28% of Online Americans Have Used the Internet to Tag Content

Forget Dewey and His Decimals, Internet Users are Revolutionizing the Way We Classify Information – and Make Sense of It

Interview: Author David Weinberger Describes How Tagging Changes People's Relationship to Information and Each Other

By Lee Rainie, Director, Pew Internet and American Life Project
January 31, 2007

Just as the internet allows users to create and share their own media, it is also enabling them to organize digital material their own way, rather than relying on pre-existing formats of classifying information. A December 2006 survey by the Pew Internet & American Life Project has found that 28% of internet users have tagged or categorized content online such as photos, news stories or blog posts. On a typical day online, 7% of internet users say they tag or categorize online content.

These people said "yes" to the following question: "Please tell me if you ever use the internet to categorize or tag online content like a photo, news story, or a blog post." The wording was designed to capture the growing use of tagging on sites such as http://del.icio.us/ (a site for sharing browser bookmarks), http://www.flickr.com/ (a photo sharing site), http://youtube.com/ (a video sharing site) and http://technorati.com/ (the blog search engine).

Tagging is gaining prominence as an activity some classify as a Web 2.0 hallmark in part because it advances and personalizes online searching. Traditionally, search on the web (or within websites) is done by using keywords. Tagging is a kind of next-stage search phenomenon – a way to mark, store, and then retrieve the web content that users already found valuable and of which they want to keep track. It is, of course, more tailored to individual needs and not designed to be the all-inclusive system that Melvil Dewey tried to create with his decimal-based scheme for cataloguing library materials.

In a book to be released on May 1, *Everything Is Miscellaneous: The Power of the New Digital Disorder* (Times Books), David Weinberger, a fellow at Harvard's Berkman Center for Internet & Society and a prominent blogger, describes how people are putting ideas, information and knowledge together now that the digital age has encouraged alternatives to organizing information from hierarchical systems like the Dewey Decimal method. An online interview with Weinberger is featured at the end of this article.

How tagging works

This is the first time the Project has asked about tagging, so it is not clear exactly how fast the trend is growing. Tagging is one of the emerging Web 2.0 activities around which there is debate about what should be officially counted as tagging. That is, what sets tagging apart as a distinct activity, say, from creating a browser bookmark?

To add to the complexity of the issue, there are probably people who have created a tag who would use a different term for the activity. For example, some sites invite users to apply "labels" to content and don't use the word "tag." Other sites enable tagging so effortlessly that people might not be conscious they are doing it.

Tagging is the process of creating labels for online content. The mechanics are simple on most tag-centered websites. After creating an account on a site like flickr.com you can upload and label your own pictures to the site – for instance, labeling a picture of a sunset as "sunset." You can also search the site using keywords and when you find photos posted by others that you like enough to want to retrieve later, you can apply your own tags to them. That might mean that you call someone else's picture "sunset" even though he originally labeled it "clouds."

Then, from any internet-connected computer you go back to flickr.com and find all the material you have tagged – both yours and the material from others that you have labeled your own way.

Not only can tags be personally useful to people who want easier ways to retrieve information that appealed to them, but tags also have a social dimension. Your tags on flickr are added to the millions of other labels on the site and that allows flickr to organize information better for other searchers who use those keywords – making this a classic example of bottom-up building of categories instead of top-down imposition of categories.

Your tags also allow flickr to highlight the most popular listings. These "tag clouds" illustrate the material that was tagged by others and tag sites usually showcase the most popular tags by increasing the font size and boldness of the type as flickr does here: http://www.flickr.com/photos/tags/ .

Who the taggers are

Taggers look like classic early adopters of technology. They are more likely to be under age 40, and have higher levels of education and income.

Taggers are considerably more likely to have broadband connections at home, rather than dial-up connections. Men and women are equally likely to be taggers, while online minorities are a bit more likely than whites to be taggers.

The act of tagging is likely to be embraced by a more mainstream population in the future because many organizations are making it easier and easier to tag internet content. For instance, Gmail users can label their email content and Amazon users can apply the labels of their choosing to books and other published material.

Yahoo has added web applications that make it easy to tag and store web pages. Some sites have buttons on their web pages that allow their content to be stored on tagging sites with a simple click of a mouse.

There are even reports that some web users now have made tagging sites their home page, making these sites at least nominal competitors to big media companies that hope users will start their online experiences on their main page.

