TOC: Metadata is not going away, and there is no one simple solution to how to add metadata and maximize its value. So let's take a look at some of the basic issues around adding metadata to unstructured content and explore a range of approaches that various groups and software vendors are trying.

To Metadata or Not To Metadata

—TOM REAMY

To metadata or not to metadata, that is the question.
Whether 'tis nobler in the mind to suffer the slings and arrows of outrageous search results
Or to take up metadata against a sea of irrelevance
And by organizing them, find them?

With all due apologies to the Bard, the questions of whether to add metadata to unstructured content and how much effort is really justified to do so have been raised with increasing frequency and vigor in the last year.

These issues and more were explored last year at the Dublin Core Metadata Initiative (DCMI) 2003 Workshop. While some participants argued for a drastic reduction in metadata efforts or at least rethinking those efforts, other participants offered new ideas of how to create valuable metadata and how to generate value from metadata.

A couple of things have become increasingly clear: Metadata is not going away and there is no one simple solution to how to add metadata and maximize its value. Consequently, what we are going to do in this article is take a look at some of the basic issues around adding metadata to unstructured content and explore a range of approaches that various groups and software vendors are trying. We will then examine how a broader view of metadata, beyond simply adding keywords to documents, is leading to a more sophisticated, multi-dimensional or infrastructure-based approach to metadata that supports a smarter balance of both more and less metadata.

TOO HARD, TOO MUCH, NO HELP

A number of issues have been raised about the effectiveness and value of adding metadata. The first issue is the cost of adding metadata, and the second is the difficulty of doing it well and the associated problem that poor-quality metadata can actually make search worse than no metadata at all.

Let's start with the cost argument. One participant at DCMI, Mike Doane, senior content analyst SBI and Company, cited his practice in which he charged between $150,000 and $250,000 for a full-scale metadata implementation. This can certainly seem like an exorbitant amount of money especially for a company that is still using a $10,000 search engine for its intranet. In addition, this expense is just for adding metadata to a large existing content repository but doesn't take into account the additional cost of maintaining and adding new metadata.

In addition to cost, another argument against adding metadata is the immense difficulty of doing it well. From my own experience and that of others, the difficulty of effectively employing metadata can easily be seen in the abysmal quality of the metadata associated with the unstructured content found on most corporate intranets. In evaluating corporate intranets, time and again we find missing metadata fields, missing values from the fields that have been defined, very poor quality values in even such simple fields as the title (ex23a.pdf is not very illuminating as a document title), inconsistent values among similar documents, and inconsistent values among authors. One interesting aspect of bad metadata is that it doesn't just detract from getting full value from the effort to add metadata; bad metadata seems to make search function worse than having no metadata at all.

Given the rather pathetic record that many metadata efforts have racked up, it is little wonder that organizations have begun to question the entire value of adding metadata. However, there is another side to the story. First, the cost of adding metadata can be reduced in several ways. For example, the $200K for a metadata initiative performed by outside consultants can be greatly reduced by not starting from scratch in each case, but rather starting with existing metadata standards and controlled vocabularies and taxonomies. The cost of a unique custom job will always be higher than one that at least starts with predefined components.

In addition, the cost of doing metadata has to be weighed against the cost of not doing metadata. Assume for the moment that adding metadata would solve all of the problems associated with search. One estimate from IDC puts the cost of bad search at $6 million for a 1,000 person company. Now it is unlikely that adding metadata will solve all search problems, but even if it only solves half, that is still a savings of $3 million per year. In this context, $200,000 for metadata doesn't seem so exorbitant.

SIMPLIFY

Now let's assume that you have decided that it is worthwhile to at least explore different approaches to adding metadata, how do you proceed? Three approaches are guaranteed not to work: One is to hire consultants, but this has a high upfront cost and an ongoing maintenance costs. Two is to ask your authors to create metadata as they publish, but this leads to very low quality metadata, especially keywords, which require a special skill that has nothing to do with subject matter expertise (not to mention the difficulty of getting them to actually add it at all). Three is to use automatic metadata generation software, but the software often costs as much a consultants and does a worse job.

