THE GILBANE REPORT

Content Technology Works!

Vol. 12, No. 4
May 2004

www.gilbane.com

Published by:
Bluebill Advisors, Inc.
763 Massachusetts Ave.
Cambridge, MA 02139
USA
(617) 497.9443
Fax (617) 497.5256
www.bluebilladvisors.com

Editor:
Frank Gilbane
frank@gilbane.com
(617) 497.9443

Editors Emeriti:
Tim Bray
tbray@textuality.com
(604) 708.9592
David Weinberger
self@evident.com
(617) 738.8323

Senior Editors:
Mary Laplante
mary@gilbane.com
(617) 497.9443
Bill Trippe
bill@gilbane.com
(617) 497.9443
Bill Zoellick
billz@gilbane.com
(617) 497.9443
Contributing Analyst:
Mark Walter
mark@gilbane.com

Recent Contributors:
Sebastian Holst
sebastian@gilbane.com
Thad McIlroy
thad@arcadiahouse.com
Ann Rockley
rockley@rockley.com
Kathleen Reidy
kathleenoreidy@yahoo.com
Bob Doyle
bobdoyle@skybuilders.com

Production Assistant:
Sarah G. Dionne
sarah@gilbane.com
(617) 497.9443

Subscriptions:
sales@gilbane.com
(617) 497.9443
Customer Service:
customersevice@gilbane.com

Consulting Inquiries:
consulting@gilbane.com
(617) 497.9443

Gilbane Report White Papers:
Mary Laplante
mary@gilbane.com
(617) 497.9443

Events:
conferences@gilbane.com
(617) 497.9443

UNDERSTANDING TAXONOMIES & SEARCH FOR CORPORATE APPLICATIONS

We have written about search technology for enterprise applications a number of times, but the interest in this space continues to accelerate, and the sophistication of the inquiries keeps on increasing. As companies grapple with ever-expanding amounts of (especially unstructured and semi-structured) content and the resulting difficulty of even finding information they know they have *somewhere*, they become more willing to consider the effort of organizing information so that it can be found, and found quickly.

This means that IT strategists and many business managers now need to understand what taxonomies are, what their value to search is, how they get developed, what is involved in their design and use, what technology can do *vs.* what humans still have to do, and what they need to consider before they get started.

This month contributor Lynda Moulton joins us to provide an introduction to taxonomies and related concepts. Lynda's article is designed to serve as an introduction suitable for anyone implementing or managing a corporate search, portal or knowledge management application, but her advice, gleaned from years of helping companies better manage their information, will also be valuable to those of you who already understand taxonomies and their value.

CONTENTS

UNDERSTANDING TAXONOMIES & SEARCH FOR CORPORATE APPLICATIONS

The content management software industry has discovered that promoting *taxonomy* delivers significant visibility. It has the desired effect of letting the market know that a vendor is a serious player in the content management market, while also driving prospects to their consulting practices. Taxonomy is one of those words that is so bandied about that everyone is sure to feel the need for one – whatever it is, whatever it does. Like many good ideas, useful business tools, or enabling components, *taxonomy*, when affiliated with a product, is given impossible hype. The projected outcomes of building or deploying taxonomy go far beyond what professionals who build them and professional search experts who employ believe they can contribute.

In this paper, we'll examine some concepts that surround taxonomy, reasons for building them, and most important, methods for deploying taxonomies once you build them. As in any hot new area of business, once you start cribbing terminology from other disciplines, new meanings for terms evolve and present problems for those trying to understand how an old term, with its original definition, fits the new application. In order to sort out some of these confounding ideas, some context for the subject sets out the scope of taxonomy in the enterprise content management field. [*Content Management Strategies: Integrating Search*, Gilbane Report *Vol 11 Num 7*, September 2003]

The concepts we'll explore are set forth for the **enterprise user** with a need to find content (search), the **IT staff** tasked with building or deploying an application that employs search, and the **manager** that observes the need for more efficiency storing and retrieving information. **Vendors in this market** may benefit, as well, by observing the ways in which we position taxonomy and search for these audiences. Confusion and mystery around the basic topic only makes product marketing and successful implementation more difficult. Product suppliers can improve the experience for buyers by conversing at the customer level, addressing the specific needs, and establishing realistic expectations. The concepts are simple but the products and execution are anything but.

CONTEXT FOR TAXONOMY

We are in an era when *taxonomy,* related to technology, refers to a business tool rather than an ordered nomenclature of organisms. The more recent business meaning relates to a list of terms divided into groups, categories or clusters, and is usually paired with *search*. Search, as a function, has play in many products and services. In fact, search is so ubiquitous that when we use it we often aren't aware or we don't give thought to what search technology is behind the experience. The search engine is usually "hidden" from the novice searcher in a way that makes it unnecessary to know how it works.

Indexing drives search

Underlying any search engine is computerized indexing; it is the method of indexing content that results in the richness, or lack, of the search experience. Our individual experiences with indexing dates from the time we learned to use a phone directory or an index at the back of a textbook. These examples of human created indexes illustrate two fundamental challenges still facing computer-indexing algorithms today.

Indexing a list of names to enable finding associated information, addresses and phone numbers, presents the organizer with **sequencing or alphabetizing challenges**. Does upper case belong before lower case? Does *USA* come before *United Parcel Service*? In a recent Verizon phone directory *US Airways* comes before *U File Discount Document Center*, which comes before *US Computer*, which comes before *United States Government*. Even a professional indexer would be hard pressed to articulate the alphabetizing rules in this example. Clearly, there are rules being applied to which many of us don't subscribe. In this example, we can see that norms of sorting are not universal in the business world. This carries over to search because search is partially governed by the sequencing of lists for browsing and the way results are sorted for viewing.

Our second common use of non-automated indexes, at the back of a book, presents another challenge in automated search, namely the **choice of words** that should be used to identify prominent concepts in a book. The most common practice is to use the author's language, but that results in a dilemma when the author varies his or her language to express a single idea. In *Management: Tasks, Responsibilities, Practices* by Peter Drucker, viewed by many as the first significant work to describe knowledge management concepts, the term *Knowledge workers and work* appears in the index with several pages listed, but *Motivation* also appears in the index with a subheading of *Knowledge workers*. Different page numbers appear at each entry. This type of indexing has been assigned the name "keyword indexing" in an automated index. Finding information in a keyword index depends primarily on the skill of the searcher to guess the term or terms the author used. It also burdens the searcher who must look at all locations in the book that are referenced by the index.

Taxonomy supports indexing

One of the solutions to normalizing language in search is to create control lists of terminology and to sort out language rules such as sequencing, cross-references, and usage guidelines. Before continuing the discussion of how taxonomy supports indexing for search by managing what gets into a searchable index, a reader's guide to the terminology in the remainder of this article is in order.

Glossary

Bibliographic data fields —	The elements that make up citations to a in a list of books, articles, maps, document items and other materials.
Categorization —	The process of grouping materials into one or more classes or topical areas.
Classification —	A structured and reasoned system of organizing materials according to their **single** strongest attribute. One attribute may represent a single facet of the material
Content (Information) —	The informational matter in a collection of materials or a single domain.
Controlled Vocabulary —	An authorized and standardized list of terminology used to define the content of one attribute or facet of materials in a collection. A collection may be defined by more than one

	controlled vocabulary. (e.g. Subject, Corporate Affiliation, Publisher).
Cross-reference —	In a controlled vocabulary list, a directive from one term in the list to use another.
Facetted Classification –	Categorization based on multiple aspects of a domain.
Full-text —	The entire content of material being indexed for search and retrieval.
Index (Search) —	A finding device. A set of information that directs the user to the object of the listing. In a term index, each entry points to one or more materials by a virtual or physical location.
Keywords —	Significant term (word or phrase) being searched that may or may not belong to a controlled vocabulary list
Metadata elements —	Structured categories defined to contain information about one aspect or attribute (e.g. publisher, subjects) of an information resource (e.g. book, document, image).
Ontology —	A structural specification for expressing all possible relationships among concepts.
Taxonomy (Content Management) —	A list of terms for classifying one attribute of information resources (e.g. subjects, names). A controlled vocabulary with a graphical structure for visualization of structures.
Thesaurus (Information Science) —	A hierarchically structured controlled vocabulary of terms that are used to describe information resources. A more comprehensive form of taxonomy with deeper relationships and cross-references.
Validation List —	Any controlled vocabulary that a computer algorithm uses to verify acceptability in a database field.

Classification to controlled vocabulary

In search technology, there is renewed interest in a traditional method of categorizing content (e.g. articles, books, papers). This interest is in the use of a *controlled vocabulary* as an indexing language, as opposed to simply keyword indexing. In its simplest form, a controlled vocabulary is a validation list, usually few in number, of terms that can be assigned to an entity. This is not unlike the labels assigned to the shopping aisles of a store. Similarly, libraries use a code, a *classification number*, to indicate wherein a collection a book "belongs." The classification number represents the **strongest** subject content for the book. A classification number may reveal books strongly in that classification but does not identify weaker or alternative subjects in the book. In both the grocery store and library you have a dilemma for both the classifier and the searcher. Do dried fruits belong in the produce section or near other packaged goods such as canned fruits? Does a book on fossil fuels belong with geology (in Science) or with petroleum processing (Technology)? The distinctions are not always clear.

