25.0%	50.0%	25.0%	45.0%	42.1%

UNSTRUCTURED DATA FORMATS

TEXT DOCUMENTS

Various permutations of text (word processing files, simple text files, emails etc.) make up the largest amount of unstructured data currently in the enterprise. Many firms are in the process of implementing unstructured data management projects to find useful information from the immensity of corporate email.

Content Management Systems exist partially to help an enterprise manage and derive information from the data contained in unstructured text documents. Most of these systems leverage metadata to provide an extra layer of classification allowing for easier searches and enhanced reporting.

WEB PAGES

Web pages are unique in the world of unstructured data. In fact, an argument can be made that HTML markup itself provides a measure of structure. Web sites that are primarily data-driven might even use a fully normalized database as their back end. In addition to pure HTML markup, many of these Web-based applications are written in Java, PHP, .NET and other development frameworks that render HTML output.

The proprietary algorithms utilized by search engine providers use HTML meta tags in addition to in-text keyword analysis to help tailor search results for the user. Of course, some of that same logic also works in generating Internet advertising for Web surfers, with the ad revenue leading to financial success for the advertising providers.

The rapid growth of social media and the interactive Web is also creating large amounts of unstructured data as well as opportunities for those companies able to manage and derive value from that data. The rising popularity of graph databases optimized for finding relationships between social media users and their consumer preferences (along with others in their social networks) is arguably due to companies hoping to leverage revenue from unstructured data analysis.

MEDIA FORMATS (Audio, Video, Images)

Audio, video, and image files are all forms of unstructured data. Intelligent real-time analysis of audio data is commonplace in digital audio recording and processing, and also used to a lesser extent with the other two formats.

Metadata is a must with media files; it provides necessary additional classification. A trip around the iTunes music store provides an excellent opportunity to see this metadata in action, as band names, genres, and related artists drive Apple's Genius and other music recommendation services like the Pandora Internet radio station.

OFFICE SOFTWARE DATA FORMATS (PowerPoint, MS Project)

Files created using Microsoft Office or any other office suite run the gamut of data formats. Microsoft Access creates and manages fully structured database files in its own format. Excel and PowerPoint files both provide challenges to organizations looking to include information from these formats in their corporate reporting mechanisms. Microsoft's VBA language provides a measure of functionality in parsing meaningful information from Office files.

Commercial applications with proprietary data formats include various customer management (CRM) tools, larger enterprise resource planning (ERP) applications like SAP, or even architectural drafting applications like AutoCAD. In some cases, this software leverages a semi-structured data format such as XML for data exchange between applications.

CAPTURING AND MANAGING UNSTRUCTURED DATA

Before any enterprise can derive meaningful information from the mass of unstructured data, the process to capture that data needs consideration. Data scraping or text parsing involves the extraction of information from data at its most basic level, but other methods and tools provide a more measured approach.

DATA SCRAPING AND TEXT PARSING

Data Scraping is a technique where human-readable information is extracted from a computer system by another program. It is important to distinguish the "human readable" aspect of data scraping from the typical data exchange between computers which can involve structured formats.

The technique first came into vogue with screen scraping, used during the advent of client-server computing, as many organizations struggled with the volume of data residing in legacy mainframe applications. The data from these mainframe apps, parsed from terminal screens, was usually imported into some form of reporting or Business Intelligence application. Data Scraping can also be necessary when attempting to interface to any legacy system without an application programming interface (API).

In recent times, Web scraping has followed a similar technique as screen scraping, allowing meaningful data from Web pages to be parsed, scrubbed, and stored in a relational database. There are currently many applications using some form of Web scraping, especially in the areas of Web mashup and Web site integration, although in many cases APIs provide a more structured process.

Report Mining is similar to data and Web scraping in that it involves deriving meaningful information from a collection of static, human-readable reports. This technique can also be useful in an enterprise's software development QA process, such as facilitating regression test result analysis.

DATA SCRAPING AND TEXT PARSING CASE STUDY

A FLEDGLING FINANCIAL SERVICES COMPANY LEVERAGES TECHNOLOGY; PLAYS WITH THE BIG BOYS

In the mid-1990s, equipped with investment dollars, ARM Financial Group went on a buying binge, acquiring a collection of small insurance companies that specialized in retirement products, such as annuities and structured settlements. Its short-term quest focused on rapidly increasing the dollar value of its assets under management and then profiting from improved efficiencies around the administration of these assets.

