Folksonomies are one of today’s hottest Internet trends. They are but one part of Web 2.0, which, in part, refers to the ability of Internet users to add, change and improve World Wide Web content... p. 7

SPECIAL SECTION

Folksonomies

7] Introduction: Folksonomies and Image Tagging: Seeing the Future? by Diane Neal, Guest Editor

12] Why Are They Tagging, and Why Do We Want Them To? by P. Jason Morrison


21] Image Indexing: How Can I Find a Nice Pair of Italian Shoes? by Elaine Ménard

26] Flickr Image Tagging: Patterns Made Visible by Joan Beaudoin

FEATURES

30] Them! Google’s Ambivalence toward Library and Information Science by Shawne D. Miksa
In an earlier column (February/March 2007) I outlined some of the challenges that societies like ASIS&T face in keeping their publication programs topicaly relevant, current in delivery and financially healthy. The changing nature of electronic access means that we have to provide the degree of freedom of access that our members expect, while maintaining the intellectual and financial integrity of our publications. This year the Board and the Publications Committee have been devoting a lot of time to publication-related issues, and since this is the last President’s Page that I will write for you, I’d like to end my term with an update on this activity.

Several major changes took place this year. With the June/July 2007 issue, the Bulletin of the American Society for Information Science and Technology became an electronic-only publication. Editor Irene Travis and her team have done a great job of bringing you the Bulletin with a new look: enhanced format, more content, more images and colors, and improved readability. So far the feedback has been overwhelmingly positive. However, as Irene notes in her column, the all-electronic Bulletin remains a work-in-progress, so comments and suggestions are welcome.

JASIST has been available online for some time, but unless you have checked the backfile recently, you may not have noticed how the material available has grown. In April of this year Wiley completed digitization of the journal’s backfile, to volume 1 issue 1 of American Documentation, published in 1950 – that is 57 years of information science online. No excuse now for not citing that early work in your area!

While the full run of JASIST is available in electronic form from subscribing libraries and in the ASIS&T Digital Library, as a society publication produced by a commercial publisher JASIST has been slow to develop a policy on open access. That situation is about to change, prompted by a motion from the JASIST Editorial Board. After receiving advice from a task force chaired by Carol Tenopir to consider the implications, the ASIS&T Board recommended that authors be authorized to post their pre- and post-acceptance versions of their papers on their personal and institutional websites. This policy and its implementation are now being negotiated with the publisher and the details will be announced soon.

An internal change with our JASIST publisher, John Wiley & Sons, has also had an impact on us. In February, Wiley acquired Blackwell Publishing and assigned JASIST to its Blackwell division. While this transfer may seem like just so much corporate restructuring, it has made...
a difference to us. Blackwell is known as a publisher of society journals, working with over 650 societies and has extensive experience with the non-corporate sector. Their understanding of the Society’s role as a publisher has smoothed our interaction with them. We look forward to working with Blackwell for the remainder of our Wiley contract.

Our current publishing contract with Wiley will end in 2010. Our Publications Committee, chaired by former president Sam Hastings, is actively engaged in a process to identify the publisher who will carry the publication forward to 2017. A Request for Proposal (RFP) was developed by the Publications Committee in conjunction with consultant Morna Conway, who has extensive experience with scholarly society publications. The RFP has been circulated to a number of established and well-regarded publishers (including, of course, the current publisher, Wiley Blackwell), to create a pool of candidates from which to select the publisher to take us through the next decade. We expect to make this decision by early next year. This is an important decision for us since our publishing program contributes a significant percentage of the operating funds of the Society. We’re fortunate that we have a number of members with substantial expertise in the publishing industry to help us in this important process.