To overcome the limitations of giving an item only one classifying category for ease of shelf browsing, librarians devised a second system of assigning controlled vocabulary to express the "aboutness" of a book. They added the concept of controlled vocabulary to that of classification. In this second system, a book could be categorized by any number of controlled vocabulary terms, subject headings or topics. There are numerous controlled vocabulary lists that have evolved from institutions as diverse as the Library of Congress to the American Chemical Society. The terms in these lists range from the language of the generalist to that of highly specialized researchers with their own languages.

Controlled vocabularies that were developed for library subject categorization or subject indexing became highly evolved over the past century. In particular, they addressed the problem of synonymous concepts or related terminology by adding structure beyond alphabetizing rules. In the 1970s an ANSI standard (Z39.19) was issued that set forth a hierarchical structure for building up controlled term lists much like a taxonomic structure of organisms, called a *thesaurus*. The relationships included Broader Term (BT) with its reciprocal Narrower Term (NT). Unlike a biological taxonomy, thesaurus structure provides for synonymous relationships with Use (U) and its reciprocal Used For (UF) indicating the preferred controlled vocabulary term. Finally, for relationships that are merely associative but can't be termed broader or narrower (e.g. causal as in *vapor trails* and *aircraft*) a Related Term (RT) relationship is included.

Scores of professional associations, scholarly society publishers, and government agencies have developed and applied ANSI standard thesauri when indexing specialized content in their fields. To name a few: NASA, Department of Energy, National Institutes of Health, American Society for Metals (ASM International), American Petroleum Institute (API), each developed a subject specialized thesaurus. Their lists have been used for decades to assign multiple subject categories to the individual pieces of content they publish: journal articles, papers, etc. Societies often publish for searching specialized indexes based primarily on their controlled subject lists, as well as author names, titles, and other finding categories.

The point is, thesauri control indexing and indexing enables search. Sometimes a limited vocabulary, taxonomy, is sufficient in an organized body of content to service indexing and to support a browsable search structure. Once collections approach the size of a major society publisher, a thesaurus of thousands of terms with relationships is needed.

On the horizon, *ontology*, like taxonomy, has its roots in another discipline, *philosophy*. Ontology has been superimposed on a newer and more complex method of relating subjects. In fact, ontology deals with whole concepts composed of terms and relationships among terms that are much more complex than thesaurus hierarchy. Ontologies provide semantic richness that imbues terms with meaning when relationships connect them. Because of the infinite combinations of terms that can be used to form concepts, **ontologies cannot be thought of as controlled vocabularies**. They are **frameworks** for the possibilities of language and term relationships that might be encountered in a specialized domain. The biomedical field has the largest such knowledge representation, the Universal Medical Language System (UMLS) developed by the National Institutes of Health. Commercial, government or private development of ontologies is in very early stages and is only beginning to find its way into experimental search systems. Exploration of this area is worth considering for future search options.

When Search is Deployed Controlled Lists may be Employed

Interactive computer applications require a finding function to locate the records in the database. Search can be as simple as locating a record by its primary key (e.g. an ID, a name, or a record number). It can be as complex as a menu of options to many indexes, one for each field in a structured database. Finally, search may mean that all content or records associated with an application are fully keyword indexed. In a structured database some of the indexed fields may be controlled by term lists that govern what data may be added to the field, while other fields may contain large amounts of "free text" that is then keyword indexed. Ease of access to records in an application depends on how simple or intuitive the search options are; you may not even be aware of whether or how you are using search.

Some examples of search illustrate the difference between a controlled vocabulary search and a free text search.

- Specialized on-line applications have replaced library catalogs since the early 1980s in public, academic, school and corporate libraries. These and publishers databases are the original database applications that made use of controlled vocabularies to categorize and index library resources. A couple of examples of these library databases for specialized collections are at Project Management Institute Library or Miami Dade County Library.

- Databases of publishers of specialized content (e.g. Index Medicus from the National Library of Medicine, Chemical Abstracts from the American Chemical Society) provide structured access to all bibliographic fields (e.g. author, title, subject, publication date), as well as full text searching of the actual content. In these databases the Subject Headings are highly controlled by specialized thesauri using language suited uniquely to each field. Because of high development and maintenance costs, these quality indexes have fees associates with use for "premium search."

- In a bookkeeping application (e.g. Quicken) dropdown lists of Accounts or Categories are validation lists used to uniformly index all payments.

- When you search for a file on your computer you are accessing a keyword index of file and folder names that you create when you label items.

- Call center management for a large technical enterprise will undoubtedly choose an application that indexes customer organization names, customer contact people, date of calls, products used, among other structured fields, plus keywords associated with the call description.

- Google, which we use to search keywords and key phrases from content across the Web, also provides structured search in the form of categorizing. You can confine your search to types of Web sites, search for images-only, or choose *Directory* for category search at http://www.google.com/options/index.html.

- Most e-commerce Web site categorize by types of product and also provide keyword searching so that you can look for product names, product numbers, or product descriptions (e.g. http://www.hp.com/ for finding Hewlett Packard products)

- Specialized industry applications have search operations (e.g. Contractor's Blue Book a contractor's bidding site)

- New search engines are emerging that specialize as content aggregators that you pay to search. Examples would be Factiva (Factiva.com) for business content including news feeds, or KNovel for scientific and technical text books (Knovel.com).

Why employ a controlled vocabulary?

Controlling index content depends on the type of search experience you require, the size of the content collection, the audience, and the complexity of the content. Each of the previous examples has a different audience and purpose. When we use a software application designed for a specialized job function, **we should expect that searching characteristics will anticipate the routine tasks needed to perform the job efficiently**. When software designed to make our jobs easier don't operate as quickly or effectively as when a function is performed manually, we feel frustrated and let down by lack of features. Among the most common complaints about business software are:

- The need to re-type the same information for each transaction completed

- The lack of support for adding consistent entries that would make future retrieval easier

- The inability to find information known to be in the database

These examples highlight places where an approved list of terms can benefit the worker who needs to add information to a content repository or database and the worker who will be required to find information at a later time. It is not sufficient for a database designer to simply provide a field for a particular category of information. If the data to be entered should be limited to one of 20 or even 100 possible terms, the software must provide a list in the form of a validation table. Consistent entries are needed for browsable lists, categorized reporting, and quick searches. The norms we expect people to follow in business practice, standardizing, become much more difficult without controls.

A common validation list is a postal code list for state names. A data entry form can confine the entries to two characters but the list further constrains the possible entries to 50 valid codes. The list will be more useful with translation from state name to code ensure that the correct code is used. In the case of Maine, many assume that the code is MA, which is Massachusetts. A feature that lets the data entry person or searcher type the name of a state and having the software translate it to the correct code is helpful. This constitutes a U relationship (i.e. Maine Use (U) ME)

If address possibilities include countries, provinces, and regions, a tree structure would be helpful to pinpoint qualifying characteristics needed to give a complete address. Where countries have been renamed, cross-references from the old name to the new should be included. This type of enhancement begins to change simple validation lists to something closer to a taxonomic structure.

We are often confronted with business language that is far more complex than geography or products, however. We need to ensure that indexing language is well defined and suitably expanded with highly specialized terminology for classifying documents to instill confidence they will be found when technical searches are performed. Nowhere is the investment in content greater than in research and development to create new and innovative products. Leveraging R & D, not just in the era when it is performed, but throughout the lifetime of an organization. It is best to capture results and insights for future use from experts when they are still actively engaged in research. Indexing with correct and consistent language is an activity that, currently, humans can do best but which automation will increasingly do well and more economically.

The better the controls on language used to categorize research, the better the search experience. A search engine that is built to take advantage of a controlled vocabulary thesaurus with cross-references can insure that a search on *high blood pressure* will also retrieve content that only contains the term *hypertension*, or on *diuretics* will find content containing *hydrochlorothiazide*. This will happen only as long as the taxonomy or thesaurus provides a USE relationship from term one to the other. Knowing that a taxonomy has built-in cross-references to encompass variations on terminology or to bring narrower concepts under the umbrella of a single uniform term strengthens confidence that a search engine will find all relevant content when a search is executed.

While controlled term lists are important to validate data entry, assigning topical categories to describe content in terms that even the author might not have used, they will also be a useful for browsing. By displaying terminology used to consistently index content the interface presents options for the searcher to select the most specific or broadest term that encompasses his knowledge quest.

Deploying taxonomies for metadata maintenance

Librarians describe information content through the assignment of bibliographic descriptors (e.g. Authors, Title, Publication Date) to form a complete bibliographic record called a citation. The citation was presented in the form of a catalog card until thirty + years ago when citations became available electronically. Publishers of bibliographic databases presented online variations of citations, all easily read by researchers.