Getting accounting systems under control is normally the first step when acquiring any company in the financial sector. This is usually followed by bringing the policy administration systems online with the new company. ARM faced a dilemma concerning the older mainframe systems from a newly acquired firm.

The company made the choice to go client-server for their annuity processing systems; they were one of the first firms in the insurance sector to do so. While massive projects grew around converting policy records from the old mainframe system to the new client-server based system, a more elegant solution was quickly developed to handle the customer service role.

Screen scraping was used to grab policy data from mainframe terminals and store the data in a simple relational database, providing customer service agents with a means to access policy data when on the phone with policyholders. This screen scraping project was online months before all the policy data were converted and stored into the new client-server system's database.

While the policy data in the older mainframe system was in a structured format, the use of screen scraping allowed the rapid development of a needed solution for the customer service function at the new company.

As ARM continued to grow, improving operational efficiency became vital in maximizing profits. Despite the company's state-of-the-art, Web-based system for new annuities, nearly all business was in the form of paper -- both for new policies and changes to existing policies.

Implementing an imaging and workflow system was crucial in re-engineering processes for controlling all aspects of an annuity's policy lifecycle. With most imaging systems, automated text parsing is vital in grabbing information from a document and storing it in some form of database. The workflow elements of the new imaging system at the financial services company were also highly dependent on the manual scraping of important metadata from essentially unstructured paper policy documents.

This metadata was essential when routing a policy through the processing workflows at the company. The imaging and workflow system also used a SQL Server database to persist important data captured from the policy. While some of the policy data was the same as what was stored in the annuity administration system, having it also stored in the workflow system with a SQL Server back end allowed the rapid development of applications in Visual Basic to support customer service and policy management functions.

These new systems proved to be an early form of Business Intelligence application made possible by the leveraging of both manual and automated screen scraping and text parsing to derive useful information from a mass of unstructured paper documents.

In the late 1990s the concepts around what is now called unstructured data were still in their infancy. The technological successes of a small financial services company in pioneering insurance processing allowed it to compete with much larger firms. These successes had a lot to do with recognizing what concepts like screen scraping and text parsing could do to make annuity processing more efficient by allowing the rapid development of new systems to support both customer service and management roles.

METADATA

Sometimes defined as "data about data," metadata is often a useful tool when dealing with unstructured data residing in text documents. In some cases, metadata serves a similar role as a data dictionary by describing the content of data, while other times it is used in more of a markup or tagging purpose.

While some metadata is embedded directly in the document it describes, metadata can also be stored in a database or some other repository. Many enterprises build a formal metadata registry with limited write access. This registry can be shared internally or externally through a Web service interface.

Many metadata standards specific to various industries have developed over time. One example of this is the Dublin Core Metadata Initiative (DCMI), used extensively in library sciences and other disciplines for the purposes of online resource discovery.

Metadata schema syntax can be in a variety of text or markup formats, including HTML, XML, or even plain text. Both ANSI and ISO are active in developing and enforcing standards for expressing metadata syntax.

A type of metadata, meta tags are used in HTML to mark a Web page with words describing the content of the page. In the past, search engines relied mostly on meta tags when building result sets based on a search query, although their relevance has waned in recent times.

LEVERAGING METADATA TO IMPROVE UNSTRUCTURED DOCUMENT SEARCHING

The Education Department of a large Midwestern state faced a problem: the state's teachers needed a system to assist them in organizing appropriate instructional materials for the classroom. The materials primarily resided as documents in the Education Department's Documentum system. The system had a small amount of metadata in it, but there was no functional tool that allowed the easy searching and collecting of that content.

The acquisition of a Google Search Appliance device promised to provide some measure of search functionality, but Google's interface is predicated on one single text box allowing for full-text search without the ability to narrow or filter those search results. The teachers needed the ability to search by criteria such as grade level, subject area, or the type of content (lesson plans, content standards, etc.).

Developing a solution for the state relied on a multi-tier approach that first built a more robust Web search interface that added categorized collections of check boxes for the narrowing of search results, in addition to Google's standard text box for full-text searching.

A school backpack metaphor was used as the "shopping cart" for this Web-based application. As a teacher searched through the instructional materials, anything they wanted to use in the classroom was simply saved in their backpack for later retrieval. Different backpacks could be saved for each teacher depending on their needs.