A search for a new editor for *JASIST* is also on the horizon. Of course a significant indicator of the success of our publications is the quality of the content we as a field produce. With Don Kraft serving as editor since 1985, *JASIST* has been among the top-ranked journals in its ISI categories. With Don’s impending retirement, a new editor is needed. We are no longer asking for a lifetime commitment — in common with most of our peer journals, we are introducing a limited term appointment for *JASIST* editors. The Board has decided on a five-year term, renewable once. A task force has been appointed to develop the selection process, identify and assess candidates, and make a recommendation to the Board in time for the March 2008 Board meeting.

Initiatives in other areas are underway, which you will be hearing about in the coming months. But this will be my last column as ASIS&T president, as my term comes to an end in October and I slide, gratefully, into the role of past president. Nancy Roderer will take over as president. It has been an exciting year and I have a greater appreciation of what makes a society like ours successful. I’ve benefited from the dedication and hard work of members who serve on the Board and on ASIS&T committees and task forces. My requests for help have been met, more often than not, with a willingness to contribute, for which I am very grateful. I’m also very grateful for the support which I received from our knowledgeable and hard-working executive office, whose day-to-day operations help to channel the volunteer effort from our members. Thank you all for your encouragement and support!
Dealing, as it consistently does, with the impact of information technology change, this issue of the Bulletin has a speculative tone reflected in the title of our special section, “Folksonomies and Image Tagging: Seeing The Future?” Diane Neal is our guest editor. Supplying text descriptions for images remains critical to their retrieval because content-based systems (those that search by features of the image such as color or texture) have relatively limited capabilities. One means to obtain such descriptions is to encourage image owners and users to add text tags, both for their own purposes and to enhance image retrieval generally. Therefore, the question of what motivates users to tag is important; James Morrison offers a discussion of the topic. Where there is social interaction such as that on collaborative tagging sites there is also the potential for friction. Chris Landbeck considers modes of conflict resolution in these venues. Finally, Elaine Ménard and Joan Beaudoin each analyze a sample of tags supplied by users and the relationship of user tagging to professional indexing.

If there is text to search, the performance and characteristics of search engines are, of course, critically important. Search engine sites, particularly Google, have had a major impact on libraries, a relationship that Shawne Miksa examines in a feature article. But websites themselves are not immune. Marianne Sweeney warns that search engine technology may soon be preferred to other user-centered site navigation techniques that have been developed by information architects, user experience designers and others. She urges these groups to become more involved in search engine development and more sophisticated in integrating search technology into their sites.

Given the complexity of changes such as those reflected in the reports of our other authors, Miriam Vieira da Cunha believes that interdisciplinary teamwork is key to meeting the challenges of our current information environment. Reporting from Brazil in our International Column, she contends that information scientists are still not assertive enough in joining and participating in such efforts, which are crucial to realizing the potential benefits of the information society – a mission that is central to information science in the 21st century.

Information science in the 21st century leads us back, in conclusion, to our own Society, and the future of its publications, which is the subject of Edie Rasmussen’s final President’s Page. In it she reports primarily on the Society’s Journal of the American Society for Information Science and Technology – on publisher relations and on plans to allow the posting of pre-and post-prints of articles from the Journal in institutional repositories.
Election Underway for ASIS&T Officers and Directors for Seats on Board

As this issue of the Bulletin of the American Society for Information Science and Technology is prepared for electronic publication, the deadline is approaching for all voting members of the Society to complete their ballots for the election of new officers and directors. For the first time, the ASIS&T election is being conducted almost exclusively by online voting. Only those who make a specific request to have ballot materials mailed in hard copy will vote that way.

Full results of the election will be available at the ASIS&T website when balloting is completed.

Running for the presidency of ASIS&T, to take office in October as president-elect, are Donald O. Case and Barbara H. Kwasnik. Two candidates, Vicki L. Gregory and Jon W. Simons, are running for a three-year term as Society treasurer. In addition, four candidates are vying for two available seats as director-at-large: Efthimis N. Efthimiadis, Beatrice R. Pulliam, Victor Rosenberg and Barbara M. Wildemuth. All newly elected members of the ASIS&T Board of Directors will take their seats at the conclusion of the ASIS&T Annual Meeting in Milwaukee.