In the late-1980s a new standard began to evolve for electronic citations, called the Dublin Core. It set forth generic categories for content descriptors called *metadata tags* in which some categories (e.g. *Title*) are the same as in bibliography, and others (e.g. *Creator* instead of *Author*) are more generic to describe the numerous possibilities for types of content. Dublin Core Specification

Metadata categories have been prominent in *content management systems*, similarly to the way bibliographic data elements are employed in *automated library* *catalog systems*. When it comes to search functions, both types of systems have exactly the same purpose, to categorize and index the important elements that describe a specific item of content (e.g. book, document, patent, photograph, news item, company annual report, laboratory notebook).

Data structures of either type of system must provide fields for all elements of metadata or bibliography needed for the content domain. The database must also accommodate one or more tables for the taxonomy or thesaurus that will be used to validate data entry fields. These must be embedded in the application to validate terms as they are added to individual content records. There needs to be support for adding new terms. One mechanism for adding new terms might be a periodic batch load from another source or, at the high end, a mechanism to permit the addition of new terms flagged as provisional entries during the content indexing process by subject experts. A good design will facilitate the reconciliation of provisional terms or modification of existing terms, plus global modification of the records that have used those terms.

A highly developed interface for those who categorize or index specialized content that enables them to interact easily with taxonomy is strongly recommended. This removes one of the serious barriers to indexing large volumes of material that will be useful as searchable resources in the future. The more awkward and cumbersome an indexing process, the less likely content will be routinely added. Some level of human interaction with documents when they are being added to a searchable database is necessary or the quality of indexing will suffer.

Deploying taxonomies in search

Once a substantial body of material has been indexed in a content management or library system there are several methods of search that might be offered. The first and most simple approach features a text box into which the searcher can type a word or phrase. The rules that govern the format of the text differ among systems but usually a help function describes when and how terms can be truncated (e.g. abbreviated to a searchable stem as in *telephon** where all words beginning with this string will be found.). This is a keyword search approach most commonly offered in Web applications. While help files may define what is being searched, the typical user has no idea what the searchable body of information looks like. It may be an ordered list of human assigned terminology or it may be the entire content of all documents associated with the application. It may be language supplied by the author, specialized terms assigned by an expert (that may or may not appear in the text), or it may be that search engine rules are built in to retrieve related content. In this example, a cross-reference may include any content about *phones* in the search results.

A second common type of search is a form containing spaces for typing text that you would expect to find in the bibliographic or metadata fields. At any point in the form, there may be access to taxonomy that controls the field. Being able to type a word or phrase to trigger a scan of the taxonomic term list is a feature that some systems offer.

Finally, systems often provide taxonomy in a form that peals back layers or exposes narrower concepts as you select categories under which you believe your strongest search interest may be indexed. This type of “browsable” structure is available through full public Web-based search engines (e.g. Google and Yahoo) but also in specialized applications that focus on a narrow domain of content as in the HP Web site. This type of application requires that the taxonomy be maintained for currency of language, have sufficient cross-references from popular terminology to controlled terms, and be devoid of terms that have no links to content.

A *browse* structure for searching depends on the use of taxonomy for prestructuring the content by using the taxonomy to categorize or index the content. The taxonomy becomes the organizing structure for content, a visual guide to the knowledge resources. Its success as a search aid depends on the graphic design used to display terms, ease of navigation and suitability of taxonomy language to the searcher. You can see an example of a browsable taxonomy for graphic arts at the Library of Congress site LC - Graphic Thesaurus (type *ink* and uncheck the Content box, then click the *TERM* button) or for a British Maritime site at Maritime. A subset of a Lawrence Livermore thesaurus can be viewed at LLNL.

Search engines that support customized taxonomies require a linkage between the taxonomy term list, which is stored in the database, and content resources to which the terms point. Without embedded linkage, search look-up (i.e. matching the term selected from a browsable list with terms in the content) is very inefficient and will result in poor performance. The best structures have a count posted with terms in the browsable taxonomy; as new content is added to the database with metadata linked to the taxonomy the document count is updated dynamically. If indexing is fully automated with no metadata supplied by human indexers, links between the taxonomy and content need to be automatically updated; a posted count with the terms is still desirable. Posted record counts with taxonomies have the benefit of letting the searcher know exactly how much content is available associated with a term. Then the searcher can decide whether to seek narrower term concepts to select.

It is probably worth noting that Web portals are often divided into sections, each governed by different taxonomic structures. This is an example of faceted classification (e.g. Products, Geographical Regions, Market Segments), each area with its own control list representing different views of how an enterprise organizes its content.

TAXONOMY DEVELOPMENT GUIDELINES

At the end of this section are links to some published material on developing taxonomies or thesauri. Briefly, there are five components outlined here with the fundamentals for a committed development process.

- **Existing resources** are always the best starting point. This includes published glossaries, taxonomies, thesauri and internally generated term lists. If the organization has been indexing materials manually or even in a simple database for a period of time, using keywords to categorize the materials, those keywords can form a useful basis for “fleshing out” a published term list. Internal terms are likely to be far more specific and expertise friendly for the domain than published lists. Data mining a large corpus of internal content, or training a categorizer with a few targeted and clearly written topical reports are also possibilities for building up relevant terms.

- **The scope** of the list you will build depends on both depth and breadth of content to be managed. A small repository of a few thousand documents does not require a long term list. In most cases a few hundred terms will suffice but a very diverse range of subject areas may demand more. It is easy to add terms when a topic becomes over-used by the indexing process but you then need to revisit content to update the metadata with more precise terms if they are added. A corpus of hundreds of thousands of documents in a specialized field may have a thesaurus of two or three thousand terms; however, be cautious about embracing an entire thesaurus from a published source that is many times larger. The language will be too general and there will likely be many sections you would never use. Choose what your enterprise really needs and leave the rest out.

- **Subject matter experts** are needed to validate the terms you will include and the relationships you will build among terms. They will help you decide which, among a group of synonymous terms, are the better choices for your audience. If you lack sufficient expertise in an area of content, utilize commercial categorization software for several passes to extract terminology from your content to build up a candidate term list. Once you see synonyms appearing, use a term count function to determine the most heavily used terms. Make cross-references from the little used terms to USE the popular term.

- **Tools** are software applications that come with your content management software or library software to build up term lists, assign relationships (Broader and Narrower), make cross-references and capture notes about term usage. There are standalone software applications available that let you build term structures fairly easily. They may expedite the initial build but can be a burden with on-going maintenance. External taxonomies need to be batch loaded periodically and the content re-indexed to take advantage of new terms. The key is to have tools that are highly integrated for real-time updating and maintenance with a minimum of human intervention. Library systems have been functioning in this real-time mode for decades; content management systems lag in comparison.

- **Method** is how you use the software tools you have, resources, subject experts and content repositories to build taxonomy most efficiently. The common aspects that all such projects share are that the process is highly iterative and it requires a high degree of human intelligence and focus. Consider short-term goals carefully and build small sections that can be tested early in the process. Employing a variety of methods and software tools for small lists will gradually reveal the most efficient methods for continuing to build up your vocabulary. Working in teams with frequent "sanity checks" of each other's work is also useful.

Grimes, Seth. *The Word on Text Mining; Text analytics provide concept discovery, automated classification, and innovative displays for volumes of unstructured documents*, Intelligent Enterprise 12/10/2003.

Knox, Rita E. *What Taxonomies Do for the Enterprise*. 2p. Gartner Group 09/10/2003

Moulton, Lynda W. *Why do You Need a Taxonomy Anyway? And How to Get Started*, LWM Technology Services 06/01/2003.

WHERE IS TAXONOMY HEADED? CONCLUSIONS & RECOMMENDATIONS

Serious research demands discipline and structure, regardless of whether the work is performed in a laboratory or by searching in books or databases of articles. Automation has turned the search for information and discovery on its head through the shear volume of content that is being produced and **replicated** daily. The same information is disseminated in hundreds of formats and stored in repositories of institutions and on publicly available servers throughout the world. One finds the same full text of published documents available globally through the Web in numerous formats and repositories. Structured databases are searchable through portals that disguise content structures and source, often making the context difficult to discern.

The masses of redundant information available in numerous unstructured or loosely defined formats beg for new order in the information world of research. Seekers are desperate for systematic approaches to information gathering that are at the same time comprehensive and non-redundant. They demand search interfaces that can search all possible repositories (file systems and databases) concurrently and return results in a relevancy-sorted list. This is the claim of *federated search.*

The new order and federated search are several years away from popular usability. Current federated search reflect search engine characteristics of the repositories where the data resides. Besides economic barriers to the commercialization of new federated search technology, there exist numerous economic barriers to building the components needed to have them work effectively. One key to federated searching in a subject specialized knowledge domain is the existence of ontological frameworks for the subject discipline. While a single ontology exists in medicine and health sciences (UMLS), which was government funded, other disciplines are years away from having ontologies fully developed. There are excellent thesauri in a number of defense areas, energy, metallurgy, chemistry, and some softer sciences. But activity to build ontologies for these areas and to create applications that will use them to provide federated and semantic (natural language) query searching is just beginning.

