Metadata and content management systems: an introduction for indexers

Fred Leise

Indexers need to know about metadata and content management systems, first because although more and more companies are implementing systems that need proper metadata associated with them, there is a dearth of people with expertise in metadata development. Second, creating metadata uses the basic skills of concept identification and term selection that are central to indexing, although those choosing to move into metadata creation will also need to acquire a new skill set and become familiar with a new set of tools.

About content management systems

Content management systems (CMSs) are software tools that allow for the efficient electronic handling of large collections of documents. As illustrated in Figure 1, they have three major subsystems: a content authoring system, a document library or data store, and a publishing system. Let us examine each of the three systems in turn.

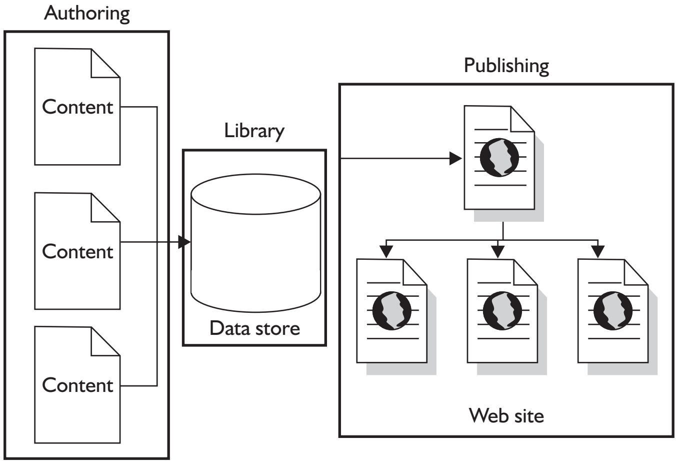

Figure 1 Content management system.

Content authoring system

In large companies, there are significant difficulties associated with the creation of documents. First is the issue of standardized formats. Many documents must be used enterprise-wide, with an audience of hundreds or thousands. Often there will be several or dozens of people creating documents of the same type – contracts, for example – and each individual might be predisposed to use a different style. Some authors include their name as author, some do not. Some authors include an 'abstract,' while others create a 'summary' section.

Second, there is the problem of version control. If several people are working on a document at the same time, it is all too easy for one person to overwrite another's changes and corrections. It might be difficult to know which of several versions is the most up-to-date.

Finally, there is the question of workflow. That is, what happens once a document is produced? Does the document need to be approved by someone? Does it need to go to legal counsel? Does it need to be routed to another department for approval? Often a document's travels can be circuitous, to say the least. And what happens when the person who is supposed to approve a document is out of town? Or when a document gets buried on someone's desk?

All of these problems can be mitigated using a CMS. Content management systems have the ability to provide customized document templates so that all content authors must enter the same kind of information for the same kind of document.

A simple document template might look like the one shown in Figure 2.

Using such a template, content authors do not have to worry about structure or formatting. The CMS handles the formatting automatically through coding that assigns a style to each field in the template. Fields in the template may be required or optional (for example, if document title is a required field, all authors have to enter a document title). Some fields (for example, date and author) may be filled in automatically when the content author logs on to the system.

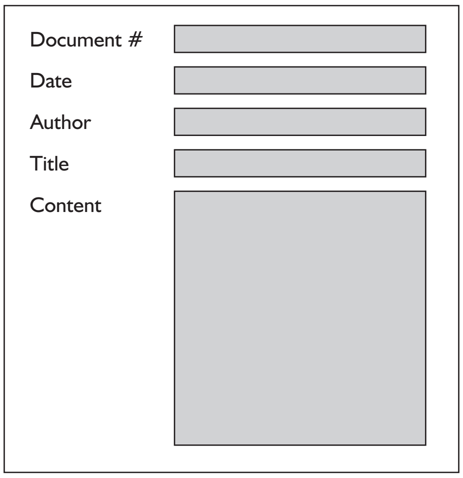

Figure 2 A simple document template.

One important way in which CMSs control versioning is through a checkout feature. That is, once a document is in the system, an author who wishes to change it must 'check it out.' The CMS then locks the file so no one else is able to make changes while the first author is working on the document. The CMS might also save a different copy of the file every time a document is checked out, thus allowing an author to go back to any previous version.

Many CMSs enable 'workflows' to be programmed. This means that when a specific content author creates a document and marks it as 'complete,' the system will automatically send a notification to the appropriate individuals to let them know that the document is ready for approval. If necessary, a supervisor can re-route the document back to the author with suggestions for changes. This whole process can be accomplished using electronic rather than hard copies.

Document library

The central portion of the CMS is the document library, sometimes called the document repository or data store. Here, documents are kept in electronic form and made available to anyone with access to the system and the appropriate clearance. Sometimes the document library will have a subject browsing structure, allowing navigation to specific documents. Some CMS libraries have text-searching capabilities. Others retrieve documents based on author-supplied keywords. Often documents can be retrieved by searching on a specific field, such as author name, document type, or creation date.