At the DCMI 2003 Workshop, a different approach to navigating the metadata dilemma was discussed at some length—

the content-value-tier model offered by information architecture expert Lou Rosenfeld. The idea was fairly simple: Focus on a practical solution, focus on high-value content, and don't try to solve all the world's problems. High-value content can be specified using a variety of criteria like authority, popularity, currency, strategic value, and reusability. Then you can choose to add full metadata to high value content and less or none to low value content.

Unfortunately, even this approach has its shortcomings. One problem is that it doesn't really solve the problem of how best to add good metadata; it simply tries to limit the problem. A second problem is that, in my experience, it creates a number of new problems, first of which is the political dimension. If you think that wars over placement on the home page can be vicious, trying to manage who gets metadata and who doesn't can be worse. And then there is the issue of who gets to decide what is of high value, which is another political minefield.

The use of relatively objective measures can help, but such measures have poor track records themselves. For example, popularity does not really correlate all that well with value, particularly in an intranet environment. It should be pointed out that attaching determination of the value of content based on criteria like authority, popularity, and the like is adding metadata to content. It just utilizes different metadata fields and applies metadata to collections instead of documents.

While I don't believe that this model provides an ideal solution, it does point in the right direction. It is based on looking at and differentiating content, and it uses multiple approaches such as a set of criteria for high value content. Finally, it is a possible, practical solution in certain cases, particularly when developed within an articulated strategic vision.

INTELLECTUAL INFRASTRUCTURE

The first step in finding the right solution, or rather, the right set of solutions, is to examine the issue of metadata within a broad context of information and knowledge needs, or what I call the intellectual infrastructure context. It is important to look at metadata within this broad context to enable the full set of answers to how to add metadata and how to utilize metadata. In some cases, that might mean less metadata, but in others it will mean more. By viewing metadata as an add-on to a search engine project, you are essentially guaranteeing that you won't come up with the best set of solutions.

This intellectual infrastructure includes all kinds of content—structured and unstructured, internal and external, document-based and tacit knowledge inside the heads of employees. It includes metadata, taxonomies, controlled vocabularies, database schemas, persona models, and other knowledge organization structures. It also includes the publishing policies and procedures as well as the people who develop and support the creation and utilization of all the kinds of content. And finally, it includes information technologies like search engines, content management, portals, categorization and visualization software, and other applications that information and knowledge workers routinely use.

[Figure 1: Inxight Tree]

For example, content management is an essential part of any attempt to add metadata. Good content management software can support the integration of your metadata standards and, more importantly, controlled vocabularies. Good content management can also support various automation and workflow capabilities that can be used to increase the quality of metadata and decrease the cost.

Another component of an infrastructure approach to metadata is the makeup and services of a central team of—yes—people. This team should be a cross-organizational team with library science well represented, but also business analysts, user-focused individuals (anything from usability people to cultural anthropologists), and software specialists. This team can perform a number of functions that will lead to better and cheaper metadata. First they would be in charge of creating, acquiring, and evaluating taxonomies, metadata standards, and controlled vocabularies. This team would also research metadata theory like the latest RDF proposals.

Another function of the team would be to work with authors, evaluating metadata quality and methods for facilitating author-created metadata on the one hand, and analyzing the results of using metadata, tracking how the enterprise's communities were utilizing the metadata, on the other. Finally, this central team would also perform an essential, but often overlooked, role: socializing the benefits of metadata and helping to create a content and user-centric culture to replace the technology-centric culture too often found in information groups.

INFORMATION INFRASTRUCTURE

A lot of the discussion about metadata has really been about one metadata field: keywords. This is probably due to the unfortunate fact that keywords are perhaps the most difficult metadata field to implement, or at least to do well enough to get the results people expect. For example, when selecting keywords for a document, should you select words that are frequently found in the document, terms that are unique to the document, or terms that try to express the "aboutness" of the document. There are arguments for all three, but none are completely compelling nor do they uniformly produce good results, and what is worse, it usually becomes an individual author's choice which means an unorganized mix of answers and results.

So does this mean that keywords are of no real value? No, it means that they have to be approached from an infrastructural perspective. And the essential first step to producing good keywords is to develop a controlled vocabulary or, even better, a taxonomy-based set of controlled vocabularies with which to populate the keyword field. As we have seen, asking authors to create good keywords simply does not work very well, but asking them to select the right keywords from a predetermined list is much easier and produces better and more consistent results.

