



## Information Architecture for Publishers

Produced by Really Strategies, Inc.

618 S. Broad Street

Lansdale, PA 19446

January 2005

[www.reallysi.com](http://www.reallysi.com)

# Contents

<b>Information Architecture for Publishers.....</b>	<b>1</b>
Information Architecture for Your Web Site .....	1
Why is information architecture important for your customer?.....	1
Inventorying and organizing content.....	1
Labels and navigation.....	2
Making the Connection.....	3
Factoring in the Back End.....	3
Inventorying content (identifying content types).....	3
Organizing content (metadata) .....	4
Standards for content organization.....	5
Normalizing terminology.....	6
Developing controlled vocabularies and taxonomies.....	6
Developing a taxonomy .....	6
Final Thoughts.....	7
Additional References.....	8
<b>About Really Strategies .....</b>	<b>9</b>

## Information Architecture for Publishers

Discussions about “information architecture” — a popular buzz word for several years — typically focuses on the organization of web site content. But this is too narrow a focus for publishers. For publishers, the delivery of content to a web site is usually highly automated, and there’s no way to support the web site information architecture unless the content itself is intentionally designed to do so.

Therefore, a complete information architecture must extend into the publisher’s back end — the systems used to create content. And, as it turns out, well-designed content benefits internal systems and processes just as it does front-end delivery systems like web sites.

This paper considers information architecture for publishers: the front end, the back end, and the connections between the two.

### Information Architecture for Your Web Site

Front-end information architecture refers to how you present your content to your customer, such as visitors to your web site. Publishers know that this is not a new topic; they’ve been worrying about print presentation for hundreds of years. Of course, online presentation has additional challenges due to its interactive qualities.

#### Why is information architecture important for your customer?

Without sensible organization, labeling, and navigation, a web site can have the most informative content in the world but in all likelihood your customers will have a very hard time finding and using it. The Web has been compared to the largest library in the world where all of the books have been pulled off of the shelves and thrown into a pile in the middle of the floor. The information is there, but it’s the user’s job to find it.

Without well-designed information architecture, your site — and your content — is just one pile among many. With good information architecture, your content is back on the shelf, properly labeled, and searchable using the online card catalog, for which your users will thank you.

#### Inventorying and organizing content

When designing the information architecture for a web site, one of the first steps should be performing an inventory of the site content. It’s important to know the universe of the types of content that you need to deal with as opposed to knowing how much content you have in total. For example, a typical journal publisher’s site will include content types such as articles, editorials, letters to the editor, and book reviews. The key in the beginning is to know *what kinds* as opposed to how much.

Once the content has been inventoried, the information architecture team can consider how the content should be organized. Depending on what kind of content was identified by the inventory, the difficulty of this process could range from very easy to very hard. If your content is very homogenous, it will probably be fairly easy to organize; given a large set of diverse types of information, it can be hard to think about how best to organize it. One way to start is by considering five major modes that can be used to organize information: location, alphabet, time, category, and hierarchy<sup>1</sup>, also known by the acronym LATCH. An explanation of each of the five modes may help you understand how they may relate to your particular content.

- **Location** is a useful way of describing information that come from or is related to different sources or places. For example, a listing of training courses might be organized by state, country, or other geographic region.
- Organizing information **alphabetically** is useful for presenting large amounts of infor-

<sup>1</sup> Wurman, Richard Saul, *Information Anxiety 2*, Indianapolis, IN: Que, 2001.

mation. The directory of courses could be organized alphabetically by name, for example.

- **Time** can be used to organize content that is published on a regular schedule such as annual reports or the archives of a publication. It works well with calendars, schedules, or other event-driven content. Obviously, content on news-oriented sites or site areas is highly time sensitive.
- Organizing by **category** can be very useful, provided the categories are well-understood and consistent in their organization and their application to content. Categories tend to organize content by type, purpose, topic, issue, and so on. Categorical organization is especially important to publishers, whose audiences often come to their sites for highly topical content. Detailed categorization schemes are often referred to as “taxonomies” or “ontologies” (see Page 6 for more information on developing a taxonomy).
- **Hierarchy** refers to organizing content by importance, rank, magnitude, or some other similar measure. For example, in a related article list, articles may be arranged by the most to least relevant.