CANDIDATES FOR PRESIDENT-ELECT

Donald Case is a professor in the College of Communication and Information Studies at the University of Kentucky, where he previously served as director of the School of Library and Information Science. He holds a Ph.D. in communication research from Stanford University and an MLS from Syracuse. He has served as director-at-large on the ASIS&T Board of Directors for the past three years. Among his professional honors is the ASIS&T Best Information Science Book Award in 2003 for Looking for Information: A Survey of Research on Information Seeking, Needs and Behavior.

Barbara Kwasnik is a professor in the Information School at Syracuse University, where she has taught in the areas of knowledge organization and research methods for 20 years. She has served as director of both the MSLIS and Ph.D. programs. She holds a Ph.D. from Rutgers University and an MLS from Queens College, CUNY. Her research interests include issues of adapting classifications and the use of genre information for enhancing information retrieval. Among many honors, she was recognized with the ASIS&T Outstanding Information Science Teacher Award in 2001.

CANDIDATES FOR DIRECTORS-AT-LARGE, 2008-2010 (2 POSITIONS)

Efthimis Efthimiadis, associate professor in the Information School at the University of Washington, teaches in the areas of information retrieval, database design and web search, and he conducts research on user-centered design and evaluation of information retrieval systems. A longtime member of ASIS&T, he has served the Society in numerous ways on national committees, awards juries, conference committees and in chapters and SIGs. He holds a Ph.D. in information science from City University, London.

Beatrice Pulliam, assistant professor and reference librarian, Providence College, is responsible for reference, user instruction and coordination of library applications for government clients. Previously, he directed the development and maintenance of ORBIT information retrieval systems at several companies. He has been active member of two ASIS&T chapters, including service as treasurer for Potomac Valley Chapter.

CANDIDATES FOR TREASURER, 2008-2010

Vicki Gregory, professor in the School of Library and Information Science, University of South Florida, has served on the ASIS&T Budget and Finance Committee for the past four years. She has also served in many local and national positions with chapters and Special Interest Groups.

Jon Simons, senior systems analyst at Westat, an employee-owned research corporation, develops and manages publication clearinghouse database applications for government clients. Previously, he directed the development and maintenance of ORBIT information retrieval systems at several companies. He has been active member of two ASIS&T chapters, including service as treasurer for Potomac Valley Chapter.
technology. She holds a master of science degree in library and information science from Simmons College. She joined ASIS&T as a student and has remained involved in a wide variety of chapter and SIG activities since, as well as on national awards juries.

**Victor Rosenberg**, associate professor, School of Information, University of Michigan, previously taught at both University of California at Berkeley and the Universidade Federal de Minas Gerais in Brazil. His broad research interests include electronic commerce, information retrieval, information policy, technology in the humanities and entrepreneurship. His own experience as an entrepreneur includes founding and serving as CEO of Personal Bibliographic Software, Inc., developer of ProCite and BiblioLink software products.

**Barbara Wildemuth** is a professor in the School of Information and Library Science at the University of North Carolina at Chapel Hill, where she previously served as associate dean for undergraduate programs. Her research projects include a collaborative digital library curriculum development effort; a test of the feasibility of recruiting and training medical informationists; the Open Video project, examining people’s interactions with a digital video collection; and an assessment of people’s needs for personal health records. She holds a Ph.D. from Drexel University.

The **Southern Ohio Chapter of ASIS&T** (SOASIST) offered Eric Miller, president of Zepheira, The Art of Data, speaking on the Semantic Web as its September meeting. Known for his vision and leadership roles in the Dublin Core Metadata Initiative at OCLC and the Semantic Web Initiative for W3C at MIT, Miller discussed how organizations can achieve data interoperability for seamless integration of online products by leveraging Semantic Web standards.