In addition to doing keywords better, it is important to realize that there is more to metadata than just keywords. Other metadata fields can often produce high value and can be much easier and cheaper to produce. It is important to focus on achieving value from all fields, such as titles, descriptions, publisher, author, and the like. And often even more valuable are fields like audience and DocumentObjectType (with values like an FAQ document, a policy document, and so on).

While software that claims to solve all your metadata needs is still illusory, there are a number of products, like Entopia's K-Bus, that can generate a great deal of very useful metadata and thus reduce the overall cost of metadata projects, as well as support the development of the sophisticated search applications.

[Figure 2: Entopia metadata]

INFRASTRUCTURE CONTEXT

Implementing metadata initiatives as a fundamental component of the intellectual infrastructure of an organization rather than simply as keywords used to influence relevance ranking supports a wide range of interesting and valuable applications that go beyond simple search, and, at the same time, enhances the search experience in a variety of ways.

One such application was the faceted metadata display presented at the DCMI Workshop by Marti Hearst, associate professor at SIMS (School of Information Management and Systems at Berkeley). In this application—Flamenco—search results are mapped to a large number of facets, which basically function as a well-structured set of advanced, tightly defined searches. This allows the user to select likely areas from which to browse to the document they seek. The well-defined facets like Products, Geography, Health Effects, and Document Characteristics work much better at limiting the results in meaningful ways than the usual mixed, broad categories you find in browse applications.

Research at SIMS with the Flamenco Search Interface Project has shown that even though the display is complex users find it quite easy to master. We all know that advanced search using metadata fields doesn't work for the simple fact that users won't do it. Advanced searching is an advanced skill which most users (outside of the library) don't have. But as Igor Perisic, chief scientist of Entopia, put it, "give them a simple, empty search box and then add structure to the results," which what Entopia's K-Bus does. Just as selecting keywords from a list works better than making up your own keywords, so selecting from the results of multiple advanced searches works better than making up your own.

WHAT IF I CAN'T GET THERE FROM HERE?

Rosenfeld pointed out that in his experience, not many organizations are willing to commit to such a huge undertaking as developing a corporate taxonomy, an enterprise-wide metadata standard, and associated controlled vocabularies, and then implementing that standard in 100,000 documents or more; integrating that metadata into an entire range of projects and technologies like search, content management, portals, and the like; and, at the same time, creating a complete metadata or knowledge architecture team to manage the whole thing.

As Rosenfeld so aptly expressed it, "It is a worthy pursuit, but we can start with other easier, low-hanging fruit, before taking on the huge honking thing like an enterprise thesaurus."

So how should one proceed on a practical level?

I would argue that in my experience (and Rosenfeld agrees) the best results start with creating the overall infrastructure vision including metadata standards. While actually implementing this vision can be expensive (though likely not as expensive as not doing it), creating the vision itself is a relatively small project. What having the strategic vision does, however, is to create the right context within which to implement and justify any and all piecemeal or smaller projects, avoiding reinventing the wheel for each information project and leveraging each project as a foundation for the next project.

The next essential step is to create a team, which need not be a large team nor is it essential that it be a full-time, dedicated team. It can be a virtual team made up of members from a library, IT, business partners, and so on. What is essential, however, is that the team has some sort of official recognition, including incorporating their central team functions into their job descriptions and reward structure. Another early step could be a content management initiative—before the initial reaction to your new portal project changes from rave reviews to user complaints of still not being able to find anything.

As far as metadata itself, I would recommend that you not start with keywords if you don't have the resources to develop them with controlled vocabularies. Instead, focus on getting value from other metadata fields. Another option is to buy and customize an existing taxonomy and/or vocabulary. Finally, don't focus on trying to tweak relevancy rankings with keywords, but try such approaches as best bet metadata, browse or dynamic classification, or faceted metadata interfaces.

There are many other tips and techniques for implementing a full-scale, enterprise-wide infrastructure solution to metadata, but that would take us too long, and they will vary from organization to organization. So let me sum up the approach with this slogan and a question:

Think Big, Start Small, Scale Fast.

