Tagging and Why It Matters

David Weinberger
Fellow, Harvard Berkman Center for the Internet and Society
May 13, 2005

Tagging has become the latest craze among the digeratti. While it certainly has been hyped, there are reasons to think it is not only going to go mainstream, it will have effects beyond the realm of mere digital convenience.

The technology couldn't be much simpler. In previous incarnations it was known as "key-wording," the attaching of a simple phrase or two to some digital object such as a document or a photo. Of course, the key words didn't have to be attached in any physical sense; a database could keep track of the object and its associated key words. This logical attachment made it easier for people to find resources. For example, a digitized issues of the Deadwood newspaper from 1885 would not contain the phrase "Old West," but someone scanning in old copies might well use that as one of the key words so that anyone searching for the phrase would find the old articles.

If key wording has been with us from the beginning of the digital era, why is it only now becoming a hot topic? In part it's because there is so much more information available now. But that is not a sufficient reason. Indeed, Google from the beginning has ignored the built-in way HTML pages can contain key words – the "meta" tag – because it's too tempting for authors to increase the popularity of their pages by using bogus key words such as "sex" or "Michael Jackson." Instead, two differences explain the current upsurge of interest in tagging.

First, readers, not just authors, get to tag objects. An author is an authority when it comes to what she intended her work to be about, but not about when it means to others. When it comes to searching, what a work means to the searcher is far more important than the author's intentions. For example, to an author, a Web page may be about natural language processing, but the page may interest one reader because she's writing about adoption curves for new technologies and another because she's researching the social importance of linguistic ambiguity. These readers will tag the page differently from the author and differently from each other because the page means something different to each of them.

Second, tagging is social. For example, at http://del.icio.us, users enter bookmarks (URLs) they want to remember, adding a word or two – tags – so they can sort them later. Del.icio.us users can see not only everyone else's bookmarks, but also all the bookmarks tagged with a particular word. For example, if you care about Emily Dickinson, you can see all the Web pages del.icio.us users have tagged with "Dickinson" or "Emily Dickinson," a great tool for researchers.

These two aspects make tagging highly useful. But there may be another reason why tagging has so excited the technologically-adept early adopters: It sticks it to The Man, especially if The Man happens to be a traditional taxonomist. The idea that to know a field is to see its structure is coextensive with Western history. We have spent an inordinate amount of time encouraging experts and authorities to construct huge

structures of classification, from trees of life to trees of knowledge. This urge to tidy up has shown itself most recently on the Web as the Semantic Web. The tagging movement says, in effect, that we're not going to wait for the experts to deliver a taxonomy from on high. We're just going to build one ourselves. It'll be messy and inelegant and inefficient, but it will be Good Enough. And, most important, it will be *ours*, reflecting our needs and our ways of thinking.

There are about 35,000 users of del.icio.us. Flickr.com, a popular photo sharing site that also features tags, has about 375,000 users. That's a lot of people but a tiny portion of the Net. So, will tagging remain a tool only for a small group of early adopters?

Perhaps, but there are reasons to think otherwise. First, Yahoo bought Flickr.com in March, 2005. As I write this, we don't know what Yahoo is going to do with its purchase, but since Yahoo already has a photo sharing capability with more users than Flickr has, it seems quite possible that Yahoo bought Flickr in part because of Flickr's tagging expertise. Tagging may show up in various of Yahoo's offerings. Yahoo is the prototype of mainstream Internet applications.

It's also quite possible that tagging will sweep through the corporate world faster than expected. We know corporations see that they have a problem sharing information internally. That's why knowledge management (KM) became the buzzword du jour in the mid and late '90s. But the biggest obstacle to KM achieving its vision of making all information available to everyone in the organization was in fact the difficulty of building and maintaining large classification systems. Even then, such systems never represent everyone's way of thinking about things. Tagging, on the other hand, doesn't require a team of Information Architects to argue for years over whether the right term is "natural language processing," "language parsers," or "nlp." Users can use whatever term works for them. This may lower the barrier sufficiently to engage corporate users. If people get used to the benefits of tagging when on the corporate intranet, it will greatly boost the chances that they will demand it from the world-wide Internet as well.

But will people find the benefits worth the trouble? If so, it will be for two reasons. First, there are environments where tags provide a highly useful, flexible way of organizing one's own stuff, including one's computer desktop and archives of digital photographs and music. These are areas in which many of us adopt several points of view several times a day: Sometimes we want to see our files arranged alphabetically, sometimes by date, sometimes clustered by topic, sometimes by project, sometimes by the other files they refer to. Tags allow us that type of flexibility.