Since you'll likely have different types of content to manage, you may need to use several different organizational models throughout your site. Also, it might make sense to organize some content in more than one way, allowing visitors to sort articles both by date and by topic, for example. These decisions affect the complexity of the design of the web site, so it's better to consider them as early in the process as possible so you don't have to re-engineer the site to accommodate the unanticipated needs of your customers that you discover post-launch.

Which brings up an important point: You may not know how best to organize your content to serve your audience, so it might pay to ask them. Organizations sometimes have a myopic view of their content. They may only think about their web site within the framework of the organization

that produces it. For example, many publishers naturally organize their web sites based on their traditional print publication sections. The odds are pretty good, however, that your visitors will want to explore content in new ways that weren't possible in print, and that span multiple print products. So while it may seem natural to organize the site that way, it might actually frustrate your customers, particularly those who don't read your printed publications. One of an information architect's jobs is to learn how your customers would like to interact with your content by conducting surveys and other studies to gain a better understanding of how customers might view and use your content.

### Labels and navigation

Attending to the information architecture of your web site also entails paying attention to how content is labeled and described. Some of the labels that are worthy of careful consideration on any web site include headings, titles, navigation elements, and index terms. These labels are important because they are the tools your users will rely on to build a mental map of how your site is organized, where they currently are located within the site map, and how to find the next piece of information they want.

Headings and titles are important pieces of information for understanding the content that comes after them. In general, they are the sign posts that guide a user through a document. Proper use of titles and headings can break large amounts of information into much more digestible parts.

Navigation labels are a key component of any site's information architecture because they are often the primary reflection of the overall organization of the site. If the labels for your navigation are clear, users will have a much easier time finding their way around than if the labels are ambiguous or vague. The problem is, what is perfectly clear to one person may be completely inscrutable to another, so long before the navigational labels of a site are cast in stone (or Photoshop mockups), it is best to have vetted the labels

through user testing or other empirical methods to be sure they make sense to your target audience.

Finally, index terms are the labels that are used to describe individual pieces of content on a site. Most people are familiar with index terms as the keywords used by some search engines to determine the relevance of a particular piece of content to a particular topic. In some situations, the information architecture team may be responsible for defining a controlled vocabulary of index terms that are used to describe content on the site. They may also be responsible for maintaining a thesaurus of related terms.

Once again, it is important to consider the target audience when assigning index terms to a piece of content. Often, the terms used to describe something will vary depending on who you ask. For example, if content originally developed for a professional market is displayed on a consumer web site, the professional index terms might not be very helpful (“bilirubin” may make sense to your pediatrician audience, but the parent visiting your site would probably find “jaundice” more useful).

## Making the Connection

Designing your site organization, labeling, and navigation is certainly a challenge, but actually getting content to your site to support that architecture can be an even greater one. Most publishers start by manually preparing content for their site, even manually tagging documents as HTML. This work — adding HTML meta tags for categories, navigational headings and links, and so on — is laborious, especially if your information architecture has significant detail and depth. Slowly, publishers add steps to ease the process, such as scripts for a basic conversion of Quark XPress files, or a simple database for entering HTML meta tags. Eventually publishers realize that they need a more comprehensive approach with better automation so that content flows seamlessly from their editorial or production environment and onto the web. “Content

management” becomes a hot topic around the editorial water cooler.

But regardless of the content management technology you might use to support your site, the first step in creating this smooth connection is extending your information architecture to the back end. How can you automate or semi-automate the categorization of documents into site sections unless you capture that information somewhere during your editorial or production processes? How can you make sure that the dates you need for sorting are captured unless you define exactly what those dates mean and where they originate? Those are the kinds of questions you answer when you complete your information architecture for your back-end environment.

## Factoring in the Back End

What is “back-end information architecture?” Just as you need to organize and categorize content for your customers, you need to do the same for your internal processes and systems. You need to:

- Identify your content within the context of editorial and production environments. (Inventory your content.)
- Develop metadata that allows for internal control as well as external delivery. (Organize your content.)
- Normalize terminology. (Define labels.)
- Develop controlled vocabularies and taxonomies. (Refine those navigation paths that are not web site-specific.)