This front-end search interface and persistent backpack model would not function without improving the metadata stored on each document in Documentum. A side project involved retrofitting each instructional material document with the relevant metadata for grade level, subject, and content type.

Luckily, Google's Search Appliance API allowed additional data to be passed into the search request. Testing revealed the combination of full-text search with additional filtering provided by the Web interface and enhanced metadata returned the relevant instructional materials in the search result set.

While Microsoft's IIS and Active Server Pages were the preferred Web development solution at the Education Department, they were a few years behind in fully implementing the .NET framework. Because of that, the project became an interesting hybrid of Classic ASP at the front end with .NET code used to provide the backpack storage functionality, along with the Web service interface for the Google Search Appliance.

Considering the additional search criteria, it was determined the standard Web page output of the Google Search Appliance was not sufficient to serve the needs of the application. The appliance's search result set was available in XML format, so back-end code was written to enhance the output with graphics, a detailing of the search criteria, and a link to add the specific instructional material to the teacher's backpack. This additional output combined nicely with the standard document summary normally provided by Google's search.

Even though a variety of pieces went into creating the best solution for teachers, leveraging improved metadata supplied the project's ultimate success. Straight out of the box, the Google Search Appliance did not provide the necessary search result filtering needed to return the correct instructional material from a mass of unstructured documents. By enhancing the metadata stored in Documentum, Google's search functionality greatly improved, and the rest of the project was able to proceed.

TAXONOMY

While the term taxonomy is traditionally related to the world of biological species classification, it also plays a similar role in classifying terms for any number of subjects. Information Management taxonomies play a vital role in making sense out of unstructured data, primarily as a method for organizing metadata.

Taxonomies in IT usually take one of two forms. The first form draws from the species classification origin of the term taxonomy, following a hierarchical tree structure model. Individual terms in this kind of taxonomy have "parents" at the higher levels and "children" at corresponding lower levels.

The second taxonomy form is essentially a controlled vocabulary of the terms surrounding any subject matter or system. This might take the form of a simple glossary or thesaurus, or something more complex and resource intensive, like the creation of a fully-formed ontology. This second type of taxonomy tends to be more common in the world of information technology.

The ANSI/NISO Z39.19 standard exists for the authoring of taxonomies, information thesauruses, and other organized data dictionaries, illustrating the growing maturity of this data management discipline. Over the last decade, new companies and software packages centered on taxonomy management have come and gone, with a few earning acclaim as industry leaders, sometimes with taxonomies dealing with a specific subject matter.

Ultimately, taxonomies, controlled vocabularies, and their similar brethren serve an enterprise by organizing metadata in a fashion that helps in finding valuable business information out of unstructured data.

TAXONOMY CASE STUDY #1

IMPROVING CORPORATE INTRANET SEARCH THROUGH THE DEVELOPMENT AND DEPLOYMENT OF TAXONOMIES

In the middle of the last decade, a large publically-held electronics company had a problem with unstructured information, estimated to be about 85 percent of all corporate data. In addition, over 90 percent of the corporate data contained no tagging. Issues with the duplication of content and determining the true age of documents also hampered the company's associates when searching for information.

To get a clearer picture of the scope of the problem, this company surveyed its employees on their Intranet search habits. The responses demonstrated that the employees wanted a better search interface with more categorization and sorting options, along with a more streamlined search result set. Some respondents felt frustrated with the difficulty in finding corporate documents through the current search interface.

A corporate project team was formed with the responsibility of improving Intranet search. The core of this project was the development of various taxonomies and controlled vocabularies combined with metadata to improve the categorization of the internal unstructured information; it was meant to provide benefits for the search interface and in the search results.

Five taxonomies were developed and deployed in the first year of the project. Some were purchased externally and modified to fit the data at this corporation, while the others were fully developed from within. In all cases, certain employees were tasked as Subject Matter Experts in their relevant areas for the purposes of creating the most suitable vocabularies and metadata for the taxonomies.

The second year of the project saw the deployment of three additional taxonomies covering the areas of Human Resource, Six Sigma, and Legal. Two taxonomies were purchased externally and modified to suit, while the other was developed internally. Additionally, some improvement of the original categorization was done based on feedback after the previous year's deployments.