For now, the effort to build subject specific taxonomies and develop them into thesauri is significant and requires sufficient human effort. An enterprise should view this effort as a good first step toward future searching goals. Major efforts in building taxonomies are now and will remain primarily within the content product industry; these continue to be enhanced by enterprise efforts to refine the lists for their own internal use with search on enterprise content. When federated and semantic searching technology gives us appropriate commercial options, ontology availability will be a significant part of the offerings. It is coming.

Lynda Moulton, lmoulton@lwmtechnology.com

INDUSTRY NEWS

Current news, old news (to January 1999), and commentary is available at www.gilbane.com. Free RSS 2.0 news feeds are available at www.gilbane.com/syndication.html.

WEBORGANIC ANNOUNCES PAGESEEDER 3.0

5/28/2004

Weborganic Systems announced Version 3.0 of PageSeeder. With this upgrade users are now able to create and manage document annotations across the full range of Adobe Acrobat clients, including older versions and the Acrobat Reader client. Other features include: an interface allowing non-technical users to create private and secure review groups featuring mail list capabilities; messaging support to allow reviewers to interact using native Acrobat commenting tools, web forms or standard email clients; the ability to lodge documents for review either by uploading or emailing attachment files; support for HTML, PDF and standard Office formats and international characters (Unicode); and availability either with a web-interface or as a middleware server and API for integration into existing document management environments. PageSeeder Version 3.0 is available immediately and can be evaluated either online or as a 30-day trial download. www.weborganic.com

HOT BANANA & NI SOLUTIONS PARTNER

5/27/2004

Hot Banana Software Inc. and NI Solutions announced the signing of a partnership agreement to deliver Hot Banana's Web Content Management System to NI Solutions' customers and to allow joint marketing of the solution. www.hotbanana.com, www.nisolutions.ca

ACCENTURE & FILENET ANNOUNCE DIGITAL PEN CONNECTOR SOFTWARE

5/25/2004

Accenture and FileNet Corporation announced new technology that enables insurance field agents and other mobile workers to electronically perform tasks that once required mounds of paperwork. The technology, designed for use with low-cost commercially available digital pens in conjunction with FileNet Enterprise Content Management (ECM) solutions, allows mobile workers to capture and transmit information to a central repository, where it can be put to work to initiate business processes. The Digital Pen Connector software, written by Accenture Technology Labs, captures the pen strokes, translates them into usable data, and delivers the data to the FileNet ECM system. FileNet ECM then integrates the data in a central content repository for sharing and launches business processes according to the requirements of the business. The solution provides a low-cost alternative to equipping workers with costlier pen-based devices, such as tablet computers or personal digital assistants, and should benefit industries that rely heavily on paper-based processes. www.accenture.com, www.filenet.com

XPRIORI RELEASES NEOCORE XMS-DEVELOPER FOR LINUX BETA

5/25/2004

Xpriori announced the availability of a public beta release of Xpriori's NeoCore XMS-Developer for Linux. Available to all who participate in Xpriori's public beta program, the download is fully functional and has no time limit. Only upon deployment of solutions based on NeoCore XMS must a deployment license be obtained. NeoCore XMS is a Unified XML Information Management System/Database that can manage data and documents, as well as a variety of content types such as spreadsheets, media files, and forms in a fully transactional and access controlled environment. Support for free developer versions of Xpriori products is available on the interactive developer forum at www.xpriori.com/developers

INTERWOVEN ANNOUNCES WORKSITE MP 4.0 FOR COLLABORATIVE DOCUMENT MANAGEMENT

5/24/2004

Interwoven Inc. announced Interwoven WorkSite MP 4.0, enterprise-class collaborative document management software for multiple platforms. The new version features improved usability, better business-unit level configurations, and enhanced compliance capabilities. WorkSite MP 4.0 brings collaborative document management to front-office professionals (such as lawyers, consultants, and other domain experts). WorkSite MP is an Internet-based product suite that delivers collaboration and document management solutions capabilities to allow teams, departments, and divisions of large enterprises to collaborate, build, and share information and manage projects. The software is available in a browser-based interface, integrated with front-office applications such as Microsoft Office, Outlook, and Lotus Notes, or as portlets for the most widely used portals. Interwoven WorkSite MP 4.0 will be available in June.
www.interwoven.com

AUTHENTICA EXTENDS MICROSOFT WINDOWS RIGHTS MANAGEMENT SERVICES

5/24/2004

Authentica, Inc. announced planned support for Microsoft Windows Rights Management Services (RMS). Authentica will partner with Microsoft to extend Windows RMS capabilities beyond the user desktop level to let organizations centrally and automatically enforce document and email usage policies within and outside of corporate boundaries. Authentica will extend the Windows RMS platform with mandatory policy enforcement, e-mail content filtering integration, network folder integration, user enrollment and initialization, and heterogeneous messaging support. Authentica protects corporate IP and other sensitive data both during and after delivery with its Active Rights Management technology. Documents and email are continuously protected and secured, even while letting others revise them. Information owners can control who can access, edit, copy/paste, forward, and print documents, spreadsheets and presentations. www.authentica.com

FATWIRE SOFTWARE ANNOUNCES SPARK 6

5/24/2004

FatWire Software announced that it has expanded its integration with the BEA WebLogic Platform 8.1 through the latest release of Spark 6 pCM (portal Content Management) for the BEA WebLogic Portal 8.1 and the introduction of FatWire's BEA Control Bridge for BEA WebLogic Workshop 8.1. Spark 6 pCM supports BEA WebLogic Portal 8.1 standards, CM: SPI and JSR 168. Spark features content management and document management interfaces, and installs with four pre-built delivery portlets for job postings, corporate communications, documents and advertisements. FatWire's BEA Control Bridge provides developers the ability to create applications in BEA WebLogic Workshop 8.1. The Control Bridge allows developers to access all content and assets, regardless of data type, stored in FatWire Content Server repositories. Spark 6 pCM is $25,000 per server, and is available now.www.fatwire.com

CA & Zope to Bring Open Source Content Management to Enterprises

5/24/2004

Computer Associates International, Inc. and Zope Corporation announced plans to provide customers with open source content management solutions that are compatible with relational database technology and meet enterprise demands for performance, data persistence and manageability. These collaboratively developed solutions will combine Zope's open source content management platform with CA's Ingres Enterprise Relational Database (Ingres). Zope implementations are currently supported by the open source, flat-file database, ZODB. Software engineers from CA and Zope will build an open source RDBMS persistence module that takes advantage of APE (Adaptable Persistence Engine). As a result of this new module, Zope implementations will be supportable by Ingres. The new module could also facilitate support of Zope implementations by relational database management systems other than Ingres. CA is prepared to commit significant development resources to Zope technology. CA and Zope Corporation are expected to announce availability of a Zope RDBMS persistence engine by the end of the year. www.ca.com, www.zope.com

Microsoft Previews Insurance Forms Accelerator

5/24/2004

Microsoft Corp. unveiled a preview of the Microsoft Office Solution Accelerator for Insurance Forms. The accelerator is designed to help insurance industry customers by tying standardized ACORD forms with XML Web services and making those forms available to participants of any of the ACORD forms programs through the use of Microsoft Office InfoPath 2003 information gathering program. The accelerator comprises the ACORD forms and XML schemas designed for use with InfoPath 2003, a set of workflow transactions using Microsoft BizTalk Server 2004, and prescriptive architecture and code for Web services, security and digital signatures. The full release of the accelerator will coincide with the release of the Microsoft Office 2003 Editions Service Pack 1 later this year. The service pack will include several enhancements to InfoPath 2003. www.microsoft.com

Document Sciences & Xenos Partner

5/21/2004

Document Sciences Corporation and Xenos Group Inc. announced that they have entered into a cooperative marketing agreement. The partnership will enable organizations to use Xenos Group's GoXML product suite to transform legacy and disparate system data into personalized content that can be delivered by Document Science's xPression content processing software to multiple output channels, including Web, print, and email, while maintaining complete fidelity of the originating systems. GoXML is an integration and transaction processing solution for structured data such as EDI, EDIFACT, AL3, HL7, X12, XML, FIX/SWIFT and other industry standards. www.xenos.com, www.docscience.com