The **Northern Ohio Chapter of ASIS&T** (NORASIST) invited **David B. Robins**, assistant professor at Kent State University, to discuss **Accessible Design for the Web: Legal Requirements, Standards and Design Tips**, at the chapter’s annual business meeting. The presentation focused on how organizations with a web presence can meet legal requirements to make their websites accessible to the blind and other disabled persons. Following the program and the election of officers, those in attendance were invited to tour the newly renovated Cleveland Heights University Heights Public Library.

**Mark Your Calendar**

**October 19-24 Milwaukee, Wisconsin**

**Joining Research and Practice: Social Computing and Information Science**

[www.asis.org/Conferences/AM07/](http://www.asis.org/Conferences/AM07/)
Folksonomies and Image Tagging: Seeing the Future?
by Diane Neal, Guest Editor

Folksonomies are one of today’s hottest Internet trends. They are but one part of Web 2.0, which, in part, refers to the ability of Internet users to add, change and improve World Wide Web content. A folksonomy is created as users of a website add “tags” (keywords) to describe items on a website. The users choose their own keywords; few or no restrictions are imposed on their choices. The terms are not chosen from a previously existing controlled vocabulary, a strict taxonomy or any other officially sanctioned method of bibliographic description.

According to information architect Thomas Vander Wal, who created the term *folksonomy*, folksonomies are created when people tag items online for their own later information retrieval purposes. This makes folksonomy an extremely useful tool for online personal information management, since the tags are coined in the user’s own words, not in the words imposed by the system. The secondary benefit is the social aspect, including the ability of other users to use those tags for search and retrieval of previously undiscovered items and the community created by and centered on users’ tags.

Websites utilizing folksonomies, such as the photograph sharing website Flickr, the social bookmarking website del.icio.us and the video sharing website YouTube, have enjoyed exponentially growing numbers of registered users since their relatively recent inception. It is no wonder that this growth has occurred: tagging is fun! Your inner cataloger will love using LibraryThing to tag each item in your personal collection of books, compact discs and DVDs. One of my former students has even used it to catalog her church’s library collection. Using Digg, you can “digg” and tag your favorite online news articles.

The online retailer Amazon allows its users to tag anything for sale and to work with the tags in many useful ways. Figure 1 shows the tag section of the amazon.com page used to sell the Canon PowerShot A550 7.1MP Digital Camera with 4x Optical Zoom. Not only can you tag the camera, you can choose tags that others have already used, find discussions and people related to those tags and tag it for Amazon Search. Tags submitted for inclusion in Amazon search must adhere to a set of guidelines, and they must be approved by amazon.com.

On del.icio.us, you can store all your “Favorites” or “Bookmarks” online, tag them and access them on any computer – no need to remember whether you bookmarked a website on your home or your work computer. Figure 2 (see page 8) illustrates some websites I have included in my del.icio.us account. Using del.icio.us, you can also see which websites others have tagged, how many people have saved given websites into their favorites, and the most popular and recent tags. Hence, websites such as del.icio.us are called “social bookmarking” websites.
Unintentionally collaborative?

Figure 3 illustrates a tagged picture on Flickr. I took this photo of my husband Jason standing by a sculpture located in front of the Fort Worth Museum of Modern Art. I added the tags “Vortex,” “Fort Worth” and “The Modern.” Someone else who viewed the picture tagged it with “Richard Serra,” which is the sculptor’s name. I was pleased that the person added the sculptor’s name, since Jason and I did not know it. This demonstrates the social utility of folksonomies – as a secondary benefit, all users are able to contribute their knowledge to the folksonomy. Consequently, the tags serve as a representation of the collective knowledge of the users. Tagging and folksonomy are sometimes called \textit{social classification}.

Tag clouds, or visual representations of the most popular tags on a website utilizing folksonomies, provide insight into the collective knowledge, or at least the collective interests, of the users. Figure 4 (see page 9) shows a portion of the tag cloud representing the most active tags on CiteULike, a website for organizing and tagging online scholarly resources. The most popular tags, such as \textit{clustering} and \textit{evolution}, are in larger type, while the slightly less popular tags, such as \textit{communication} and \textit{climate}, are in smaller type.