Demographics of Taggers	
28% of online Americans say they have tagged content like a photo, a news story or a blog post	
	Proportion of all Americans in the group who are taggers
Men	29%
Women	27%
Race/ethnicity	
White, non-Hispanic	26%
Black, non-Hispanic	36%
English-speaking Hispanic*	33%
Age	
18-29	32%
30-49	31%
50-64	23%
65+	18%
Educational attainment	
High school diploma	24%
Some college	28%
College degree +	31%
Household income	
<$30K	28%
$30K-$49,999	28%
$50K-$74,999	27%
$75,000+	36%
Internet connection at home	
Dial up	23%
Broadband	38%
Source: Pew Internet & American Life Project December 2006 tracking survey. N for internet users=1,623. Margin of error is ±3%).	

Tagging sites are getting more popular

Data from Hitwise, the web-tracking firm, show that tagging sites like flickr and del.icio.us have gained in popularity as internet users become aware of them.

The data are presented as a percentage of all web traffic.

Flickr.com traffic

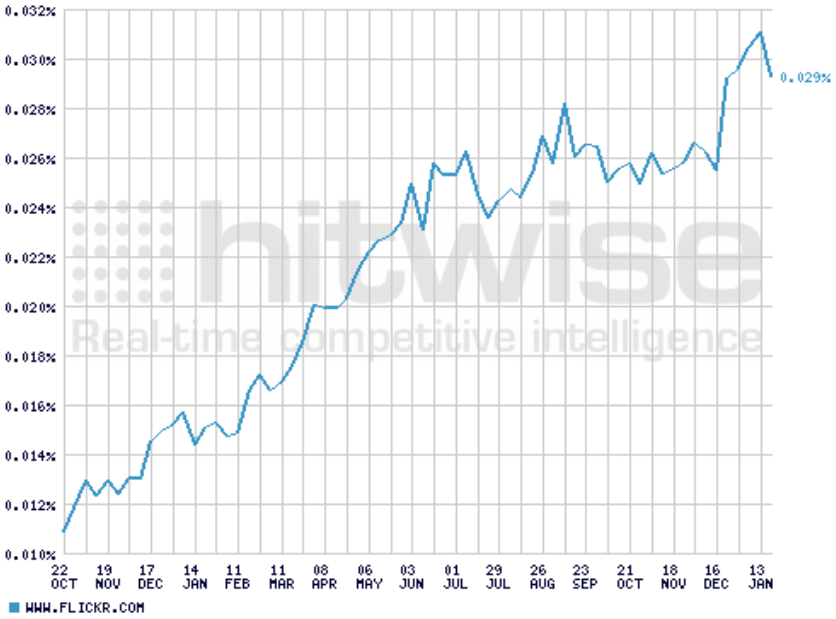

CHART OF THE WEEKLY ALL SITES MARKET SHARE IN 'ALL CATEGORIES', BASED ON VISITS.
TIME PERIODS REPRESENTED WITH BROKEN LINES INDICATE INSUFFICIENT DATA.
GENERATED ON: 01/26/2007. COPYRIGHT 2007 (C) 'HITWISE PTY LTD'.

Del.icio.us is a site where people can tag their website bookmarks and, again, share their tags with others.

Del.icio.us traffic

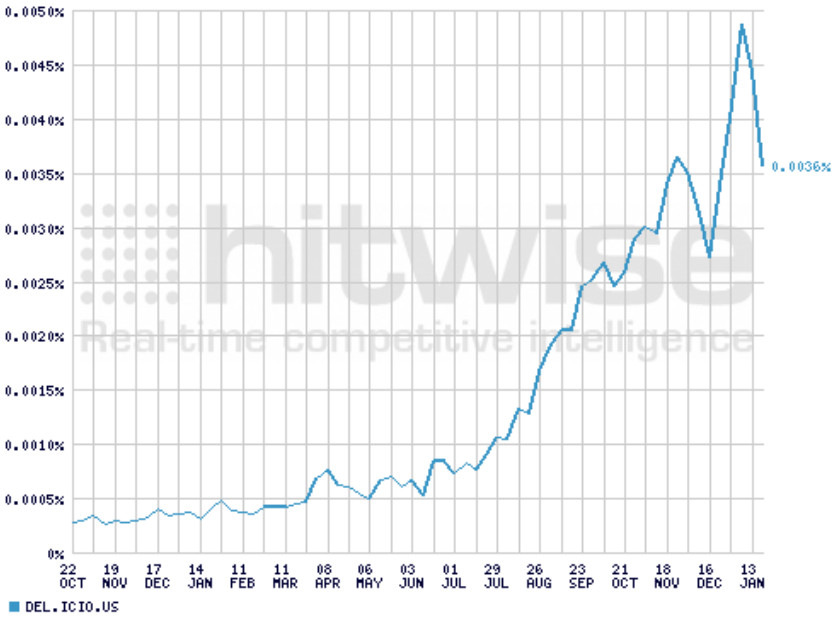

CHART OF THE WEEKLY ALL SITES MARKET SHARE IN 'ALL CATEGORIES', BASED ON VISITS.
TIME PERIODS REPRESENTED WITH BROKEN LINES INDICATE INSUFFICIENT DATA.
GENERATED ON: 01/26/2007. COPYRIGHT 2007 (C) 'HITWISE PTY LTD'.