You wouldn't think of running a company without organizing your employees, why do you think you can create access to information without organizing that information? EC

TOM REAMY (tomr@kapsgroup.com) is chief knowledge architect for KAPS Group, a group of knowledge architecture, taxonomy, and elearning consultants. He writes for *Knowledge Management*, *Intranet Professional*, and *KMWorld*, and is a frequent speaker at KM conferences.

Comments? Email letters to the editor to ecletters@infotoday.com.

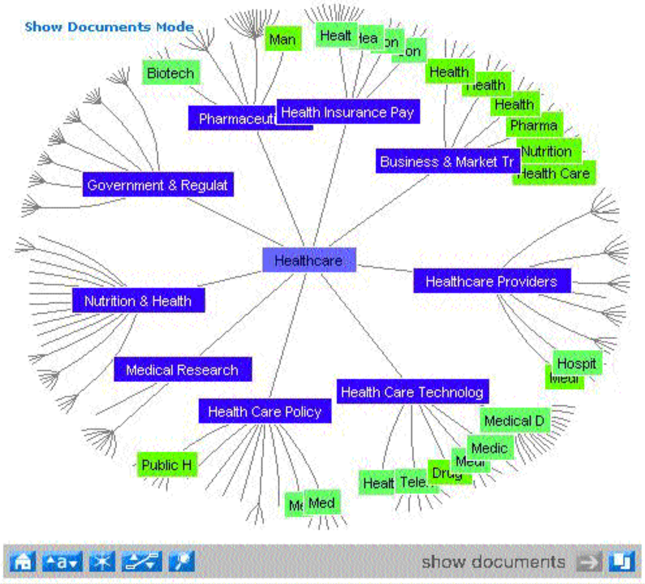
Companies Featured in This Article

Entopia
www.entopia.com

Flamenco Search Interface Project
http://bailando.sims.berkeley.edu/flamenco.html

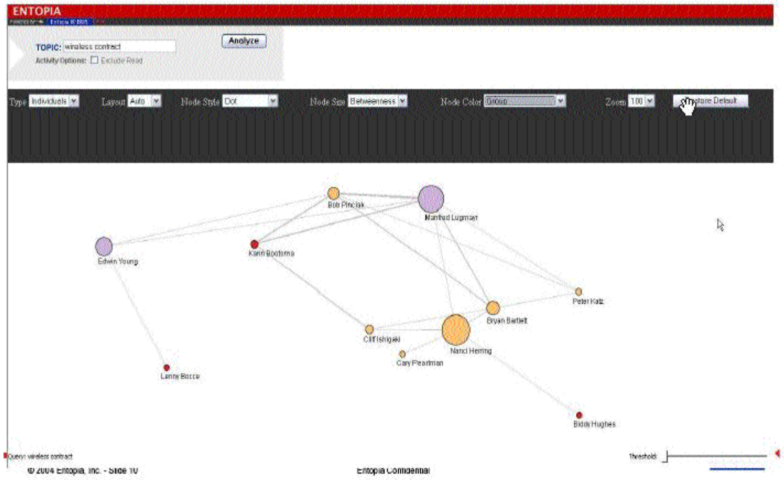
Inxight
www.inxight.com

[Figure 1: Inxight Tree]



Inxight's Taxonomy Manager. A visual display of taxonomic relationships make developing and maintaining a taxonomy much easier and with better

[Figure 2: Entopia metadata]



TOPIC: wireless contract

Activity Options: □ Exclude Read

Type Individuals Layout Auto Node Style Dot Node Size Betweenness Node Color Group Zoom 100

