

Navigating the Long Tail

by James Kalbach

James Kalbach is a human factors engineer at LexisNexis and is the author of the recent book *Designing Web Navigation* (O'Reilly). For more information about Jim and to view his blog "Experiencing Information," see <http://experiencinginformation.wordpress.com/about/>. Contact Jim by email at jim.kalbach@gmail.com

The *long tail* is a popular term used to describe power laws and statistical distributions such as Zipf's law or Pareto distributions. On a graph, the long tail looks like an inverted, upside-down "L" – there is a tall neck to the left and a long, thin tail extending to the right. It reflects the fact that in many situations, a few frequently occurring events generate most of the activity. For example, the most popular music albums represent only a small portion of all albums produced, even though they account for the bulk of the sales. Similarly when analyzing search logs we see that a small number of frequently used terms account for most searches, while a very large number of other terms account for the rest.

The long tail also has economic implications. Offline, businesses are constrained by physical limitations. There's limited shelf space in a store or limited screenings of movies in a theatre. Consequently, those business owners are financially compelled to focus on the most popular products.

But online, there is essentially no limit to what can be offered. With digital goods like music downloads or even information, there is infinite shelf space. And including more information is virtually free: you just add it to the database.

In his book *The Long Tail* [1], Chris Anderson offers the most detailed discussion of these implications on businesses. He observes that "the hits now compete with an infinite number of *niche markets*," or the smaller markets down in the long tail. Online businesses are also finding that the more they add the more people buy in those niche

markets. So, there's not only infinite shelf space but also infinite demand.

Noise

But simply adding more information isn't virtually free. It comes with a cost: noise. The long tail – with all of its special interest markets – is a really noisy place.

If this new online, long-tail economy is to work, people have to be able to navigate to the markets that interest them and filter the information quickly and efficiently. This is really the value of information architecture (IA). IA not only helps people find the information they need, but it also helps them makes sense of it by providing context.

The Navigation Layer

Chris Anderson also discusses what he calls the "navigation layer." This concept isn't new. We've known for a long time that by separating metadata from content, we're better able to navigate information. The traditional library card catalogue, for one, is an example of the navigation layer. So, too, is Sabre, the primary networked reservations systems for airlines.

As we add more information to accommodate niche markets in the long tail, and consequently more metadata, the navigation layer becomes more and more critical. In many respects, creating an efficient and effective navigation layer in the long tail is the biggest hurdle to bringing value to users and business.

Sources of Metadata

We can identify three primary types of sources for explicit metadata relevant to the navigation layer:

1. **User-generated metadata** has received a lot attention recently. For this source users apply their own labels and keywords to resources on the web. *Tagging* is a primary example.
2. Another source of metadata is **technically generated metadata**. Text mining and algorithms can do a lot of the work. For instance, *entity extraction* can pick out proper names and key topics in news articles.
3. And finally we have **owner-generated metadata** such as traditional top-down classification. A *controlled vocabulary* is an example of this.

Keep in mind that none of these metadata types is better than the other, and there are advantages and disadvantages to each, summarized in [Table 1](#).

Consider an automatically generated top 10 list on a music download site. A list for the entire site – across all genres and styles of music – would be a mixed bag of items, and it would be essentially useless to anyone. But within a particular style of music, such as Afro-Cuban jazz or Hip Hop, top 10 lists become much more meaningful. As Anderson writes, “Lists make sense only in context, comparing like with like within a category.”

Note that the context doesn’t necessarily have to come from a top-down structure. Consider this example from [Mag.nolia.com](#), a social bookmarking site. The site allows members to create special interest groups, say for web design. Once a group is created, people can see all the tags from just the members of that group. The likelihood of those tags being of use to someone also interested in web design is much higher than a tag cloud of all tags from the site. So the user-generated metadata (tags) are given context by a user-created structure (the group).