This probably sounds a lot like what we described for front-end information architecture, and in many ways information architecture for the back end *is* similar to what you do on the front. But, there are also important distinctions.

## Inventorying content (identifying content types)

Identifying your content types is just as important on the back end as on the front end. It

allows for easier identification and retrieval and specific processing and management.

The process begins with a broad survey of all the content needed to produce your final products. This might include content that is never published outside your organization but that is critical to internal operations, such as licensing information or original manuscripts. It will also include the identification of types that you probably didn't think to break out for the front end, such as images.

The steps for refining your list of content types is the same for the back end as for the front end, but it becomes even more important to remove product bias — online or print — from the set of content types. It's okay if the name of a content type is different in your back end than on a specific web site; this can be accounted for when content is processed to the site. Also, you might find that you need multiple types on the back end where you only expose a single combined type to your customers. For example, you might have workflow reasons to differentiate between specialized types of journal articles, but no reason to differentiate among them for your users. That's okay too, as long as there are legitimate business reasons for doing it. However, your back-end types must be at least as detailed or "granular" as your front-end types — you don't want to have to add complexity when delivering to the web.

The following are considerations when determining content types:

- **Remove from context:** One of the biggest challenges for a publisher is to type their content outside the context of publications and products. A publisher should ask "what is this piece of content?" apart from the publication in which it first appears.
- **Other metadata:** Don't label content types with information that is already captured through other metadata, or that can be easily inferred. For example, use "News" as opposed to "Today's News." There should be some time stamp metadata that gives the date

or timeframe. Don't use product names in content types. You'll capture the source of the content and the products it is used in elsewhere and can add the product name if necessary in delivery.

- **Structure:** Structure also plays an important role in content typing. Letters to the editor have a unique structure different than articles or tables. Content types may need their own DTDs, schemas, or templates.
- **Instinct:** The last aspect of content typing is using your gut level feel for the content. What is this content and why should you distinguish it from your other content? Is it the same as content found in one of your other products?

Content typing takes the points above and looks to find the appropriate balance between them. It is a difficult task that has no right or wrong answer. The process can often be mired in departmental politics and semantic debates, but it's worth slogging through to ensure clear communication and presentation to your customer.

### Organizing content (metadata)

The information used to organize and describe content is usually referred to as "metadata". Identifying your content types as described above is the first and most fundamental step in defining metadata. That's because the remaining metadata often varies by content type. For example, you might need a "caption" metadata for images, but not for any other content types. On the other hand, some fields are global in application, such as a unique identifier. It is especially important to model global metadata correctly as it greatly facilitates search across all objects and as well as their re-use and re-purposing (note that content type is a globally applied metadata field).

The following observations about metadata are useful in the development of a back-end information architecture.

1. Some metadata is specific to internal needs (e.g., version information) and some to exter-

- nal needs. 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 an external digital assessment management system.
2. Some metadata has different uses or modeling requirements internally (e.g., for internal re-use or workflow routing) and externally (when presented to customers). For example, internal and external content navigation methods might vary and therefore require different taxonomies or categorical metadata. Full digital object identifiers (DOIs) are sometimes assigned on export from a content management system rather than stored.
  3. It is not always easy to determine whether a metadata field should be optional or required, and this may be dependent on the lifecycle (workflow) stage of the object or on its type. In addition, different groups often have different opinions about what should be required. Careful analysis in this area is essential given that inconsistently assigned metadata can render is useless in product applications (e.g., a taxonomy is only useful if all of the content a user is searching has been tagged to the taxonomy).
  4. Some metadata can be programmatically assigned as specific points in a workflow (e.g., when loaded into a system) and some must be manually assigned or at least manually reviewed. It's important to identify these differences and consider them when determining whether a new metadata is worth including or requiring.
  5. Relationships among objects and especially among object types 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 (e.g., should an image inherit a related article's taxonomy assignments?), which can also lead to complex discussions of workflow.
  6. Some metadata should be globally accessible to all internal users; some should be available only to those with appropriate permissions.
  7. Over time, metadata is often "abused": 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 the metadata thoroughly and in a way that is understood by internal users (e.g., use their language and a lot of examples), to begin the educational process about why it is important to keep metadata "pure" over time during the analysis process, and to consider the process for extending the model from the beginning.
  8. Metadata can be modeled as XML (DTDs or schemas) or relationally, or both. It is important to agree on the modeling (and therefore documentation) approach at the beginning of the modeling effort.

