



INFORMATION ARCHITECTURE ON THE BACK END

By Ed Stevenson and Lisa Bos, Really Strategies Inc.

When developing a taxonomy for your Web site content, it's important to think about how it will be used on the back end as well as from the front end.

Discussions about information architecture typically focus on the organization of Web site content. But this is too narrow a focus for publishers and other organizations dealing with large quantities of content. For publishers, the delivery of content to a Web site is usually highly automated, and there's no way to support the site information architecture unless the content itself is intentionally designed to do so.

Therefore, complete information architecture must extend into the back end, meaning the systems used to create and manage content. Publishers can take the well-known concepts of Web site information architecture – primarily the intelligent labeling and organization of content – and apply them to back-end process and content management. This approach helps to create an information architecture stream from content creation through management and delivery to the customer.

Content Types

It all starts with content types – defining what the content “is.” Defining content types allows for easier identification and retrieval and specific processing and management of content. The process begins with a broad survey of the entire content collection, identification of types and organization of the content into those types. The process needs to consider the following:

Avoid context: One of the biggest challenges is to type content outside the context of publications or products. Ask yourself, “What is this piece of content?” apart from the publication in which it first appears. Note that sometimes the content label differs from internal management to the presentation to the customer; this can be accounted for when content is processed to the site.

Other metadata: Content labels shouldn't include information that is already captured through other metadata or that can be easily inferred. For example, “News” is a better label than

“Today's News” as there should be some time stamp that stores the date.

Structure: Because of structure variations, content types may need their own DTDs, schemas or templates. For example, letters to the editor have a different structure than articles or tables.

Instinct: A good gut-level feel for the content certainly helps answer questions such as, “What is this content and why should it be distinguished from other content?” and “Is it the same as content found in other products?”

Organizing Content

The information used to organize and describe content is usually referred to as “metadata.”

Identifying content types as previously described is the first and most fundamental step in defining metadata, as the remaining metadata often vary by content type. For example, images may need a caption (a specific piece of metadata) but other content types may not. On the other hand, some fields are global in

application, such as unique identifiers. It is especially important to model global metadata correctly as they greatly facilitate search across all objects, as well as re-use and re-purposing.

Internal processes may require their own specific metadata. For example, the notion of image or document collections might not have meaning outside the company but is helpful for internal organization. On the other hand, an image thumbnail might need to be described as metadata for external users but might simply be a built-in feature of a digital assessment management system.

Some metadata can be programmatically assigned at specific points in a workflow and some must be manually assigned or reviewed.

Relationships among objects can be difficult to model correctly in a way that appropriately enables re-use/repurposing. Relationships can also be used to determine metadata by inheritance, which can also lead to complex discussions of workflow.

It is not always easy to determine whether a metadata field should be optional or required. Careful analysis in this area is essential given that inconsistently assigned metadata can render it useless in product applications. For example, a taxonomy is only useful if all of the content a user is searching has been tagged to the taxonomy.

Over time, metadata is often abused. That is, fields are used for purposes other than they were originally intended, or required fields are populated with bogus values. Such abuse leads to inconsistency among content sets and undesirable complexity in implementing internal and external systems that exploit metadata. For this reason it is essential to:

- Document metadata thoroughly and in a way that is understood by internal users.
- Educate staff during the analysis process about why it is important to keep metadata pure over time.
- Consider the process for extending the model from the beginning.

Obviously others have gone through the thought process of determining applicable metadata for their content. Some of these thoughts have been captured in industry standards.

In their full form, these standards are sometimes more appropriate for sharing content among organizations than for use within a single organization.

However, even when not implemented internally, standards make terrific starting points for information analysis.

The metadata offered in standards are not exhaustive of everyone's needs and at the same time include unnecessary elements for some. That is okay; most metadata standards allow elements to be optional and allow for additional elements to be included. Metadata standards include Dublin Core, The Publishing Requirements for Industry Standard Metadata (PRISM) and the IMS Learning Resource Meta-data Information Model.

Content producers and production staff need internal labels that make sense to them, which may be quite different than the labels that make sense to a customer. Something referred to as a "teaser" or "first graph" by editors may become the "abstract" or "overview" to

needed in their creation and management. This is because, even if they are not displayed directly to end users, they will often be used in automated processing (e.g. "Publish only documents with list values A, B and C.").