Vignette Announces Series of Organizational Compliance & Governance Offerings

5/19/2004

Vignette Corp. announced a series of technology solutions to help companies meet organizational compliance challenges, including records and document management, process controls and business productivity applications. Building on the records and document management technologies acquired with TOWER Technology Pty Ltd., Vignette now provides a series of solutions that, together or individually, can help organizations meet the audit and reporting requirements of regulations such as the Sarbanes-Oxley Act, the Financial Services Authority and the Securities and Exchange Commission, while providing the data security and accessibility required by regulations and guidelines such as the Health Insurance Portability and Accountability Act of 1996 (HIPAA) and the Americans with Disabilities Act (ADA). Vignette's enterprise capture technology handles paper records, electronic documents, scanned images and e-mail communications, and manages the information through its life cycle, from creation to destruction. Vignette's integrated records and document management product has achieved the Department of Defense 5015.2-STD certification. www.vignette.com

DAY ANNOUNCES GROUP HAS RELEASED JSR 170 FOR REVIEW

5/19/2004

Day Software announced that the industry expert group has placed in public review the specification request for JSR 170. JSR 170, which has been developed according to the Java Community Process (JCP), is designed to improve the interoperability between content repositories and applications, allowing developers to work with a homogenous API for all content repositories. The JSR 170 standard will help companies manage content across the large-scale enterprise. In addition to Day, Apache, IBM, SAP, BEA Systems and Oracle all serve as members of expert group for JSR 170. Other industry participants include, Documentum Inc., Filenet Corp., and Vignette. The standard is now available for public review. The public review closes on July 19, 2004. www.jcp.org/aboutJava/communityprocess/review/jsr170/, www.day.com

STELLENT & PROTIVITI TO COLLABORATE ON COMPLIANCE PLATFORM

5/18/2004

Stellent, Inc. announced it has entered into a strategic agreement with Protiviti Inc., an internal audit and business and technology risk consulting firm, to create a content management-based compliance platform. Under the agreement, Stellent will deliver and market compliance solutions based on the Stellent Universal Content Management system and supported by Protiviti subject matter expertise. These solutions, to be co-marketed by Protiviti, will offer capabilities such as audit trails; document management; project management and reporting; personalized interfaces; automated process documentation collection and process testing; secure, Web-based access by internal and external users; and records management. www.stellent.com, www.protiviti.com

GLOBALSIGHT SIGNS OEM AGREEMENT WITH ALCHEMY SOFTWARE

5/18/2004

GlobalSight Corporation announced that it has entered into agreement with Ireland-based software localization solution provider Alchemy Software Development for integrating Catalyst 5.0 into GlobalSight's Ambassador product. Ambassador is an enterprise software application with localization management tools. The OEM relationship combines Catalyst's translation of software applications with Ambassador's enterprise process automation. Alchemy CATALYST 5.0 allows for the creation of a localization toolkit that contains all of the linguistic and cultural material that needs to be translated. Ambassador's workflow component and enterprise architecture will enable deployment of Catalyst across large organizations, automating the creation and distribution of developer packages containing software files for software localization. The Software Localization Module will automate the creation, leveraging and extraction of files necessary for software localization; generate statistics for all actions; and enable users to track files throughout the lifecycle of the software localization process. www.globalsight.com

GILBANE REPORT TEAM EXPANDS WITH LAPLANTE AS VP CONSULTING SERVICES & ZOELLICK AS SENIOR ANALYST

5/18/2004

Bluebill Advisors, Inc. and the Gilbane Report announced the appointment of two highly respected industry veterans to meet the rapidly growing demand for the company's consulting services. Mary Laplante has joined the company as Vice President, Consulting Services, and will be responsible for the management and operations of the Content Technology Works Program, Gilbane Report webinars, white papers, and consulting projects. Mary is well known throughout the content technology industry from her extensive industry sales, marketing, and management experience, from her work as an analyst, writer, and consultant, and from her role as Founding Executive Director of OASIS. Mary's experience and exceptional management and organizational skills will help grow the Gilbane Report CTW program and consulting activities to the next level. Bill Zoellick has also joined the Gilbane Report as Senior Analyst, and will be devoting his considerable analytical and writing skills to the Gilbane CTW and White Paper programs. Bill is a highly regarded author whose books include "CyberRegs: A Business Guide to Web Property, Privacy, and Patents", and "Web Engagement: Connecting to Customers in e-Business", both from Addison-Wesley. Bill's insightful and penetrating analysis has informed his many roles as a manager, researcher, developer and author in the content and information management industry. www.gilbane.com

SAIC'S TERATEXT SOLUTIONS SIGNS AGREEMENT WITH STELLENT

5/17/2004

TeraText Solutions, a division of Science Applications International Corporation (SAIC), announced that it has entered into a license agreement with Stellent, Inc.'s Content Components Division to embed Stellent Outside In XML Export, HTML Export, and Viewer technology for use in TeraText Database Systems. SAIC's TeraText Database Systems (DBS) is an information management system optimized for storing and manipulating large volumes of text and managing large structured text databases. Stellent's tools expand TeraText's existing capabilities by providing the ability to convert more than 250 proprietary file formats to XML or HTML for indexing and searching in conjunction with the TeraText DBS. TeraText technology was developed at Melbourne-based RMIT University. In July 2001, SAIC entered into an exclusive agreement with RMIT to develop and commercialize TeraText technology in North America and Europe. The TeraText DBS is sold as part of customized, integrated solutions developed and maintained by SAIC systems specialists. Applications include intelligence gathering, technical documentation, legislation management, publishing, and knowledge management. www.saic.com

IDIOM RELEASES WORLDSERVER GLOBAL ELECTRONIC PUBLISHING SOLUTION

5/17/2004

Idiom Technologies, Inc. expanded the WorldServer product family with the introduction of a new, application-specific version of WorldServer designed to meet the needs of global electronic publishing organizations. WorldServer Global Electronic Publishing enables cross-functional teams including authors, reviewers, editors, translators, localization specialists, and publishers to automate large portions of the document lifecycle, and deliver content in any format or language. With version control, an integrated XML repository, an extensible publishing framework, "preview in-context" capabilities, and search & query functions, the new member of the WorldServer product family enables "continuous translation" of complex documents.

The new Idiom solution supports a variety of digital publishing formats including HTML, PDF, CHM (complied help), as well as print. And it is designed to work with authoring tools, such as Adobe FrameMaker. www.idiominc.com

ALTOVA RELEASES SOFTWARE VERSION 2004 RELEASE 4

5/17/2004

Altova Inc. announced the immediate availability of Release 4 of its Version 2004 product line. Among the major enhancements are extended data-mapping functionality in Altova MAPFORCE 2004, a completely redesigned Altova STYLEVISION 2004, and numerous upgrades to XMLSPY 2004. The entire Version 2004 Release 4 (v2004r4) product line is available for immediate download. In addition to the new features in v2004r4, the Altova Support and Maintenance Package (SMP) now includes coverage for major software releases. New bundled packaging for suites was also announced. Altova Enterprise XML Suite includes v2004r4 Enterprise Editions of XMLSPY, MAPFORCE and STYLEVISION at a savings of almost $700 (USD) off regular price of products purchased separately. Altova Professional XML Suite includes v2004r4 Professional Editions of XMLSPY, MAPFORCE and STYLEVISION at a savings of almost $250 (USD) off regular price of products purchased separately. www.altova.com

DRALASOFT & XYTHOS PARTNER

5/13/2004

Dralasoft, Inc. announced that Xythos Software has chosen to partner with them to provide integrated workflow/document management solutions for enterprise use. The solutions will provide a way for organizations to access, distribute, route, process, and review documents and other file-based information according to the users pre-set rules and standards. Xythos has customized Dralasoft Workflow, Dralasofts BPM product, for use with its WebFile family of products. The new solutions will leverage a range of Xythos/Dralasoft applications including Xythos WebFile Server (WFS) and Client Technology; WebFile Document Manager, WebFile Classification Manager, and WebFile Records Manager; WebFile Client and WebFile Scan Client; as well as Dralasoft Workflow Engine, Workflow Manager, and Workflow Studio. www.dralasoft.com

COMPRENDIUM RELEASES CONTENTGATEWAY 2.1

5/13/2004

Comprendium announced the immediate availability of ContentGateWay 2.1. This version of its content integration solution extends product functionality, while making it simpler to use. Building on Comprendium's Content Services Platform concept ContentGateWay 2.1 includes additional application and infrastructure platform support; enhanced security, scalability and availability features; and a number of end-user enhancements. ContentGateWay 2.1 enhances support for essential enterprise functionality, including load balancing, clustering, caching, and replication. ContentGateWay2.1 provides a unified interface to all content repositories (e.g. XML, Java, .net) by supporting any available protocol such as Corba, SOAP, JMS, RMI, and SMI Comprendium's optimized content protocol. With the new Virtual View users no longer need to know where content resides, since all integrated repositories are available. They can also easily build their own content GUIs or integrate them into existing applications through the availability of content access visual content components (e.g. taglibs, ActiveX). ContentGateWay is based on J2EE and includes a SOAP-based interface. www.comprendium.biz