#### Standards for content organization

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. The standards typically include both a metadata model (the set of metadata elements) and usage instructions (how the information should be encoded — e.g., as XML). In their full form, these standards are sometimes more appropriate for sharing content among organizations than for use within a publisher. However, even if you choose not to implement a standard internally, they make terrific starting points for information analysis. Standards will help you consider metadata elements you may not have thought of, and categorize them into useful groupings. When considering any standard, you will find that some of the included elements do not apply to your content, and elements you need will be missing. That is okay — most metadata standards allow

elements to be optional and allow for your own elements to be included within your specific environments. Examples of standards to consider are:

- **Dublin Core** — Dublin Core defines a set of fifteen basic, widely applicable metadata elements.
- **PRISM** — The Publishing Requirements for Industry Standard Metadata (PRISM) specification defines an XML metadata vocabulary for publishers. The elements from Dublin Core are included as a subset within PRISM. PRISM is somewhat print focused, but still has a great starter set of basic metadata. It uses an RDF syntax. (RDF is yet another standard for metadata, but for its structure, not for specific elements.) PRISM is an IDE-Alliance specification.
- **IMS** — The IMS Learning Resource Metadata Information Model identifies a subset of IEEE learning object metadata (another standard!). It is useful as a starting point for educational content information architecture.

### Normalizing terminology

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 your user. Something referred to as a “teaser” or “first graph” by editors may become the “abstract” or “overview” to 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 email 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-ends labels, while keeping track of how they map to each other.

That said, information analysis of the back end often calls into question labels that are of questionable value even internally. Maybe they are print-oriented and should be generalized to more clearly describe the many ways in which the content is used on output. Maybe some have

been in use so long that you can’t even remember how they came into being.

### Developing controlled vocabularies and taxonomies

Many metadata fields must be limited to a controlled set of terms. Sometimes this is a very 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 should be taken in their creation and management. This is because, even if they are not displayed directly to your end users, they will often be used as a decision point in automated processing (e.g., by setting up rules such as “publish only documents with list values A, B, and C”).

Sometimes these lists are not “flat”; instead they are hierarchical, meaning they have multiple levels (or tree branches). Such controlled vocabularies are typically called “taxonomies”, and often have other complex characteristics as well. For example, two “nodes” (terms) in a taxonomy might have a relationship to each other — a similarity that can be helpful in guiding your users to related content. For publishers, taxonomies are typically used to describe subject categorizations, and can be one of the most important aspects of your information architecture. This is because a well-designed taxonomy can be a significant differentiator in the usefulness of searching and browsing your web site. Taxonomies can be expensive to design, to implement, and to apply to content, and require great care to ensure appropriate pay off.

### Developing a taxonomy

When defining a taxonomy (or taxonomies) for your content, the first step should be to research standards or other taxonomies used in your field. Perhaps an existing taxonomy will work for your content or at the very least give you a start.

Next determine your top level categories. If you publish information on pets, your top levels might be “dogs”, “cats”, “birds”, “fish”, “reptiles/amphibians”, and “small mammals”.

Take each of those and determine how far within each category you need to go. Under “dogs”, you might have “small”, “medium”, or “large”. Or you might list out specific breeds. There are no wrong answers; the right approach is the one that makes sense for your content in light of its internal management, your customers, and the content itself.

During this process don’t worry too much whether your 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 strawman starting point, and to improve it over time. Even when you bless a final version that you’re ready to implement, your taxonomy will not be perfect. None are. Not everyone will be happy with every aspect of the taxonomy. It is important for everyone to realize and accept this fact.

The next step is to test it out. Check how the strawman taxonomy works with real content by categorizing representative documents and images. Does all the content fit into the categories you created? Were some levels of the taxonomy filled more heavily than others? Expanding on our pets example, you might find that “reptiles” and “amphibian” should be split into their own categories because you have so much content on each. Perhaps you need another category to capture less-popular or unique pets, such as horses, pigs, or insects. Does the top level term “fish” work, or do you need a broader term to include animals kept in aquariums that are not technically fish, such as anemones, hermit crabs, or starfish?