Why Tagging Matters: An Interview with David Weinberger

In his upcoming book, *Everything Is Miscellaneous: The Power of the New Digital Disorder*, Weinberger describes how radical it is for people to move away from hierarchical classifications of information like the Dewey Decimal System, to individually- and group-arranged systems.

In Melvil Dewey's world, all information is divided into ten major topical categories that might have made perfect sense to well-educated Westerners who shared Dewey's frames of reference, but perhaps not to others. For instance, Dewey assigned the 800-899 block of numbers to literature and then assigned numbers 800-889 to American, European and classical languages. Thus, he squeezed every other bit of literature into the 10 remaining digits. Among other things, that means Russian literature did not even get its own whole number. It comes under 891.7, amidst East Indo-European and Celtic literatures.

It was also perfectly logical to Dewey that he list material relating to pets in the "technology" block of numbers in the 600s. Here's how he worked that out:

600	Technology
630	Agriculture and related technologies
636	Animal husbandry
636.7	Dogs
636.8	Cats[1]

In the 21st Century world of user-generated categories and meaning, this is not perhaps as useful or sensible as it once was. David Weinberger has thought through the many ways tagging changes people's relationship to information and each other. So, I traded emails with him on the subject and here's the result:

Q: What started the current interest in tagging?

Weinberger: The bookmarking site http://del.icio.us hit a nerve [in 2003] when it let users tag Web sites with a word or two so that they could find those sites later. And http://www.flickr.com hit the same nerve when it adopted tagging as a way to let people organize the photos they posted.

But the nerve was there, ready to be struck, because of two factors:

First, tagging lets us organize the vastness of the Web -- and even our email, as Gmail has shown -- using the categories that matter to us as individuals. You may want to tag, say, a Stephen King story as "horror," but maybe to me it's "ghost story" and to a literature professor it's "pop culture." Tagging lets us organize the Net our way.

---

[1] The general format of the Dewey Decimal System is as follows:

000s	Information (now including computer science) and general works
100s	Philosophy and psychology
200s	Religion
300s	Social sciences
400s	Language
500s	Science
600s	Technology
700s	Arts and recreation
800s	Literature
900s	History and geography

A good general description of the integers of the Dewey Decimal system can be found at http://www.oclc.org/dewey/resources/summaries/deweysummaries.pdf

Second, tagging is social. Tags used to be called “keywords,” and they've been with us for a long time. But only recently have we been making them public. That has big effects. By searching for a tag we can find material others have discovered ahead of us: At Amazon’s “most popular tags” page, (http://www.amazon.com/gp/tagging/cloud/ref=tag_sr_nss/103-6276287-3099815) a search for things tagged “horror” turns up almost three thousand books and movies.

Tagging also allows social groups to form around similarities of interests and points of view. If you're using the same tags as I do, we probably share some deep commonalities.

And, by looking over the public field of tags, we can see which tags are most frequently used and how they relate. Those patterns are called “folksonomies” -- it's a play on the word “taxonomies.” Folksonomies reveal how the public is making sense of things, not just how expert cataloguers think we *ought* to be thinking.

Q: Why do you think Internet users are drawn to tagging?

Weinberger: It's really useful. Compare your traditional computer system to organize your digital photos to using a tagging system. Instead of having to stick a photo into a single folder -- say, “trips 2006” -- you can easily tag it as “Italy,” “anniversary,” “sunset,” “mountains,” and “no kids.” You can assemble instant virtual albums of all your anniversary photos, or all your photos of all your trips to Italy, etc.

There's an altruistic appeal to tagging as well. Tagging at public sites can give you a sense that you're adding to a shared stream of knowledge. At del.icio.us, or other such sites, tag a page “robotics” and you know that it's automatically added to the list of pages tagged that way, so anyone else interested in that topic can find it.

Q: So, there are benefits beyond the individual.

Weinberger: Absolutely. Maybe the most interesting thing about tagging is that we now have millions and millions of people who are saying, in public, what they think pages and images are about. That's crucial information that we can use to pull together new ideas and information across the endless sea we've created for ourselves.

Q: Does tagging create problems?