HOT BANANA & SITEPOSITION PARTNER

5/12/2004

Hot Banana Software Inc. and SitePosition announced a partnership to co-market Hot Banana, a Web Content Management System, built from the ground up in harmony with Search Engine Optimization (SEO) best practices. Websites with content that has previously been inaccessible to search engine spiders, due to Macromedia's Flash!, or dynamic or database-driven content, can now admit search engine spiders and crawlers to access all publicly-available web content, no matter how deep. Hot Banana's ability to parse and index all existing and updated web content, information and data enables any website to achieve a balance between organic search engine rankings and a paid-ranking strategy with Google, Yahoo!, and MSN. Hot Banana is built on Macromedia's Cold Fusion MX Server, can leverage the features available to a J2EE application server, and can be deployed on IBM's WebSphere or Macromedia's JRUN. Hot Banana's database engine is Microsoft's SQL Server, and Hot Banana is .NET-, XML-, and Web Services-ready. Web server performance is taken into consideration by caching database queries, creating persistent Web page objects, and producing cached, XML-based navigation structures. www.hotbanana.com, www.siteposition.ca

EXTENSIS ANNOUNCES PORTFOLIO 7.0 FOR WINDOWS

5/12/2004

Extensis Inc. announced immediate availability of its Digital Asset Management (DAM) solution, Portfolio 7 and Portfolio Server 7 for Windows (Macintosh versions will be available in May). Portfolio enables small-to-medium sized organizations to multiple workgroups in global 2000 companies to organize, retrieve, repurpose, and distribute digital files. Portfolios new AutoSync functionality allows users to assign tasks to Portfolio Server, enabling automatic synchronization. NetPublish is an add-on module that automates the process of turning Portfolio catalogs into full featured websites. Other features include One Click CD/DVD archiving, Batch Image Conversion that allows users to convert image files to JPEG or TIFF format, with controls for resolution, size and color mode, and Round-trip embedding of IPTC and XMP metadata into JPEG and TIFF files. Portfolio 7 is available at an estimated street price of $199.95US with a single NetPublish license. Upgrades are available for an estimated street price of $99.95US. Portfolio Server's estimated street price is $3,499.95US. Upgrades to Portfolio 7 Server are available for an estimated street price of $1,999.95. NetPublish Server licenses can be purchased for an estimated street price of $1999.95US. www.extensis.com

DOCUMENTUM LAUNCHES APPLICATION LOGO PROGRAM

5/11/2004

Documentum announced its Application Logo Program that provides Documentum partners with resources and guidance for developing offerings based on the Documentum Enterprise Content Management (ECM) platform and helps customers easily identify Documentum-accredited solutions for extending their use of the platform. The Application Logo Program provides partners with design standards, architectural support and development best practices for developing out-of-the-box applications, integrations and solutions based on the Documentum ECM platform. Partners in the Application Logo Program earn a "Designed for Documentum" designation for a specific offering by conforming to a stringent set of design specifications and other criteria established by Documentum in collaboration with its partners and customers. Currently, 12 offerings that have earned the Designed for Documentum designation are available with additional offerings currently being evaluated. www.documentum.com

TRADOS Language Server Receives Documentum Accreditation

5/11/2004

TRADOS Inc. announced that its TRADOS Language Server for Documentum has received the "Designed for Documentum" logo designation. The Designed for Documentum accreditation can be earned by offerings developed on the Documentum ECM platform, as part of the new Application Logo Program. TRADOS Language Server for Documentum is an integration that adds multilingual and localization capabilities into the Documentum ECM platform. The system integrates with Documentum repositories and workflows to enable companies to manage their localization processes from inside the Documentum system. www.trados.com, www.documentum.com

Documentum Delivers New Portlet Development Toolset & Portlets

5/11/2004

Documentum announced the availability of a new portlet development toolset, new Documentum web content management portlets and enhancements to Documentum's existing portlets. The new Documentum portlets are a set of embeddable application components that deliver out-of-box content management capabilities to any portal that supports Java Specification Request 168 (JSR 168). The new Documentum web content management portlets add to Documentum's existing family of portlets for enterprise document management and collaboration. All of the portlets provide single sign-on functionality. The new WDK for Portlets, Documentum portlets and new Documentum web content management portlets are immediately available for use with portal servers from BEA, SUN and IBM. A set of developer and customization documentation is provided and additional materials, customization code and sample portlets are available through the Documentum Developer website's Component Exchange. www.documentum.com

Vignette & Partners to Provide Audit and Business Efficiency Tools

5/11/2004

Vignette Corp. announced that it has released a comprehensive set of tools to help organizations measure and improve business efficiency. Organizations interested in the Vignette Efficiency Audit will work with experienced third-party consultants to gather their business requirements, review overall vision and justification, conduct a gap analysis, monitor use of existing investments, and develop best practices. Upon completion of the audit, Vignette and the respective partner will provide a customized report that provides a detailed picture of areas throughout the organization where efficiencies can be increased, along with customized recommendations on how to improve overall information technology performance. Organizations will have an opportunity to see potential return on investment and total cost of ownership estimates for each option based on individual definitions. Vignette partners Acquity Group and Rapidigm are the first to offer the Vignette Efficiency Audit. In addition, Vignette has developed the Vignette Efficiency Assessment and the Vignette Efficiency Evaluation. Both programs are available at no additional charge upon completing the registration requirements. www.vignette.com

CAPTIVA SIGNS SOFTWARE LICENSING AGREEMENT WITH FILENET

5/10/2004

Captiva Software Corp. announced it has signed a software licensing agreement with FileNet Corp. FileNet will use Captiva's PixTools/Scan and PixTools/View products and bundle Captiva's ISIS scanner drivers within its products. With PixTools' Scan and View products and the bundling of ISIS, FileNet has the ability to support new scanners and evolving features such as support for color images, multiple data streams and distributed scanning. The ISIS technology also ensures support for a variety of industry scanners. PixTools/View toolkit is a full-featured API that provides all the necessary image viewing and printing functions including image display, rotation, scale, scale-to-gray and annotation. PixTools/Scan toolkit supports ISIS, driving peak performance from every scanning device and providing access to the full range of available scanner features. www.filenet.com, www.captivasoftware.com

TARARI RELEASES RAX CONTENT PROCESSOR

5/10/2004

Tarari Inc. announced the immediate availability of its latest XML Silicon technology -- the RAX Content Processor, which incorporates an in-silicon implementation of Random Access XML (RAX). RAX allows complex XML document analysis to be completed in "near-zero" CPU time and can process millions of XPaths per second. Random Access XML (RAX) enables network switch, server, blade, and appliance vendors to create new applications such as gigabit message classification and routing, high transaction rate publish and subscribe systems, advanced SOAP message processing, high performance XML security firewalls and real-time telecommunications billing solutions. Tarari is proposing that RAX be accepted as an industry-standard just as DOM and SAX have garnered many supporters within the W3C community. OEMs, ISVs and corporate developers interested in evaluating the Tarari RAX Content Processor should purchase the Tarari XML/Web Services Development Kit which consists of two Tarari RAX Content Processors on PCI cards, Random Access XML Agents (plus Encryption/Decryption Agents) and API documentation. www.tarari.com/rax

DOCUMENTUM DELIVERS ECM FOR HP INTEGRITY SERVERS & HP-UX 11I

5/10/2004

Documentum announced that it has certified key components of the Documentum Enterprise Content Management (ECM) platform for HP Integrity servers and HP-UX 11i on the Intel Itanium 2 microarchitecture. www.documentum.com

PUREEDGE DELIVERS XML E-FORMS FOR IBM LOTUS WORKPLACE 2.0

5/10/2004

PureEdge Solutions Inc. announced that its secure XML e-forms are now available for integration with IBM Lotus Workplace 2.0 and IBM WebSphere Portal. PureEdge's framework provides IBM users with the ability to integrate XML e-forms-based processes with back-end systems, and expands their on-demand capabilities by bringing a rich client to the forefront. Extended WebSphere Portal support is offered through PureEdge WebForm Server and the PureEdge Viewer, enabling XML-based forms with validation, formatting, layout and offline capabilities. PureEdge provides portlet support for zero footprint forms and for inter-portlet communication, using click-to-action data sharing functionality to leverage forms data to and from other portlets. PureEdge forms-based business process solutions are based on XML, Web Services and Java. www.pureedge.com

Informative Graphics Releases Brava! Enterprise 5.0 for Documentum Webtop & Digital Asset Management Clients

5/10/2004

Informative Graphics Corp. (IGC) announced the release of its Brava! Enterprise 5.0 view and markup software for Documentum Webtop and Digital Asset Management (DAM) clients, delivering viewing versatility for online access to documents, images and CAD engineering files. Brava 5.0 for Documentum creates and views IGC's new "content sealed format" (CSF), which incorporates the Visual Rights security framework. Visual Rights gives authors selective and persistent security controls over their content. Markups can be burned-in to CSF files, and a new block-out function (redaction) allows users to hide specific file content from view and text searching. CSF files are viewable by both Brava Enterprise and the free Brava! Reader. Brava 5.0 for Documentum also supports occasionally connected or offline computing, permitting Documentum users to save selected files in CSF format to their local machine, and then disconnect from the Brava Server to work offline. www.infograph.com

