

## **Why Taxonomies Are Doomed**

Sharon Flank, Ph.D.  
DataStrategy Consulting

Nearly every search solution nowadays includes a taxonomy component. We can't find what we're looking for, so we need to pay someone to help organize our office, virtually. Search vendors compete to sell us a solution that requires us to organize our world into a complex hierarchy and fit everything we do into it. These taxonomies are supposed to work like file drawers, so we know just where to look to find what we need. Unfortunately, they are annoying, expensive, and ineffective.

In the long run, search taxonomies will disappear. If a taxonomy is part of your enterprise and reflects what you do every day, then it is natural to use it for finding things. At Ford Motor Company, for example, Volvo and Jaguar are separate divisions, with their own staffs, finances, factories and dealers. So Volvo and Jaguar are natural divisions in a taxonomy, and it is logical for searching to reflect that. All too often, however, taxonomies are an afterthought glued on to solve a bad searching problem. If you are careful, you can avoid taxonomies entirely, or at least minimize the amount of time they suck away from more productive endeavors ... like finding what you need and acting on it.

Search is, of course, not limited to the files in your enterprise. Beyond Web surfing, searching is key for retailers showcasing their products, and others using the Web to find customers. Both enterprises and individuals search for media – photos and video. Users increasingly want to be able to search e-mail, both to manage their business and to avoid lawsuits. This growth in search applications suggests a growing market for optimization solutions, taxonomies among them. Gartner predicts that taxonomies will continue to draw attention, but that they will continue to perplex users and developers.

In this paper we discuss some of the weaknesses of taxonomies, how to get the most out of them if you do use them, and strategies you can expect to replace them in the long run.

### ***Individuality: My taxonomy isn't your taxonomy***

Taxonomies purport to map the way you think, so you can find what you want. Unfortunately, you don't think the way I think, and your colleagues won't necessarily think the way you think either. If your taxonomy needs to support long-term needs in a business or archive, it cannot afford personal idiosyncrasies. But idiosyncrasies are notoriously difficult to avoid.

"For humans, it's obvious how to organize things. But, learning to be explicit so that computers know what it is so we can find it is a tremendous challenge," according to Gartner analyst Rita Knox.

For both consumer and internal-facing applications, it is vital that the categories and vocabulary make sense to the person searching for information. Often the categories are created by a librarian according to principles of information science. Nonetheless, there can be a considerable disconnect between classification done by humans -- or machines -- and the terminology users employ to search for information.

Some would claim that a better taxonomy is the answer. Certainly a bad taxonomy creates additional difficulties, but even good ones have disadvantages, as we will see. And the effort required to generate and maintain the taxonomy is considerable, even if you could establish a standard descriptive language and taxonomy that is vendor neutral and accepted by the industry.

### ***Scale-up: Navigating your own PC file hierarchy isn't so easy***

When you first got your computer, you set up the hard drive so that it reflected your view of the world. Each document was in the right place, and you knew just where to find it. Several hundred files later, things are not so clear. You can't remember where you put things, especially when they seem to fit in more than one category. You use the *Find* feature, hoping that you can remember at least part of the file name. And that's just for documents you created and filed yourself. As the challenge scales to thousands of users and millions of assets, the problems multiply.

### ***Manual taxonomies take time and management***

To do a good job of creating a hierarchy, you need input from people who know your business, know how it will grow and change, and understand how different communities in your organization will try to find things. All of that requires meetings, and meetings require time. Worse, once you have a taxonomy, you aren't finished. As the world changes, you need to update it and maintain it. Does *anthrax* belong under *Weapons of Mass Destruction*, or does it just stay under *Diseases*? *Tomahawk* is a cruise missile as well as a Native American weapon... and a motorcycle and a town in Wisconsin. Not just anyone can be assigned the task of taxonomy creation and maintenance. Specialized – and expensive – talent may be required.

For managing taxonomies across an extended enterprise, such as an e-commerce site, Gartner recommends adapting in-house or commercial taxonomies with a mix of human and machine classification. In addition, they point out that enterprises need to budget for ongoing maintenance of the business vocabularies, dictionaries, glossaries and indexes necessary for useful taxonomies. Many believe that manual taxonomy creation no longer makes sense. As Teragram puts it, manual classification is slow, expensive, inaccurate and unproductive. Various researchers have suggested techniques for automated taxonomy generation; see, for example, Brewster, Ciravegna and Wilks, 2001.

### ***No multiple access points***

Rigid approaches to taxonomies can force you to put items into only one category. These single-access-point systems are frustrating to users because many things fit into multiple categories. Did you put that great photo with the Frisbee into *Vacation pictures*, *Family pictures*, or *Best pictures*? Ideally, of course, you want it to be in all those categories.

But you do not want to have the maintenance nightmare of multiple copies, so your taxonomy had better be able to accommodate multiple classifications.

### ***Does granularity really demand a taxonomy?***

Users claim taxonomies are necessary for government language, medical language, and the special terms in their business. Any decent search engine will allow exact matches on unfamiliar terms, but it is true that special effort will be required to ensure that, say, DARPA matches ARPA, or SARS matches Severe Acute Respiratory Syndrome. A synonym list may well be necessary, but a taxonomy might not. In any event, responding to the granularity issue by building a taxonomy raises business process issues, training issues, and flexibility issues. Who uses which terms? How does everyone learn them? Who is in charge of keeping them up to date?

### ***Global taxonomies***

Taxonomies are neither multilingual nor multicultural. Is an SUV a truck? Is a Mini a sports car? Will all of your customers and suppliers give the same answer? If they do not, a taxonomy may be more of a hindrance than a help. Sometimes terminology use depends on where you are or how you view the world.