The benefits of the taxonomy development were obvious; they significantly improved all aspects of Intranet search: general employees and electronic engineers wasted less time weeding through bad search results, positively improving productivity, and the company's bottom line. With the reduction of duplicated data, storage costs were improved, even when considering new data growth.

Post project surveys and metrics revealed increased use of the improved search and the use of the categories. The general employee opinions on search improved. Ultimately, the internal development team won an award for their work on the project.

TAXONOMY CASE STUDY #2

FOLKSONOMIES: A HYBRID APPROACH TO TAXONOMY DEVELOPMENT

The creation of taxonomies in any organization comes with associated costs, especially in the case of formal taxonomies where external consultants and Subject Matter Experts are involved with the process. In some cases, informal taxonomies exhibit smaller costs, but with more risk considering the relative lack of taxonomy creation experience compared with a more formalized process.

A hybrid approach utilizes experts in taxonomy creation combined with a user-centric focus on the knowledge modeling for the project. It attempts to lessen the costs associated with a formal taxonomy, while still providing an organized development process. The term "folksonomy" is used to describe this more user-centric approach.

Folksonomies leverage crowd-sourced wisdom on the relevant subject matter at hand, an approach not too different from social bookmarking Web sites like Digg or Reddit, or even a blog community focused around a specific subject.

In these kinds of hybrid taxonomy projects, users are able to submit Web sites and/or tags they would like to see included in the overall vocabulary. An expert taxonomy team reviews these submissions for appropriateness and uniqueness. Finally, the submissions end up persisted in XML format for inclusion in the enterprise search process.

The sharing of these internal social bookmarks and tags enhances the quality of enterprise search and also provides insight to the perception of an organization's internally facing content. Users feel they are an important stakeholder in the process, thus improving company morale.

Implementing a folksonomy normally involves installing some form of tagging tool, usually available as a module for an open-source CMS like Drupal or WordPress. A quality reporting system is also important in determining the efficacy of the tagging, in addition to providing metrics on how the internal content is being used.

In many cases, a hybrid approach to taxonomy development hits a proverbial sweet spot in combining ROI with lower costs when compared to a full-fledged formal taxonomy. For smaller organizations, with a limited budget, it is an approach worthy of consideration.

eDISCOVERY (Electronic Discovery)

eDiscovery providers bring an automated search focus to leveraging valuable information from an organization's unstructured data. They are similar to discovery systems used primarily in the library science and research industries. A separate section covering library discovery systems follows this current section.

eDiscovery is widely used in the legal industry as a means for finding any evidence or information potentially useful in a case. In fact,

court sanctioned hacking of computer systems is a valid form of eDiscovery. Computer forensics, normally used when trying to find deleted evidence off of a hard drive or other computer storage, is also related to eDiscovery.

Electronic discovery systems are able to search a variety of unstructured data formats, including text, media (images, audio, and video), spreadsheets, email, and even entire Web sites. The best eDiscovery applications search everything from in-house server farms to the full breadth of the Internet in their quest to find valuable evidentiary information. Investigators peruse the discovered information in a variety of formats that run from printed paper to computer-based browsing.

When potential litigation is a concern, the protection of corporate emails, instant messaging, and even metadata attached to unstructured documents becomes crucial for any enterprise. This electronically stored information (ESI) became a focal point in changes to Federal Law in 2006 and 2007 that required organizations to retain, protect, and manage this kind of data.

Yet, a 2010 study showed that only that while 52 percent of organizations have an ESI policy, only 38 percent have tested the policy, and 45 percent aren't even aware whether any testing occurred. Considering the risk of future litigation, firms need to focus on the complete development, including testing, of a robust ESI management policy.

As the eDiscovery industry matures, larger companies are bringing the discovery function in house as opposed to relying on external vendor-provided systems; other firms prefer a mix between internal and out-sourced solutions. Whatever their ultimate choice, firms need to realize the vital importance of well-defined and documented procedures for ESI archives and eDiscovery platforms.

DISCOVERY SYSTEMS

The library sciences industry also depends on electronic discovery systems to provide research and other functionality to their consumers. An array of vendors and platforms has grown around this form of discovery, residing generally separate from the eDiscovery sector in the legal industry. Recently, enterprise-based knowledge management initiatives have embraced the traditionally library-based discovery process.

While discovery systems for library sciences have previously focused on search products for traditional content (e.g. books and periodicals), in recent times their scope has broadened to include a wider range of material, including video, audio, and subscriptions to electronic resources. Additionally, content from external providers is now able to be discovered in the search.