Second, and perhaps more important, social tags result in *tagstreams* – streams of objects tagged with the same tags by a collection of people. For example, if you care about the Civil War, applications already exist to show you all of the online resources tagged by others as relevant to that topic. Everyday you can see what the rest of the tagging world has found that accords with your interests. That is a powerful benefit.

There are, of course, challenges. Many of these are aggravated as tagging becomes more popular.

First and foremost are the bugaboos of information retrieval for decades: precision and recall. For example, if you are planning a trip to London, England, a search for every digital resource tagged as "London" will fetch the works of Jack London and maps of London, Ontario, exactly the same problem faced by standard text search engines (e.g., Google). And the search will not find photos tagged "Carnaby Street," "The Thames at night," or "Londres." The first reponse to this problem is: So what? In a

world of information abundance, one usually just needs good enough responses. After all, if you find 250,000 photos of London, does it really matter that you missed another 50,000? Usually not, although there are research projects – for example, case law – where such casualness is not acceptable.

Fortunately, there are already efforts underway to increase the precision and recall of tagging systems. Statistical analysis of tag sets can often discover that there is a strong correlation between pages tagged as "London," "Carnaby St." and "Londres." Thesauri and gazetteers can be consulted programmatically to discern equivalences and relationships.

Some have proposed intersecting social networks and tagging in order to increase the precision, recall and relevancy of tagging applications. If the application knows who is in one's social group, it can weigh the tags that group uses more heavily when executing searches. For example, if my social group uses the word "enterprise" only in the Star Trek sense, then when I search for resources tagged "enterprise," I'm statistically more likely to be looking for photos of Vulcans than of IBM's headquarters. Of course, the analysis can be much more complex and subtle than that.

Perhaps most interesting, tagging systems can create "folksonomies," a term invented by Thomas Vander Wal in 2004. A folksonomy is an emergent grassroots taxonomy. For example, if I'm about to tag a photo of San Francisco that I've found, if the application tells me that 5,000 people have already tagged it, and most have tagged it as "San Francisco" and only a few as "Frisco," that encourages me to tag it (and future photos of San Francisco) the popular way. After all, if I'm tagging it so I and others can find it, giving it an "oddball" tag means only oddballs will find it. If I tag it "San Francisco," now 5,001 people have used that particular tag, adding yet more impetus to that particular tag. Thus do folksonomies emerge.

Some have complained that a folksonomy can become a tyranny of the majority, favoring one language or one set of concepts over another. But the use of statistical analyses to cluster tags means that software may be able to discover the relationships among mainstream and oddball tags so that we get the benefit of greater precision and recall without having to conform to ways of thinking that are not natural to us.

As tagging becomes more popular, it will become more attractive to spammers who purposefully mis-tag their resources in order to make them more visible. Applications that search for tags already are instituting measures to sort out the spam. Those measures are unlikely to be perfect, but, perfection is not required.

So, tagging is likely to become a common and useful part of our networked computing environment. There are some reasons why this is important beyond its straightforward utility.

First, tagging will help social groups form around shared semantics, in addition to shared semantics arising from, and helping to define, groups. Tagging makes this more apparent. To take a trivial example, already at technical conferences it's common for the organizers to suggest some particular tag be used by the bloggers in attendance: "If you're blogging, we suggest you tag your posts with 'etech2005'." This is quite useful because of the "San Francisco" vs. "Frisco" problem. But this type of agreement on common nomenclature can bind a group. Imagine a tagging system used inside of a corporation. A group of people working on a project might decide on a set of tags. People not in the group who use the tags might thereby affiliate themselves socially with

the group or, depending on the circumstances, might be branded as interlopers. By their tags shall we know us.

Second, tagging repudiates one of the deepest projects our culture has undertaken over and over again: The rendering of all knowledge into a single, universal framework. The rendering has been assumed to be a process of discovery: The universe has an inner order that experts and authorities can expose. But in a networked world we know better than ever that such an order is a myth of rationality. We can't even agree even on basic issues such as what constitutes a "major" religion or a "legitimate" state. Order and categorization, we are learning, depend on context and project. The semi-chaotic state of the "tagosphere" represents the nature of our shared world better than the cool marble columns of the old mono-order ever could.

# # #