### ***Your filing cabinet or a web search?***

Many libraries still have card catalogs in drawers, but the computer-savvy shun them. Instead of looking through drawers and categories, one computer search gives them the book they want. They are shunning the taxonomy approach and opting for search instead. Where do *you* go when you want to find something? If you know exactly where it is in your filing cabinet, it is worth getting out of your chair and retrieving the physical copy. But if you do not know exactly where to find the file in the cabinet, you will be more likely to try a Web search. For example, to find the instructions on changing the battery in your sport watch, the Internet is more convenient than scrambling through drawers.

### ***The reality of your brain***

Brains don't use taxonomies for retrieval. While there is some evidence that our brains contain a hierarchical, structured representation of concepts, we do not seem to find things by first locating them in the hierarchy. If we do not mentally approach search using a taxonomy, then it is probably not a useful metaphor. Keyboards are not neurologically natural either. Both will go the way of the buggy whip eventually. Your brain does categorize, yes, but its approach to retrieval is not top-down. Your taxonomy is not the same as everyone (or anyone) else's. You do not approach mental storage by accessing items at a single access point. Why should the computer system that is designed to help you approach the problem in a way you do not?

### ***ROI***

How do you decide what search technologies are useful in your enterprise? Will a taxonomy save you money?

Though taxonomies can improve employee productivity (less time searching, more time doing their job) and ultimately customer satisfaction (especially if they are working directly with external facing systems), quantifying those gains in terms of ROI can be difficult. Gartner's Knox recommends tying taxonomy projects to specific business goals, such as improving response rate or customer acceptance. Sometimes it is difficult to untangle the ROI of taxonomies from the ROI of simply being better at finding what the user wants, no matter what the technique.

We see claims for the benefits of taxonomies. They will tame searching, and you will see a positive return on your investment. In particular, you are asked to invest in search hierarchies tailored to your business. In theory, you ought to reap returns for making search more efficient, so that your users no longer waste time looking for things, but instead find them using the taxonomy. Knox provided an example of how the lack of a good taxonomy hurts the productivity of people looking for information as well as those serving it. Employees unable to find a certain bit of information on the intranet of their human resources department will end up calling the HR department. In turn, the person who takes the call simply provides the URL to the information and, sure enough, it was there on the intranet all along. It was just impossible to find. A taxonomy can address this problem, assuming that the taxonomy is up to date, the employee's view of the information is the same as HR's view, and the employee knows how to use the taxonomy. All of that coordination and training takes money, and it cuts into the ROI for a taxonomy.

And even if searching improves when the taxonomy is introduced, it's hard to know whether the time you have spent on training wouldn't have yielded even greater benefits if you had just taught users how to search better.

### ***The solution***

So how *do* you improve the ROI of finding what you are looking for, and finding it quickly? You can improve your search engine, your training, or your taxonomy.

### **Improving your search engine**

Good search engines nowadays can be expected to understand phrases, so they do not confuse *Venetian blind* and *blind Venetian* (example courtesy of Prof. George Miller, Princeton University). They may have other enhancements, including broader ability to understand natural language in context. Following Google's lead, they may give extra credence to more authoritative or more frequently accessed sources. Like Vivisimo, they may use clustering to put related examples together.

### **Search training**

One-word queries are notoriously ineffective, and even two-word queries can be frustratingly imprecise. As search engines improve, they are increasingly able to take word order into account, and handle longer queries. And as users improve with them, we

will increasingly be able to find exactly what we want, right away, and use it to do our jobs. That's the goal, after all.

The average search is still only two words long. It may be worthwhile to train your users how to get the most out of searching by using longer searches and taking advantage of special features or library science insights.

### **Taxonomy, if you must**

If you do choose to use a taxonomy, make sure to use a hierarchy with multiple access points, so those Frisbee pictures can be in multiple "bins." Consider relying on process consulting to create an initial hierarchy, so that it accurately reflects your business and how you expect it to grow. Leverage the efforts of others, if you can –

<http://www.taxonomywarehouse.com> is a good place to start. And make sure your taxonomy includes synonyms, so that finding the correct terms is not a guessing game. Finally, keep migration issues clearly in mind – eventually, taxonomies will disappear, and you will want to be able to keep supporting your users' searches, so they can find what they need and do their work.

### ***Bibliography***

#### **Top strategic technologies for 2005**

By Dan Farber

April 22, 2004

Tech Update, ZDNet

[http://techupdate.zdnet.com/techupdate/stories/main/top\\_10\\_technology\\_trends.html](http://techupdate.zdnet.com/techupdate/stories/main/top_10_technology_trends.html)

#### **Taxonomy today, ROI tomorrow**

By David Berlind

April 11, 2004

Tech Update, ZDNet

[http://techupdate.zdnet.com/techupdate/stories/main/Taxonomy\\_today\\_ROI\\_tomorrow.html](http://techupdate.zdnet.com/techupdate/stories/main/Taxonomy_today_ROI_tomorrow.html)

#### **A common language: a cure for the business and IT disconnect?**

By Dan Farber

November 7, 2003

Tech Update, ZDNet

[http://techupdate.zdnet.com/techupdate/stories/main/common\\_language\\_cure.html](http://techupdate.zdnet.com/techupdate/stories/main/common_language_cure.html)

#### **Knowledge Acquisition for Knowledge Management: Position Paper**

Christopher Brewster, Fabio Ciravegna and Yorick Wilks, 2001

<http://www.dcs.shef.ac.uk/~fabio/paperi/ontolearning.pdf>

#### **The Universe of Search: Navigating through the complex enterprise search marketplace**

By Andy Moore

July 1, 2004

KMWorld

<http://www.kmworld.com/publications/magazine/>