Northern Light Unveils Version 2.0 of its Enterprise Search Engine for Linux

5/10/2004

Northern Light announced that it has released a new version of its high performance enterprise search engine for Linux and Solaris. The new version includes: support for Linux, support for database indexes over 50 million documents, support for over 150 queries per second from a single software installation on a single server, Document Relevancy Booster for using editorially selected collections and specially weighted metadata to manage results list placement, and an integrated spell checker. "With Northern Light Enterprise Search Engine software at an entry level price of $2500, a $1000 computer, and a free operating system, an enterprise can process 14 million search queries a day." Other features include a configurable crawler, content filters for HTML, XML, MS Office, PDF, a 17,000 node subject taxonomy encompassing all human knowledge, a classification engine trained to automatically classify arbitrary unstructured content to the taxonomy, a clustering engine for organizing results from queries on the fly into subject categories, and sample user interfaces in PHP and JSP. The Northern Light Enterprise Search Engine for Linux can be downloaded for a 30-day free trial. www.northernlight.com

Captovation Announces Integration Between eCapture & SharePoint Portal Server 2003

5/7/2004

Captovation announced direct integration support between their eCapture software, and Microsoft Office SharePoint Portal Server 2003. With eCapture's integration, SharePoint users can scan and index paper documents directly into a SharePoint repository. The integration is facilitated through an eCapture Data Provider (EDP) which transfers document images from eCapture and places them into a SharePoint repository. The eCapture component's ecIndex, ecAutoFile Server, ecCommit Server, ecImport Server, ecNet Server and Captovation Check Capture are all capable of performing a document commit (archive) into SharePoint. Captovation's direct integration support for SharePoint Portal Server 2003 utilizes Web Services to link eCapture users with a SharePoint Portal Server system. XML is used to structure the data between an eCapture system and SharePoint. The integration includes a SharePoint Template Manager, which allows an administrator to define field mappings between eCapture and SharePoint. On a per template basis, an administrator specifies the URL of a SharePoint Portal Server. After connecting, the SharePoint Template Manager retrieves all available user lists and respective field names. The administrator then correlates field names in SharePoint with available fields in eCapture. Upon commit, the images are converted to either Multiple Page TIFFs or PDFs. www.captovation.com

SAP & Adobe Announce Interactive Forms in SAP Solutions

5/6/2004

SAP AG and Adobe Systems Incorporated announced the joint delivery of Interactive Forms based on Adobe software as part of the SAP NetWeaver open integration and application platform. The solution is available today with mySAP Business Suite. The offering is aimed at automating and streamlining paper-based communications that companies rely on to increase business agility. Both will sell, support and provide implementation services for the delivery of Interactive Forms in mySAP Business Suite. With the availability of Interactive Forms, SAP customers will be able to further enhance data capture and streamline the dissemination of business-critical form processes. Whether government to citizen, business to business, or business to consumer, SAP customers can reach a much broader set of users inside and outside the firewall using Interactive Forms. www.adobe.com, www.sap.com

Centra Unveils Version 7.1

5/5/2004

Centra Software, Inc. announced a series of enhancements to its core software platform, Centra 7. Centra 7 version 7.1 features greater integration for contextual collaboration with Microsoft Office, provides integration with Public Switched Telephone Networks (PSTN), streamlined pre- and post-meeting attendance and management capabilities, enhanced Knowledge Center reporting, increased enterprise-class administration features and enhanced branding capabilities. Created to support Centra's four solution sets, Enterprise Application Rollouts, Sales Effectiveness, Collaborative Learning and Customer Acquisition, Centra 7 version 7.1 lets users monitor the progress of application training programs, facilitates project team meetings and provides a way for sales people to conduct customer meetings and online demonstrations. High-touch events like Centra-powered online Web seminars are now easier to setup, attend and record. Centra 7 version 7.1 makes it easier for learners to build their own blended learning programs. www.centra.com

SchemaLogic & Meta Integration Technology in Agreement

5/5/2004

SchemaLogic and Meta Integration Technology announced they have formed a technology alliance where SchemaLogic customers will be able to utilize over 50 additional adaptors to create a shared, cross-system, "active" metadata repository. Meta Integration Model Bridge connects to databases or information models from IBM, Oracle, Sybase, SAS, Business Objects, IBM Rational, Computer Associates, OMG and W3C. SchemaLogic provides a framework for shared metadata based on an active repository, a unified information model, collaborative change management with impact analysis, notification and approval, plus the synchronization of approved changes to subscribing systems using XML, SOAP and Web Services. This provides a holistic view of information assets including content, data and XML: who is responsible for each asset, how they're organized (structure and semantics) and the relationships among them. Information architects, database analysts, content system managers and developers can see and control metadata definitions, taxonomies, hierarchical lists and vocabularies in one repository, available throughout the enterprise. www.metaintegration.com, www.schemalogic.com

RICOH INTRODUCES GLOBALSCAN FRAMEWORKS FOR DOCUMENTUM & INTERWOVEN WORKSITE

5/5/2004

Ricoh Corporation introduced GlobalScan frameworks for Documentum and Interwoven Worksite for use with existing Ricoh Aficio Multifunctional Products (MFPs). Using Ricoh's Aficio MFP combined with GlobalScan Sever software and the framework to Documentum, paper-based documents can be scanned for delivery to the Documentum doc-base. By combining Ricoh MFPs, GlobalScan and the Interwoven WorkSite frameworks, users can convert paper documents into a variety of file formats including PDF, Microsoft Word, Excel or PowerPoint. They can then distribute electronic files directly to worksite databases with a single scan. www.ricoh.com

ARBORTEXT SHIPS ARBORTEXT 5.0

5/5/2004

Arbortext, Inc. announced it is shipping Arbortext 5.0, its latest product release that expands the publishing software suite's functionality and ease of use. Arbortext 5.0 features four major new products and many enhancements, including: Styler, a new development tool that allows designers to create stylesheets to drive automated publishing from a "single source of style" to print, PDF, Web, HTML Help, and wireless devices; Contributor, a Web-based XML editor that runs in a browser and requires no desktop installation. Contributor provides simplified XML editing; Companion, an add-in for Microsoft Word 2003, Companion simplifies the development of XML authoring applications for Word; and DCAM, an optional component of Arbortext's E3 publishing server that lets authors create and manage the hundreds or thousands of inter-document and intra-document links that typically exist in a large collection of information, and it supports the publishing of dynamic documents that contain links. www.arbortext.com

IBM & SYNKRON A/S PARTNER

5/4/2004

IBM has entered into a strategic portal partnership with Danish CMS vendor, Synkron A/S. By integrating the Synkron.web CMS into IBM's WebSphere technology, enterprise customers can now get content management coupled with portal functionality. www.synkronweb.co.uk

KOFAX ANNOUNCES SUPPORT FOR IBM'S DB2 CONTENT MANAGER EXPRESS

5/3/2004

Kofax announced support for IBM's DB2 Content Manager Express Edition V8.2. The Ascent Capture integration module for DB2 Content Manager Express offers users access to a production capture solution, including document, data and Internet-based distributed capture, that helps input information into IBM's content management solution. IBM's DB2 Content Manager Express software is designed to help medium-sized organizations implement a holistic digital content management infrastructure. Through the integration of Ascent Capture, these organizations can now collect large volumes of forms and documents, transform them into retrievable electronic information, then deliver it into IBM's DB2 Content Manager Express. Ascent can also capture e-documents and XML streams as well as enable front-end integration with enterprise applications. www.kofax.com

Vasont Integrates with Adobe FrameMaker

5/3/2004

Vasont Systems announced that its Vasont SG content management software now seamlessly integrates with Adobe FrameMaker and its XML capabilities through its Vasont Universal Integrator (VUI) software extension. Designed for small editorial groups with limited resources, Vasont SG is a low-cost version of the Companys content management software that helps users more efficiently create, manage and repurpose their complex content, including technical documentation, product or users' manuals, and reference materials, to print, Web, CD-ROM, and wireless formats. The VUI streamlines the writing and editing process by providing a simple editorial interface so that authors and editors can access Vasont SGs content management functionality from the toolbar menu of FrameMaker. Additionally, Vasont SG tracks all reuse, repurposing, and versioning of content, even as changes are made through FrameMaker enabling small editorial groups to create, manage and publish their content from a single source.
www.vasont.com

Authentica Announces Professional Services Group

5/3/2004

Authentica, Inc. announced the formal introduction of its Professional Services Organization. The group now offers a full suite of implementation and custom consulting services to Authentica's enterprise customers. Authentica's Professional Services enable customers to maximize the benefits of their E-DRM solutions by offering: fixed-fee packages that provide rapid deployment and training for customer-specific business solutions for workgroups, departments and enterprises; extended analysis and consultation to help customers fully customize and integrate their E-DRM solutions within their enterprise infrastructure. These services include applications integration, custom security policy models and workflow enhancements. www.authentica.com