These discovery systems depend on the indexing of both document metadata along with the full text to provide a robust set of search results for the user. Content providers benefit from enhanced exposure as well as being able to control the display and delivery of the discovered materials. Obviously, cooperation between those creating the content and those creating the discovery system is paramount to ensure the proper indexing of metadata.

With competing discovery system providers, those also producing content sometimes choose to not index their documents with a competitor. This occurred when EBSCO removed its content from the Ex Libris Primo Central discovery system just before introducing its own EBSCO Discovery Service. Other pundits worry organizations that both produce content and a discovery service will skew any search results to favor their own content. Consumers, including libraries, need to take into account these issues before subscribing to any discovery system.

TEXT ANALYTICS

Text analytics is another related discipline useful in deriving meaningful information from unstructured data. The term became widely used around the year 2000 as a more formalized business-based outgrowth of text mining, a technique in use since the 1980s.

In addition to raw text mining, text analytics uses other more formalized techniques, including natural language processing, to turn unstructured text into data more suitable for Business Intelligence and other analytical uses. Considering that a majority of unstructured

data resides in a textual format, text analytics remains one of the most important techniques for making sense of unstructured data.

Professor Marti Hearst from the University of California, in her 1999 paper Untangling Text Data Mining, presciently described the current practice of text analytics in today's business climate:

"For almost a decade the computational linguistics community has viewed large text collections as a resource to be tapped in order to produce better text analysis algorithms. In this paper, I have attempted to suggest a new emphasis: the use of large online text collections to discover new facts and trends about the world itself. I suggest that to make progress we do not need fully artificial intelligent text analysis; rather, a mixture of computationally-driven and user-guided analysis may open the door to exciting new results."

Text analytics makes up the basis of some of the previously mentioned methods used in capturing unstructured data, most notably discovery systems and eDiscovery. It is also employed by organizations to monitor social media for human resources applications or personally targeted advertising.

Machine learning and semantic processing are two other sub-disciplines of text analytics at the forefront of innovation. IBM's Watson computer, famous for defeating Jeopardy! all-time champion Ken Jennings, is an apt example of the power of semantics at the higher end of computer science.

In addition to natural language and semantic processing, a typical text analytics application might include other techniques or processes, such as named entity recognition, which is useful in finding common place names, stock ticker symbols, and abbreviations. Disambiguation methods can be applied to provide context when faced with different entities sharing the same name: Apple, the Beatles record company, compared to Apple, the consumer electronics giant.

Regular expression matching is a straightforward process used in parsing phone numbers along with email and Web addresses from unstructured text. Conversely, sentiment analysis is a more difficult technique, attempting to derive subjective information or human opinion from text. Machine learning combined with sophisticated syntactic analysis techniques normally make up the basis of automated sentiment analysis.

Text analytics remains a wide-ranging disciplineused in both the business and scientific worlds. From leading edge applications like IBM's Watson to day-to-day tasks like parsing emails from text, it is an important part of capturing unstructured data.

INTEGRATING UNSTRUCTURED DATA AND PUTTING IT TO USE IN YOUR REAL WORLD

Christine Connors, Principal, *TriviumRLG LLC*

Various permutations of text (word processing files, simple text files, emails etc.) make up the largest amount of unstructured data currently in the enterprise. Many firms are in the process of implementing unstructured data management projects to find useful information from the immensity of corporate email.

Content Management Systems exist partially to help an enterprise manage and derive information from the data contained in unstructured text documents. Most of these systems leverage metadata to provide an extra layer of classification allowing for easier searches and enhanced reporting.

Now that you are capturing and storing your data from unstructured sources, what can you do with it? Where can you put it to good use? What new categories of applications are best suited to exploit it?

ASSET VALUATION

Your organization has created terabytes of intellectual property - what is its value? Value is a difficult thing to assess. A piece of information stored "just in case" today may or may not become the critical missing piece of the puzzle down the road. That is assuming of course that you can both find and use that information.

Applying structure to your data will enable two critical processes:

1) De-duplication and 'weeding' of bad or unnecessary data
2) Visualizing what remains

Once that weeding has occurred, the costs of data storage can be estimated based upon hardware and maintenance expenses. These costs are then weighed against the value of the digital assets. Value is determined by the alignment of the type and nature of the data against the organization's core goals. Is the data used as inputs to product or to measure the profits of the products or services being delivered? How many degrees of separation are there? Does the data identify opportunities or threats in the marketplace? What are the predicted profits and potential losses?