Additionally, note that the taxonomy you develop for internal management does not have to exactly match the taxonomy you expose to your site visitors, although it does need to map in some way. Below are several reasons why your internal taxonomy may morph in some way between internal management and front-end use:

- You may need to present different labels for the same concept to front-end users (internally you may be more comfortable with “canine” but present the label “dogs” to the customer).
- You may limit the level of granularity to the customer (perhaps, for some internal management reason, you decide to drill down into subdivisions of breeds, but that subdivision is not needed to organize the content on a specific online product).
- You might decide to modify entry points into the taxonomy or adjust certain levels for presentation to the customer. Perhaps for internal organization, the top levels are “mammals”, “fish”, “reptiles”, and “amphibians”, but for the site navigation, you change the top level to be “dogs”, “cats”, and “other mammals”.

These are only a few examples, but the point is that the taxonomy you develop for internal organization may differ than what you present to the end user. That happens often, and you need to be able to map from the internal taxonomy to the external one (for example, your external taxonomy cannot have deeper granularity than your internal one).

## Final Thoughts

When you approach the concepts outlined in this document, do it with a team representing different interests in the organization. Bring in editorial, production, IT, and business staff to join and take responsibility for the process. And of course don’t forget the customer. Make sure someone with customer-focused views participates, or better yet involve some customers in the process.

The following summarize some of the key concepts expressed throughout this document, especially those that connect front-end and back end-information architecture:

- Consider front-end and back-end information architecture as a continuous stream. For the most part the same principles and techniques apply to both and both must be considered for fully successful information architecture analysis.
- Your labels and presentation of information architecture concepts can vary for internal and external uses. The important thing is to be able to map the two.
- Never forget the end users, including both the editorial and production staff and the ultimate consumer of your content.
- Establish processes for maintaining the front-end and back-end information architecture over time. Misuse will destroy the good work you put into it.

### Additional References

The following resources may help you in your approach to information architecture:

- For more information on referenced standards, see [www.dublincore.org](http://www.dublincore.org) (Dublin Core),

[www.prismstandard.org](http://www.prismstandard.org) (PRISM), and [www.imsglobal.org](http://www.imsglobal.org) (IMS).

- *Information Architecture for the World Wide Web: Designing Large-Scale Web Sites*, 2nd Edition, by Louis Rosenfeld and Peter Morville is an excellent book to read to gain an understanding of how information architecture fits in to the design and development of a web site.
- *Information Anxiety 2*, by Richard Saul Wurman, while not strictly about information architecture, is full of great ideas about how to effectively turn the large amounts of data that surrounds us into information that can be used and understood effectively.
- The IA Wiki is a good place to read more about front-end information architecture and even add your own thoughts and opinions about the subject.
- The web site [www.boxesandarrows.com](http://www.boxesandarrows.com) is a good source for articles and views on information architecture.

## About Really Strategies

Really Strategies, Inc. is a privately held company that was founded in 2000 to provide world-class content solutions and services to publishers, media companies, 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. Our solutions encompass XML editorial tools, XML repositories, content management systems, and editorial and production systems. Our services include workflow reengineering; technology evaluation; DTD and Schema development; business, functional, and technical requirements development; and electronic product development strategy. As a recent recipient of the Deloitte and Touche Rising Star Award (Delaware Valley), the Philadelphia 100® Award, and one of the "Best Places to Work" in Philadelphia Award by the Philadelphia Business Journal, Really Strategies is committed to building the premier content solutions and services firm.

### Contact Information

Email: [info@reallysi.com](mailto:info@reallysi.com)

Telephone: (215) 631-3107

Mailing Address: 618 Broad Street, Lansdale, PA 19446

Barry Bealer, President/CEO

[bbealer@reallysi.com](mailto:bbealer@reallysi.com)

Lisa Bos, Executive Vice President, Chief Architect

[lbos@reallysi.com](mailto:lbos@reallysi.com)

Marcelle Soviero, Vice President, Business Development

[msoviero@reallysi.com](mailto:msoviero@reallysi.com)