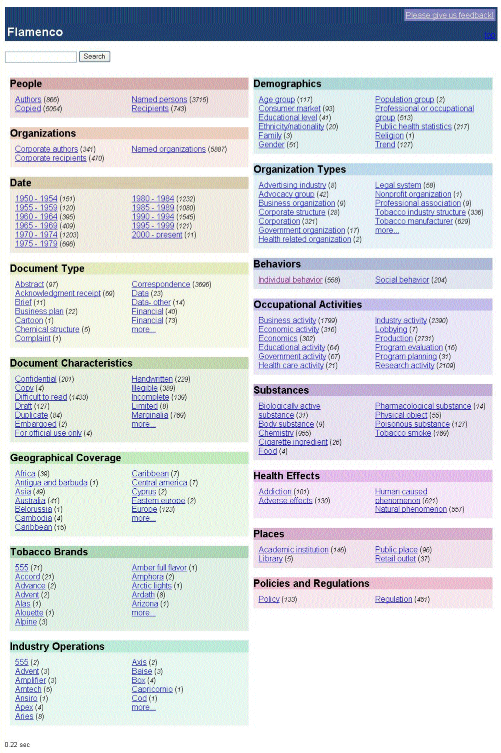
Query: wireless contract

© 2004 Entopia, Inc. - Slide 10

Entopia Confidential

A visual display of automatically generated metadata by Entopia, in this case a variant of the author field with the size of the node representing the number of documents.

[Figure 3: Flamenco Full]



People

Authors (866)
Copied (5054)
Named persons (3715)
Recipients (743)

Organizations

Corporate authors (341)
Corporate recipients (470)
Named organizations (5887)

Date

1950 - 1954 (151)
1955 - 1959 (120)
1960 - 1964 (395)
1965 - 1969 (409)
1970 - 1974 (1203)
1975 - 1979 (696)
1980 - 1984 (1232)
1985 - 1989 (1080)
1990 - 1994 (1545)
1995 - 1999 (121)
2000 - present (11)

Document Type

Abstract (97)
Acknowledgment receipt (69)
Brief (11)
Business plan (22)
Cartoon (1)
Chemical structure (5)
Complaint (1)
Correspondence (3696)
Data (23)
Data- other (14)
Financial (40)
Financial (73)
more...

Document Characteristics

Confidential (201)
Copy (4)
Difficult to read (1433)
Draft (127)
Duplicate (84)
Embargoed (2)
For official use only (4)
Handwritten (229)
Illegible (389)
Incomplete (139)
Limited (8)
Marginalia (769)
more...

Geographical Coverage

Africa (39)
Antigua and barbuda (1)
Asia (49)
Australia (41)
Belorussia (1)
Cambodia (4)
Caribbean (15)
Caribbean (7)
Central america (7)
Cyprus (2)
Eastern europe (2)
Europe (123)
more...

Tobacco Brands

555 (71)
Accord (21)
Advance (2)
Advent (2)
Alas (1)
Alouette (1)
Alpine (3)
Amber full flavor (1)
Amphora (2)
Arctic lights (1)
Ardath (8)
Arizona (1)
more...

Industry Operations

555 (2)
Advent (3)
Amplifier (3)
Amtech (5)
Ansiro (1)
Apex (4)
Aries (8)
Axis (2)
Baise (3)
Box (4)
Capricornio (1)
Cod (1)
more...

Demographics

Age group (117)
Consumer market (93)
Educational level (41)
Ethnicity/nationality (20)
Family (3)
Gender (51)
Population group (2)
Professional or occupational group (513)
Public health statistics (217)
Religion (1)
Trend (127)

Organization Types

Advertising industry (8)
Advocacy group (42)
Business organization (9)
Corporate structure (28)
Corporation (321)
Government organization (17)
Health related organization (2)
Legal system (58)
Nonprofit organization (1)
Professional association (9)
Tobacco industry structure (336)
Tobacco manufacturer (629)
more...

Behaviors

Individual behavior (558)
Social behavior (204)

Occupational Activities

Business activity (1799)
Economic activity (316)
Economics (302)
Educational activity (64)
Government activity (67)
Health care activity (21)
Industry activity (2390)
Lobbying (7)
Production (2731)
Program evaluation (16)
Program planning (31)
Research activity (2109)

Substances

Biologically active substance (31)
Body substance (9)
Chemistry (955)
Cigarette ingredient (26)
Food (4)
Pharmacological substance (14)
Physical object (55)
Poisonous substance (127)
Tobacco smoke (169)

Health Effects

Addiction (101)
Adverse effects (130)
Human caused phenomenon (621)
Natural phenomenon (557)

Places

Academic institution (146)
Library (5)
Public place (96)
Retail outlet (37)

Policies and Regulations

Policy (133)
Regulation (451)

0.22 sec

A faceted metadata display from the Flamenco project at Berkeley SIMS shows a results set from a simple search with the number of documents within each node of each facet, allowing a browse to be launched from each node.