EXPERTISE LOCATION

Consider all of the resumes your HR department has collected; they are a wealth of unstructured data regarding the talents of your employees. Using text analytics, a profile can be built of each person. The application of structure to such a collection of existing data means that project heads can easily identify potential team members who have the experience needed to tackle a particular challenge. Add to that the HR-approved information collected during performance reviews, and the team leader has a better handle on what strategies to employ managing the team's efforts.

Christine Connors has extensive experience in taxonomy, ontology and metadata design and development. Prior to forming TriviumRLG Ms. Connors was the global director, semantic technology solutions for Dow Jones, responsible for partnering with business champions across Dow Jones to improve digital asset management and delivery. In that position, she managed a worldwide team responsible for the development of taxonomies, ontologies and metadata that are used to add value to Dow Jones news and financial information products. Ms. Connors also served as business champion for the Synaptica® software application, including managing a US- based team of software developers, and supported Dow Jones consulting practices worldwide, which deliver end-to-end information access solutions based on taxonomies, metadata and semantic technologies. Prior to joining Dow Jones Ms. Connors was a knowledge architect at Intuit, where she was responsible for introducing semantic technologies to online content management and search. And before that, she was a Metadata Architect at Raytheon Company and Cybrarian at CEOExpress Company. At Raytheon Company she oversaw knowledge representation and enterprise search, delivering large-scale taxonomies, metadata schema and rules-based classification to improve retrieval of petabytes of internal information via a multi- vendor retrieval platform.

BUSINESS INTELLIGENCE

It's always good to know how exterior forces can impact your organization. Perhaps your best customer has joined the board of a non-profit organization -- a board on which an executive at your top competitor is already a member. How will that new network connection affect your relationship with your customer's organization?
Or perhaps one of your top engineers has been asked by her alma mater to work with a lead professor, his students, and another alumnus on a project that could benefit the school. Could you lose this top performer to a new venture? Is the research worth investing in? Would you like to set up an alert, using a discovery system, to keep track of the internal memos and external press regarding the project?

NEW PRODUCT DEVELOPMENT

Your researchers and editors have crafted fabulous publications. System analysis reveals that your subscribers are only using bits and pieces of this published work, reading a page or two, reading only the abstract, or going right to charts and visualizations. How can you break out these sub-sections of content with minimal overhead? How can you start quickly by using existing content?

By identifying entities in your content, you can re-use or create new graphs, charts, and even user-filtered data visualizations. The entities are identified by analyzing the existing content, extracted or tagged, and then indexed for re-use.

Doing this work allows you to publish sub-sets of data within a single publication or across publications. You can use wholly owned, licensed or a combination of content as contractually permitted. You can integrate your data in multi-media tools or social networking sites.

REQUIREMENTS FOR UNSTRUCTURED DATA PROJECTS

By **Christine Connors**, Principal, *TriviumRLG LLC*

As with any undertaking, requirements are needed for an unstructured data project. It isn't about simply exposing the contents of the documents. It is about making that content useful to the systems and people who need to use them. Or as many experts have said in other applications: making the right content available to the right people at the right time.

There may be documents exposed that you didn't know were there, that shouldn't be publicly available, and are available because of an error somewhere in the applications. Imagine the trouble if your new system indexed an HR spreadsheet with salaries, addresses, and social security numbers, while being backed up onto a shared drive the user thought was secure?

Consider the content collections that will be part of the program. Do you anticipate any of it having restrictions? If so, then what are those restrictions? How will authorized users authenticate and gain access? Will you restrict access by entity type? By rules-based classification? By system access and control policies? These are important things to consider.

Given that you might find documents you weren't expecting, how will you architect the back end to scale effectively? Will it be easily repeated on additional clusters? What OS and software will it need to run? Will it fail over? Can it scale to handle the number of users, documents, and entities predicted for the anticipated life of the hardware?

Once you've determined that, how will users interact? What will the front end need to provide? Typically users manage Create - Read - Update - Delete rights as permissioned within a system. They also search, browse, publish, integrate, migrate, and import to and from

other systems. What tools are needed to support these actions? Should select users be able to perform administrative tasks via a client or browser interface? How about the ability to generate reports? What operating systems does this interface need to support?