Publishing system

Finally, most CMSs have a publishing system through which documents or content objects are automatically posted to a website. In fact, the entire website might be created 'on the fly' from stored content, rather than each page being hard-coded. The CMS uses a series of rules that tell the system where to publish content. A website's Media Center might have a page listing recent media releases. The CMS publishing system might have a rule saying something like: 'If document type = "media release" and date is greater than (after) May 1, 2004, list document title in this box, sorting all titles by date.'

Such rules need to be established for each area of the website, making the customization of the publishing system a significant part of the cost of implementing the CMS. But the benefit is that changes to the site happen automatically. In the instance cited above, if the publishing system updates the website every four hours, any new media releases would automatically have their titles displayed on the appropriate page in the Media Center section of the site. Clearly, automating this process saves considerable time and effort.

Also, because website pages are produced automatically, a single piece of content residing in the content library can be used in multiple places in the site. When the document is updated and re-stored in the document library, the updated information appears in every place that the document was used on the site. This ensures greater consistency than is possible by making manual changes in each location where the document appears.

Now we turn to a description of metadata and why it is important for content management systems.

About metadata

Basically the term 'metadata' refers to data about data, but that definition does not fully explain the concept. More specifically, metadata is a set of terms that can be used to tag documents or other content objects. We use metadata all the time in our daily lives. When shopping for clothing, for example, we often first look at a garment's size or price or designer, which are all information about the actual piece of clothing we are considering buying. This information describes the clothing and informs our decision whether to buy it.

If we are considering buying a car, we want to know its engine size and gas mileage, manufacturer and model. We want to know whether it is new or used – its 'ownership status.' When we take a trip we put a luggage tag on our bag, identifying it as ours. That luggage tag contains metadata: that is, it describes the bag in a way that lets the airline or other carrier know to whom it belongs. The airline also puts a computerized tag on the bag that identifies the flight number and destination, which is metadata designed to make sure the bag reaches the right place.

When we go to a doctor, a nurse often obtains basic information about us: age, weight, height, heart rate, temperature. These measurements are metadata about our body, and they help the doctor determine our health status. If we are renting a movie, we might seek out the work of a particular actor or director, or perhaps a specific genre: we might feel like a Hitchcock film tonight, or one of Bette Davis's classics. That information about the film helps us match it to our preference.

In essence, then, metadata is anything that describes an object that is not the object itself. Metadata generally comes in two parts. First are the metadata attributes themselves, sometimes referred to collectively as the metadata architecture or the metadata schema, which is the list of individual metadata types that will be used to describe an object or class of objects. A basic metadata architecture might include the following:

- author
- date
- title
- subject
- document type.

Second are the specific values allowed for any particular metadata attribute. Some attributes, such as 'title' may be 'open': that is, no limit is placed on the titles that authors can create. Other attributes, such as 'document type,' might be 'closed,' with a predetermined list of approved values from which authors must choose. Such a list of approved values is a controlled vocabulary.

Importance of metadata

All information management systems need to keep track of metadata to function properly. Your home computer or laptop, for instance, must know the address of a file on your hard drive, so it can find it when you ask that the file be opened. It must also 'know' what kind of file it is so that it can open it with the proper software. Your computer uses the file suffixes for that purpose. It knows (because somewhere along the way, someone told it) that PPT files get opened by PowerPoint, that CDX files get opened by CINDEX and that XLS files get opened by Excel.

Metadata may also be used in search systems to improve recall and precision. Recall is the number of relevant documents retrieved compared with the total number of relevant documents. Think of recall as measuring 'How much of the good stuff did I get?' Precision is the number of relevant documents retrieved compared with the total number of documents retrieved. Think of precision as 'How much of what I got is good stuff?'

Metadata in content management systems

Although it was not explicitly stated above in the description of CMSs, such systems need metadata to operate. During the authoring stage, the CMS assigns each document a unique ID so that it is not confused with any other document. The CMS also usually tracks a document's creation date, version number, and title. For workflow purposes, the CMS needs to know the author and document type as well. In the document library, the CMS must know the location of the document in storage to be able to retrieve it for revisions or approval. For publishing, the CMS must know the subject of a piece of content, so the CMS can post the content in the appropriate places on a website, for example. The metadata can also be used to create a browsing hierarchy, leading users to the information they need quickly and efficiently.

Metadata and search

Usually, recall and precision operate in an inverse relationship. That is, it is possible to improve the total number of relevant results, but only by vastly increasing the total number of documents retrieved, thus lowering precision. However, document subject tagging makes it possible to improve both recall and precision at once. Here's why.

Free text searching only tells us when a document contains a certain word or phrase. It does not let us know whether a document is 'about' the concept that those words represent. When metatagging is done properly, users are assured that the document actually discusses the topic they are searching for.

Properly tagged content also allows users great precision in finding the information they are looking for. Let's say I want the contract that Josie prepared last week for Company Z. A good content management system in this context will have been set up with document type, author, date, and vendor fields on which I can search. So all I have to do is enter the appropriate terms in each relevant field, and I get the exact document I want.