Sometimes these lists are not flat; instead they have multiple levels. Such controlled vocabularies are typically called "taxonomies" and often have other complex characteristics as well. For example, two nodes or terms in a taxonomy might have a relationship to each other – a similarity that can be helpful in guiding users to related content. For publishers, taxonomies are typically used to describe subject categorizations and can be one of the most important aspects of information architecture. Taxonomies can be expensive to design, implement and apply to content and require great care to ensure appropriate pay off.

When defining a taxonomy, the first step should be to research standards or other taxonomies. Often a publisher can find an existing taxonomy that works for

IT'S IMPORTANT TO ALLOW FOR THIS VARIATION BETWEEN FRONT- AND BACK-ENDS LABELS, WHILE KEEPING TRACK OF HOW THEY MAP TO EACH OTHER.

the customer. The content may even receive different labels dependent on the context of its presentation. That same teaser may be an "alert" in an e-mail newsletter, an "overview" in a table of contents and a "summary" in a search results list. It's important to allow for this variation between front- and back-end labels, while keeping track of how they map to each other.

Vocabularies and Taxonomies

Many metadata fields must be limited to a controlled set of terms. Sometimes this is a short list. Sometimes it's a list that can grow to hundreds of terms, with different term lists for different internal user groups or content types. In either case, these can be described as "controlled vocabularies," and great care is

its content or at the very least provides a good starting point.

The next step is to determine the top-level categories. For a publisher of pet information, the top levels might be "dogs," "cats," "birds," "fish," "reptiles/amphibians" and "small mammals."

Moving from the top level down, the publisher needs to determine exactly how far within each category to go. Subcategories under "dogs" could include "small," "medium" or "large." Or the taxonomy might list out specific breeds. There are no wrong answers; the right approach is the one that makes sense for the content in light of its internal management, customers and the content itself.

During the development process, publishers should not worry too much

whether choices are exactly right. It is easy to get bogged down in the details and to halt progress while agonizing over minutia. The objective is to develop a starting point and to improve it. Even when the final taxonomy – or least one ready for implementation – finally evolves, it will not be perfect. None are. Not everyone will be happy with every aspect of the taxonomy. It is important for publishers to realize and accept this fact.

The next step is to test it out. One of the best ways to do this is to see how the taxonomy works with real content by categorizing representative documents and images. Does all the content fit into the categories? Were some levels of the taxonomy filled more heavily than others? Expanding on the pets example, the publisher might find that “reptiles” and “amphibians” should be split into their own categories because there is so much content in each. Perhaps another category is needed to capture less-popular or unique pets, such as horses, pigs or insects. Does the top-level term “fish”

work, or is a broader term needed to include animals kept in aquariums that are not technically fish, such as anemones or hermit crabs?

Additionally, note that the taxonomy developed for internal management does not have to exactly match the taxonomy exposed to the Web site visitors, although it does need to map in some way. There are reasons why your internal taxonomy may morph between internal management and front-end use, such as:

- To present different labels for the same concept to customers. Editors may prefer “canine,” but customers may be more comfortable with “dogs.”
- To limit the level of granularity to the customer. Perhaps internal processes require a subdivision into breeds, but that subdivision is not needed on a specific online product.
- To modify entry points into the taxonomy or adjust certain levels for presentation to the customer. Internally, the top levels may be “mammals,” “fish,” “reptiles” and “amphibians,” but for a

site navigation, it is changed to “dogs,” “cats” and “other mammals.”

These are only a few examples, but the point is that the internal taxonomy may differ than what is presented to the end user.

We can learn from front-end information architecture methods and are better off when front-end and back-end information architecture are considered as a continuous stream. For the most part the same principles and techniques apply to both, and both must be tied together for success.

Ed Stevenson is director of consulting services and Lisa Bos is executive vice president and chief architect for Really Strategies, which provides content solutions and services to publishers and other content-centric companies. From content creation to delivery, Really Strategies helps bring strategy, content and technology together to analyze, architect and implement appropriate tools and technologies. For more information, visit www.reallysi.com.

Half page ad
Veleo IP