Once you've got your content under control, how are you going to package and publish it? What other applications need to use the data? What are the interoperability requirements?

The ongoing identification of your unstructured data is critical to avoid undertaking such a project again. One method is via Metadata Management. What requirements do you have there? What kinds of information remain important to manage, in addition to the metadata elements? Will you need taxonomy? Do you need an external tool or is there a module within your current CMS, DMS or portal solution that will suffice?

There are many questions here, but most of the overall process is not too different than anything led by a competent project manager. These tasks can be completed in parallel or serially in combination with usability tests, surveys, and focus groups. Taxonomy development, if needed, will benefit from the guidance of an expert, as it is not typically a linear process like that of software development.

SUMMING UP THE OPPORTUNITIES CREATED BY UNSTRUCTURED DATA

The existence of vast quantities of unstructured data at any organization is not necessarily a problem. In fact, it needs to be considered as an opportunity for success. The various case studies contained within proved that projects focused on deriving information out of seemingly unrelated data in many cases allowed firms with a proactive attitude to gain a competitive advantage when compared to firms who fear or ignore such unstructured data.

This paper provided background on the various types of unstructured data along with a collection of time-honed techniques for capturing and managing that data. The real world case studies provided inspiration in solving potential unstructured data issues. The list of applications and organizations with tools related to unstructured data are an excellent starting point for researching the wide issues in this sector of data management.

Additionally, the included survey data revealed that most firms are taking an active approach with unstructured data, so no one should feel alone when considering their own data issues. It is a fascinating area of expertise that continues to change and evolve; there are many opportunities for organizational success, personal success, and knowledge growth, with so many transformations occurring all the time.

APPENDIX: OPEN SOURCE AND COMMERCIAL APPLICATIONS AROUND UNSTRUCTURED DATA

TERADATA: Teradata produces enterprise Business Intelligence and Data Warehousing software. Their suite also provides functionality that facilitates data extraction from various unstructured and structured sources into a proprietary relational database. The company recently added text analytics (from Attensity) to its software suite to better analyze certain types of unstructured data, including text documents and spreadsheets.

Teradata Corporation
10000 Innovation Drive
Dayton, OH 45342
Phone: (866) 548-8348

ATTENSITY: Attensity specializes in the extraction of meaning from unstructured data through the use of text analytics. They recently partnered with Data Warehousing provider, Teradata, to add text analysis to the latter's software suite.

Attensity Group
2465 East Bayshore Road
Suite 300
Palo Alto, CA 94303
Phone: (650) 433-1700

DATASTAX: DataStax is known as a leader for commercial implementations of the Apache Cassandra database. Last year's introduction of DataStax Enterprise combines Cassandra with the company's OpsCenter product, all running on the Hadoop framework.

DataStax HQ - SF Bay Area
777 Mariners Island Blvd #510
San Mateo, CA 94404
Phone: (650) 389-6000

TAXONOMY PROVIDERS

SMARTLOGIC: Smartlogic is an enterprise taxonomy provider primarily known for Semaphore, a content intelligence platform promising improved control of and easier access to an organization's unstructured data.

Smartlogic US
560 S. Winchester Blvd, Suite 500
San Jose, California, 95128
Phone: (408) 213-9500
Fax: (408) 572-5601

SYNAPTICA: Synaptica is an innovation leader in the areas of enterprise taxonomy and metadata. Their platform integrates with Microsoft SharePoint, providing a method to store Synaptica's taxonomy within SharePoint, facilitating unstructured document search.

Synaptica, LLC
11384 Pine Valley Drive
Franktown, CO 80116
Phone: (303) 298-1947

TERAGRAM: Recently acquired by SAS, Teragram is an expert in the world of linguistic search. They help organizations to better manage unstructured content in a variety of languages allowing an enterprise to further develop its international presence.

SAS Institute Inc.
100 SAS Campus Drive
Cary, NC 27513-2414
Phone: (919) 677-8000

CONTENT MANAGEMENT SYSTEMS

DOCUMENTUM: Documentum remains one of the largest content management system (CMS) platforms in the industry. The software facilitates the management of business documents as well as a host of other unstructured data types, including images, audio, and video. Documentum is now owned by IT services conglomerate, EMC Corporation.