Weinberger: What doesn't? Tags work because they're so simple, but because they're so simple, they can be ambiguous. The tag “roman,” for example, might refer to an Italian fountain, the director Roman Polanski, or the French word for “novel.” So, there's a possibility for misunderstanding. And if you search for photos tagged “San Francisco,” you may not see photos tagged “sf” or “Golden Gate.” So, if you need to find everything about a topic, you often can't rely on tags.

More broadly, some worry that folksonomies can be a type of “tyranny of the majority,” in which the prevalent group's way of thinking about the world overwhelms the local and the quirky. That's something to watch out for, but by analyzing tag sets we can also build a tag thesaurus that knows that the tag “roman” may be equivalent to the tag “novel” in some circumstances.

Q: What's the future of tagging?

Weinberger: Because it's useful when there's lots of information and the information is truly meaningful to individuals, it'll be adopted more and more widely. But we're also going to invent new ways to harvest tagging. Flickr, for example, is already able to cluster photographs by subject with impressive accuracy just by analyzing their tags, so that photos of Gerald Ford are separated from photos of Ford Motor cars. We'll also undoubtedly figure out how to intersect tags

with social networks, so that the tags created by people we know and respect have more “weight” when we search for tagged items. In fact, by analyzing how various social groups use tags, we can do better at understanding how seemingly different worldviews map to one another.

About the Pew Internet & American Life Project

The Project is a non-profit, non-partisan initiative of the Pew Research Center that is funded by the Pew Charitable Trusts to do research about the social impact of the internet. The Project takes no positions on policy issues. Its reports and data can be found at http://www.pewinternet.org/.

Methodology and question language

December 2006 Tracking Survey Final Topline 1/5/07

Data for November 30 – December 30, 2006

Princeton Survey Research Associates International
for the Pew Internet & American Life Project

Sample: *n* = 2,373 adults 18 and older
Interviewing dates: 11.30.06 – 12.30.06
Margin of error is plus or minus 2 percentage points for results based on total sample [n=2,373]
Margin of error is plus or minus 3 percentage points for results based on total internet users [n=1,623]

Asked of internet users

WEB1 Please tell me if you ever use the internet to do any of the following things. Do you ever use the internet to…/Did you happen to do this **yesterday**, or not?[2]

	TOTAL HAVE EVER DONE THIS	DID YESTERDAY	HAVE NOT DONE THIS	DON'T KNOW/ REFUSED
Categorize or tag online content like a photo, news story or blog post	28	7	70	2

This report is based on the findings of a daily tracking survey on Americans' use of the Internet. The results in this report are based on data from telephone interviews conducted by Princeton Survey Research Associates International between November 30 to December 30, 2006, among a sample of 2,373 adults, 18 and older. For results based on the total sample, one can say with 95% confidence that the error attributable to sampling and other random effects is plus or minus 2.3 percentage points. For results based Internet users (n=1623), the margin of sampling error is plus or minus 2.7 percentage points. In addition to sampling error, question wording and practical difficulties in conducting telephone surveys may introduce some error or bias into the findings of opinion polls.

The sample for this survey is a random digit sample of telephone numbers selected from telephone exchanges in the continental United States. The random digit aspect of the sample is used to avoid "listing" bias and provides representation of both listed and unlisted numbers (including not-yet-listed numbers). The design of the sample achieves this representation by

[2] Prior to January 2005, question wording was "Please tell me if you ever do any of the following when you go online. Do you ever…?/Did you happen to do this yesterday, or not?"

random generation of the last two digits of telephone numbers selected on the basis of their area code, telephone exchange, and bank number.

New sample was released daily and was kept in the field for at least five days. The sample was released in replicates, which are representative subsamples of the larger population. This ensures that complete call procedures were followed for the entire sample. At least 10 attempts were made to complete an interview at sampled households. The calls were staggered over times of day and days of the week to maximize the chances of making contact with a potential respondent. Each household received at least one daytime call in an attempt to find someone at home. In each contacted household, interviewers asked to speak with the youngest male currently at home. If no male was available, interviewers asked to speak with the youngest female at home. This systematic respondent selection technique has been shown to produce samples that closely mirror the population in terms of age and gender. All interviews completed on any given day were considered to be the final sample for that day.

Non-response in telephone interviews produces some known biases in survey-derived estimates because participation tends to vary for different subgroups of the population, and these subgroups are likely to vary also on questions of substantive interest. In order to compensate for these known biases, the sample data are weighted in analysis. The demographic weighting parameters are derived from a special analysis of the most recently available Census Bureau's March 2006 Annual Social and Economic Supplement. This analysis produces population parameters for the demographic characteristics of adults age 18 or older, living in households that contain a telephone. These parameters are then compared with the sample characteristics to construct sample weights. The weights are derived using an iterative technique that simultaneously balances the distribution of all weighting parameters.

The final response rate is 27 percent.