Just as there are three types of sources of metadata, there are also three types of structure: user-created, technically created and owner-created structures. The advantages and disadvantages for each are similar to those for the types of metadata in [Table 1](#).

Down with Taxonomy

Do taxonomies and controlled vocabularies have a place in the online digital world?

Many have asked this question, including Clay Shirky in his oft-cited article “Ontology is Overrated.” [2] However, a close reading of this article shows that the author is not against the use of taxonomies wholesale. He writes,

TABLE 1. Advantages and disadvantages of metadata generated by different types of sources

| Source of Metadata | Example | Advantages | Disadvantages |
|-----------------------|-----------------------|---|---|
| User-generated | Tagging | <ul style="list-style-type: none"> • Low entry costs • Low maintenance • Personal, flexible to be usable | <ul style="list-style-type: none"> • Comprehensiveness • Requires <i>incubation time</i>* • Inconsistent terms |
| Technically generated | Entity extraction | <ul style="list-style-type: none"> • Scales up • Inexpensive once in place | <ul style="list-style-type: none"> • Difficult to implement • Can be inaccurate |
| Owner-generated | Controlled vocabulary | <ul style="list-style-type: none"> • Consistent terms • Comprehensive and complete | <ul style="list-style-type: none"> • Rigid and impersonal • Costly to build and maintain |

*Incubation time: the time before an item or a collection accumulates enough tags to make the item findable or to provide enough tags to satisfactorily collocate or differentiate items in a collection.

Providing Context

For metadata to be useful, it needs structure and context, even technically generated metadata. Categories, subdivisions and filters help us make sense of metadata and navigate it effectively.

“Ontological classification works well in some places, of course. . . . So what you want to know, when thinking about how to organize anything, is whether that kind of classification is a good strategy.”

Shirky goes on to explain when ontological classification (as he called it) is a good strategy for organizing information. He speaks of *bounded domains*, which have formal categories, stable entities and clear edges. Participants in bounded domains may have an authoritative source of judgment and be expert users. And there is often coordination of users into a community where a common technical language emerges.

Notice that the definition of a bounded domain approximates that of a niche market. Therefore, as we move down the long tail into niche markets, as Chris Anderson shows is happening right now, we move into bounded domains. And in bounded domains, traditional taxonomy and information architecture become more and more important, because it’s within bounded domains that top-down classifications make most sense.

Taxonomies instruct and guide and web navigation tells a story. The value of IA is to help people understand the information in those niche markets by providing the context they need using these tools. In a sense IA has emerged as a response to a need for structuring information in those niche markets in the long tail.

Second Order Design

IA will increasingly become what could be called a second-order design activity. Instead of manipulating web pages or metadata directly, IA in the long tail is about giving people the tools they need to create their own context. It’s about facilitating great information experiences, if only indirectly.

Combining the different types of metadata with the structure mentioned above gives rise to a matrix, shown in [Table 2](#). We can use this framework to consider new ways to provide context. Where would a thesaurus fit in? Where does the example from [Mag.nolia.com](#) mentioned above fit in? Is it possible to provide technically created structure to technically generated metadata?

TABLE 2. A matrix for considering different ways to provide context

| Source | Structure | | |
|-----------------------|--------------|---------------------|---------------|
| | User-created | Technically created | Owner-created |
| User-generated | ? | ? | ? |
| Technically generated | ? | ? | ? |
| Owner-generated | ? | ? | ? |

The point is that designing navigation for the long tail calls for any and all types of sources of metadata and all types of structure to provide context. It’s not about one or the other, but about what’s right for the situation. In some situations, a traditional taxonomy may be the best thing; in others, tagging works great. A mix is needed, and those practicing IA will have to experts in them all. ■

Resources Mentioned in the Article

- [1] Anderson, C. (2006). *The long tail*. New York: Hyperion.
- [2] Shirky, C. (2005). *Ontology is overrated*. Retrieved October 26, 2007, from www.shirky.com/writings/ontology_overrated.html.