Adobe Acquires Q-Link Technologies

5/3/2004

Adobe Systems Incorporated announced that it has acquired Q-Link Technologies, Inc., a provider of business process management software. The acquisition provides Java-based workflow technology that will be integrated with the Adobe Intelligent Document Platform to help customers reduce document processing cycle times, eliminate bottlenecks and integrate business processes more easily and efficiently across the extended enterprise. The terms of the acquisition were not disclosed, although Adobe said the acquisition will not have a material financial impact on the company. A privately held company based in Tampa, Florida, with customers in government, financial services, manufacturing and telecommunications, Q-Link provides a J2EE component based workflow architecture that enables developers to visually assemble complete applications. From workflow to forms, UI design, rules, web services, integration and Business Activity Monitoring (BAM), Q-Link offers a solution to build process-driven applications.
www.adobe.com

SAQQARA Introduces CDM for e-Procurement

5/3/2004

SAQQARA Inc. announced SAQQARA Commerce Data Management (CDM) for e-Procurement. SAQQARA CDM is a managed service that enables enterprises to reach their spend management (ESM) objectives by providing and maintaining commerce data. CDM for e-Procurement replaces SAQQARA's Catalog Management Solution (CMS). SAQQARA manages the entire catalog management process including supplier engagement, content acquisition, content transformation, exception management, filtering, and an aggregated supplier catalog for use through e-procurement systems. SAQQARA CDM for e-Procurement works transparently with e-Procurement systems from Ariba, Microsoft, Oracle and SAP. Features include supplier enablement and on-boarding of catalogs, supplier management, product content management, catalog management and an optimized purchasing interface. SAQQARA CDM for e-Procurement will be available in Q2. Pricing is based on platform access and per-supplier charges. www.saqqara.com

The Gilbane Conference

ON CONTENT MANAGEMENT TECHNOLOGIES

Nov 30 - Dec 2, 2004 • Boston, Massachusetts

New Technologies & Best Practices

http://www.gilbane.com/CM_conference_Boston_04.html

Everybody knows they need “content management”. Content management technologies are now mainstream and need to be part of all major enterprise applications, and integrated into IT architectures and infrastructures. But what does that mean? To some people content management is all about web publishing, to others it means managing multiple, perhaps all, types of unstructured data. Still to others, the term might be associated with a particular aspect of managing information, such as searching for it, organizing it, transforming it, or sharing it. We have organized our Boston conference into six tracks that are focused on the most critical components of content management. Attendees can immerse themselves in specific areas directly related to a project, or gain a broad understanding of all the relevant technologies in order to design a well-informed content strategy for an enterprise or department. The tracks are:

TRACK DESCRIPTIONS

Content Management (CM track)

Our content management track covers what most people consider content management issues and technologies. This includes both web content management (WCM) and enterprise content management (ECM). Whether you are a project manager, IT strategist, or business manager, you will find every aspect of planning, selecting, building, deploying, enhancing or replacing a content management system covered in this track. The faculty includes a wide range of well-known and respected content management consultants and authors, so you can be sure what you learn is based on experience and expertise.

Document & Records Policy & Management (DM track)

Document and records management are receiving dramatic increases in attention these days. The need for both has not diminished with the increase in web applications. Instead, there are more (electronic) documents and records than ever to be managed. And with ECM vendors acquiring these capabilities and marketing their newly combined (or not) solutions, companies are being re-educated in the need to manage *all* their content. Combine this and with new security and compliance requirements associated with Sarbanes-Oxley, HIPAA, etc., and the result is a need for companies to review the rules and policies regarding many different types of documents, records, and digital content. This track will look at how to deal with these issues in today’s more complicated digital corporate environments.

Digital Asset Management (DAM track)

There is certainly a large overlap between Digital Asset Management (DAM) and ECM. While analysts and practitioners may debate the *degree* of overlap in functionality, we don't know of anyone who would argue that there are not applications that require a set of capabilities that are not usually found in ECM systems. This track covers DAM issues and technologies for enterprises, with an emphasis on marketing, publishing, and brand applications. If you are interested in DAM, you should look at this track as well as the CM Track.

Enterprise Search, Knowledge Management & Collaboration (KM track)

Each of these topics are important on their own, but when you look at the way most companies implement KM or collaboration or a portal, you will always find some combination of these approaches and technologies, usually also combined with categorization and taxonomy efforts. In this track you will learn about these technologies, as well as how people actually use them to share and leverage existing and emergent knowledge, especially in R&D environments.

Enterprise Information/Content Integration (EII track)

The problem is simple and familiar; you have built multiple content, document, and database repositories across your organization. They may all work just fine on their own, and you may have even connected a few using EAI, or implemented a way to share information between them using some kind of a conversion tool. But you have realized that to accomplish the kinds of productivity, efficiencies, and ROI you expected from your individual applications, you need *information* integration, which is much more complicated than *application* integration, to be the rule rather than the exception. To accomplish that, you need easy, fast, reliable, and cost-efficient ways for applications to share content. This track will look at the issues and technologies and approaches associated with this increasingly important area.

Content Technology Works (CTW track)

This track focuses on analyses of successful deployments of content technologies from the enterprise user's perspective. While we include some case studies in other tracks, the case studies in the CTW track have been looked at much more closely by our analysts. There are many implementations of content management technologies, but the industry is still young enough that accepted best practices are difficult to find. Our CTW program is designed to identify and verify emerging best practices, and in the meantime, to share successful implementations in a marketing-free environment so that more companies can gain the confidence to implement and benefit from content technology.

SUBSCRIPTION FORM

You can also order on our *secure* website www.gilbane.com.

☐ Please start my electronic subscription to the Gilbane Report for only $99. (10 issues/year). Subscription includes access to HTML and PDF versions at www.gilbane.com. (Call for print subscriptions, site license prices, and back issues.)

☐ I am eligible for an affiliate discount* ______________________ Affiliate organization________________ Tracking #

☐ My check for $__________ is enclosed ☐ please bill me
Please charge my credit card ☐ MasterCard ☐ Visa ☐ American Express

Name as on card: _______________________________________Number _______________________________________
Signature ____Expiration date _________________________________

Name____ Title____
Company____ Department__________________________________
Address____
City___________________________________ State/Province________________Zip/Postal Code_________________
Country________________Tel.____________________Fax_________________E-mail______________________________

Checks from outside the U.S. should be made payable in U.S. dollars.
Funds may be transferred directly to our bank, please call for details.
Mail this form to: Bluebill Advisors, Inc. 763 Massachusetts Ave., Cambridge, MA 02139, USA.
You can also place your order at www.gilbane.com or by phone (+617.497.9443), or fax (+617.497.5256).

CALENDAR *(SUBSCRIBERS: LOGIN TO THE GILBANE.COM SUBSCRIBER SITE FOR YOUR CONFERENCE DISCOUNTS!)*

IQ Boot Camp - Optimize Your Website, *28 June - 9 July, London*. iQ Content's learning events offer private and public sector organisations the opportunity to maximize their web potential. Choose from 10, one-day workshops covering essential aspects of web optimisation, including accessibility, usability, web writing, email newsletters, intranet strategy and online marketing. Workshops are strictly limited to 12 places and feature expert instruction, interactive exercises, extensive tools and resources and peer-to-peer networking. **Gilbane Report subscribers receive a 10% discount!** Full details are available at www.iqcontent.com/solutions/boot_home.htm

Seybold San Francisco 2004. *August 16-19, 2004 - Moscone West, San Francisco.* In 2004, Seybold SF introduces a conference program format that focuses on continuing education for professionals in the four key aspects of content creation and publishing: Creating content, Publishing content, Marketing content, and Managing content. **Gilbane Report Subscribers receive a $200 discount!** Use this Education Voucher that can be applied to either the Platinum or Gold Passports and waives the Pavilion entrance fee. To redeem, simply register with the priority code on the voucher. http://www.seybold365.com/sf2004/

The Gilbane Conference on Content Management Technologies. *Westin Copley Place, Boston MA, November 30 – December 2, 2004*. Our Boston event is being launched to complement our other content management conferences with an anchor event that covers all major content technologies. Our other conferences focus on topics for businesses embarking on a content management project. Our 3-day Boston conference will still include everything a project team needs to know, but will also offer a look ahead at upcoming technologies, "new" best practices, and a broader look at technologies necessary to supplement core content management applications. We are accepting proposals for speaker presentations and panels through May 15. www.gilbane.com/CM_conference_Boston_04.html or www.lighthouseseminars.com

The Gilbane Report is published 10 times a year (monthly, except for August and December). The Gilbane Report is an independent publication offering objective analysis of technology and business issues. The Report does not provide advertising, product reviews, testing, or vendor recommendations. We do discuss product technology that is appropriate to the topic under analysis, and welcome product information from vendors. Letters to the editor are encouraged. Send to: editor@gilbane.com. Visit our web site at www.gilbane.com

ISSN 1067-8719