Metadata also helps prevents concept scattering. Studies have shown that few people use the same words to identify the same concept. In fact, the most frequently used term may be used by only 15 percent of the people searching for that concept. Using consistent content tagging as well as variant terms (synonym rings) in the search system helps overcome the problem of scattering. For example, a good search system will know that the terms 'cats' and 'felines' are equivalent, so no matter which of the two terms a searcher enters, all relevant documents are returned.

Metadata can also help distinguish between homonyms. A search system might include at the top of its results set something like: 'Do you mean "pitch" as in tar or "pitch" as in baseball?' That allows searchers to increase precision by narrowing the search to only the meaning of the word they want.

Developing metadata for a content management system

Background research

Before actually developing a controlled vocabulary (CV) or metadata schema, it is important to understand the context in which it will operate. That context has three major components: (1) the business rationale, (2) the content, and (3) the users. We look at each of these briefly in turn.

Business rationale: What is the purpose of the system that is using the metadata or controlled vocabulary? Is it to improve employee productivity? Is it to sell products? Is it to enhance the company's image? Each of these roles affects how the metadata/CV is developed.

Content: How many documents does the system include: hundreds or millions? Is the document collection growing rapidly? Is it a changing collection of documents or is it relatively static?

Users: Who will be using the system? Experts in the field or laypeople? Employees only? Suppliers? Each user segment will have different vocabulary needs.

There are three basic techniques for identifying important information about the three contextual components. For the business rationale, it is customary to interview opinion leaders or project sponsors to develop a good picture of the business environment for the system. Usually an interview of 45 minutes to an hour is sufficient to get the information necessary. It is also a good idea to interview between 8 and 10 individuals to ensure a range of perspectives and to make it possible to begin to identify trends.

Content audits or content analyses can be used to discover important facts about the content itself. Audits identify each piece of current content and analyze its structure, format, and content. Analyses use a representative sample of content to identify important content types, as well as other general information about the content.

Finally user research, including user interviews and user testing, gives information about what actual users are looking for, their information-finding techniques, and whether they can use the current system or not.

Metadata schema development

Once the business context, the content, and the users are known, it is possible to identify the metadata attributes that will be necessary for the system to function well. Almost all systems need basic data, such as author, title, creation date, and subject. A good place for finding such basic metadata attributes is the Dublin Core Metadata Element Set, one of the basic international standards in metadata. You can find information about Dublin Core on the Web at: `http://dublincore.org/documents/dcmi-terms/`

It is also necessary to consider additional metadata attributes that respond to the needs of the particular system. These might include such attributes as 'customer segment,' or 'product,' or 'internal service.' After an initial attribute schema has been developed, it is necessary to validate it internally within the company, by having appropriate individuals review and comment on the schema. Then comes the time to start developing any necessary controlled vocabularies.

Controlled vocabulary development

First, candidate terms should be captured. This can be done by reviewing actual content (this is where indexing skills come to the fore), as well as through a review of search logs containing terms that users actually input into the system. After the terms have been gathered, it is necessary to start identifying synonyms, selecting preferred and variant terms, creating term hierarchies (broader and narrower terms), and identifying related terms.

To ensure the users understand the concepts and terms selected, user testing should be performed using a variety of techniques. It will be necessary to go through several, often many, iterations of vocabulary development and validation before arriving at the CV's final form. Finally, the CVs need to be implemented in the particular software used for the system under development.

Conclusion

This overview only touches the surface of what metadata is and how it is created and used. More detailed information and additional experience will be needed by those who wish to hone their skills for use in metadata and controlled vocabulary development. One important resource is the online magazine BoxesandArrows.com. Developed originally for the information architecture community, this e-journal offers an ever-expanding selection of articles on a variety of topics, including user testing, usability, content management, and metadata and controlled vocabularies.

In the interest of full disclosure, I should note that I have a vested interest in this resource. Two colleagues, Karl Fast and Mike Steckel, and I have been writing a series of articles for BoxesandArrows on the subject of controlled vocabulary development. The first article in the series is available at:

```
http://www.boxesandarrows.com/archives/
what_is_a_controlled_vocabulary.php
```

This is a good place to start to find out more about the world of metadata. You might also want to look at the thesaurus of controlled vocabulary terms we developed:

```
http://www.boxesandarrows.com/archives/
controlled_vocabularies_a_glossothesaurus.php.
```

Fred Leise is owner and principal of ContextualAnalysis, LLC, providing consulting services in metadata and controlled vocabulary development, user testing, information architecture, and website indexing. Since 1995, he has worked as a freelance indexer specializing in scholarly works in the humanities. He currently serves on the board of directors of the American Society of Indexers, and regularly presents workshops on indexing and controlled vocabularies. In indexing circles, he is well known for his First Rule of Indexing: 'There are no rules. There are only contexts.'