EMC Corporation
176 South Street
Hopkinton, MA 01748
Phone: (866) 438-3622

MARKLOGIC: MarkLogic makes enterprise software to help organizations manage unstructured data. Their system is based on the use of XQuery for fast retrieval of documents marked up with metadata in XML format, and thus scales nicely when accessing Big Data stores.

MarkLogic Corporation Headquarters
999 Skyway Road, Suite 200
San Carlos, CA 94070
Phone: (877) 992-8885

WORDPRESS: WordPress is one of the most popular open source blogging and content management platforms. A robust community has grown around the platform which leverages the MySQL and PHP open source solutions for database and scripting functionality.

DRUPAL: Drupal is another open source content management platform, but without the self-blogging focus of WordPress.

DISCOVERY SYSTEMS

VERITY K2: K2 is an enterprise search platform, or discovery system, used by organizations wanting to provide intelligent searching of the mass of corporate unstructured data. Verity was acquired by Autonomy, which in turn was recently acquired by HP.

Autonomy US Headquarters
One Market Plaza
Spear Tower, Suite 1900
San Francisco, CA 94105
Phone: (415) 243 9955

SERIALS SOLUTIONS' SUMMON SERVICE: The Summon service from Serial Solutions is a Web-scale discovery system used primarily by libraries. Summon provides search functionality on a full range of media, including audio, video, and e-content, in addition to books.

Serial Solutions North America
501 North 34th Street
Suite 300
Seattle, WA 98103-8645
Phone: (866) SERIALS (737-4257)

EBSCO DISCOVERY SERVICE: EBSCO's Discovery Service facilitates discovery of an institution's resources by combining pre-indexed metadata from both internal and external sources to create a uniquely tailored search solution known for its speed. Although known for their database and e-book services, EBSCO's Discovery Service primarily supports research institutions and libraries.

EBSCO Publishing
10 Estes Street
Ipswich, MA 01938
Phone: (800) 653-2726 (USA & Canada)
Fax: (978) 356-6565

OCLC WORLDCAT LOCAL: WorldCat Local is a library-based discovery system provided by the Online Computer Library Center (OCLC). The system provides single search box access to over 922 million items from library collections worldwide. OCLC is also the organization responsible for first developing the Dublin Core Metadata Initiative.

OCLC Headquarters
6565 Kilgour Place
Dublin, OH 43017-3395
Phone: (614) 764-6000
Toll Free: (800) 848-5878 (USA and Canada only)
Fax: (614) 764-6096

EX LIBRIS PRIMO CENTRAL: The Primo Central Index is the centerpiece of Ex Libris' Primo Discovery and Delivery discovery system focused on providing access for the research scholar audience to hundreds of millions of documents. Ex Libris is a leading provider of automation solutions for the library sciences market.

Ex Libris
1350 E Touhy Avenue, Suite 200 E
Des Plaines, IL 60018
Phone: (847) 296-2200
Fax: (847) 296-5636
Toll Free: (800) 762-6300

METADATA

DUBLIN CORE METADATA INITIATIVE: The Dublin Core Metadata Initiative is a metadata collection primarily used by libraries and education institutions worldwide. It was originally developed by the Online Computer Library Center (OCLC), located in Dublin OH.

ESRI ARCCATALOG: Esri's ArcCatalog is a tool within their ArcGIS software suite used for the development and management of GIS-related metadata. Esri is a worldwide leader in the management of geographic data.

Esri Headquarters
380 New York Street
Redlands, CA 92373-8100
Phone: (909) 793-2853

SCHEMALOGIC METAPOINT: MetaPoint is a metadata tagging and management tool developed by SchemaLogic. Its primary audience is companies with large investments in the Microsoft-based document tools, Office, and SharePoint. MetaPoint promises to provide the missing connection between the two software products. SchemaLogic was recently acquired by taxonomy systems provider, Smartlogic.

Smartlogic US
560 S. Winchester Blvd, Suite 500
San Jose, California, 95128
Phone: (408) 213-9500
Fax: (408) 572-5601

ABOUT DATAVERSITY

We provide a centralized location for training, online webinars, certification, news and more for information technology (IT) professionals, executives and business managers worldwide. Members enjoy access to a deeper archive, leaders within the industry, knowledge base and discounts off many educational resources including webinars and data management conferences. For questions, feedback, ideas on future topics, or for more information please visit: http://www.dataversity.net/, or email: info@dataversity.net.