Social communities form around tags. On Flickr, for example, users can join groups focused on individual tags such as \textit{beautiful}, \textit{LOVE IT!} and \textit{Ulster} (a province in Northern Ireland). In other cases, the people who can tag in a given area are limited to a particular online community. Users can only join some Flickr groups via invitation or moderator approval.
Speaking of community, many libraries provide outlets for their staff and patrons to participate in tagging and folksonomy creation. The Ann Arbor District Library’s catalog allows users to tag its items. A tag cloud of the tags applied to the Edmonton Public Library’s subject guide pages is visible at CiteULike. Libraries can opt to use LibraryThing tags within AquaBrowser Library, a faceted visual search engine. AquaBrowser Library provides a demo of this mashup for the Queens Library. Libraries such as the Thomas Ford Memorial Library in Western Springs, Illinois, have tagged photos on Flickr that can feature anything of interest to its community. Examples include a book sale, a guest speaker or digitized photos from the library’s historical collection. The library staff, library patrons and any other viewer of the library’s photo collection can tag the pictures. For the library, a Flickr account provides free marketing and excellent publicity. For its patrons, it provides a way to participate in local community and a way to connect with the library in the domain where many patrons already live – online!

Tagging methods focused on people’s specific interests are emerging. For example, geography enthusiasts can participate in geotagging. The practice of geotagging records geographical information about a document. For example, geotagged Flickr photos provide information about where the photo was shot. They have the tag “geotagged” and may include verbally oriented place information such as “Germany” or “my sister’s house.” A more technically accurate geotagging method involves listing the latitude and longitude associated with the item. Using this approach, a picture of Durham, North Carolina, would be geotagged with “geo:lat=5.910440” and “geo:lon=-78.906380.”

Not just for technogeeks?

Attendees of the ASIS&T 2006 Annual Meeting in Austin may remember June Abbas’ and Jennifer Graham’s poster session titled “So let’s talk about tagging, user defined/applied descriptors: A research and curricular agenda.” Abbas and Graham provided markers and sticky notes for conference attendees to place tags on their poster. In the end, we tagged the poster, ourselves, our colleagues, objects around the hotel and anything else we could find. For the remainder of the conference, I enjoyed wearing a sticky note on my badge displaying a tag I created – me.

Many Web 2.0 components, such as folksonomies, are not simply singular technologies, but rather technically based manifestations of human tendencies. When Vander Wal created the word folksonomy, he combined the words folk and taxonomy. In a folk taxonomy, members of a society create words, categories and classifications for things in order to describe the world in a way that holds relevant meaning for them. Similarly, a folksonomy allows members of a group, such as one that uses the same OPAC or one that likes pictures of cute puppies, to search for and retrieve documents, communicate with each other and share their knowledge online in a way that means something to them.

The ability to create tags that hold personal meaning is important to Internet users. We are probably more likely to remember our own tags for searching than terms belonging to a predetermined system. (Could our own sets of tags be considered individualized controlled vocabularies?) Some folksonomy researchers have concluded that the emotion and affect elicited by a document play a large part in users’ tag choices; they may tag and search for beautiful or expansive pictures. In “Two Kinds of Power: An Essay on Bibliographical Control,” Patrick Wilson wrote, “What seems to stand out to us depends on us as well as the writing, what we are ready to notice, what captures our attention.” Document description subjectivity has been well-documented in information science research, so it makes sense to allow users to apply their own tags to their own documents.
You can’t see it that way?

Representation of visual documents, such as images and videos, is particularly susceptible to subjectivity. Although “a picture is worth a thousand words,” no words accompany a picture or video to guide our tag choices. Vision research has demonstrated our tendency to first look at what interests us the most in a picture. So different groups of visitors to an online museum exhibit – an artist, a casual visitor and an art scholar, for example – may find different elements of interest in the same painting.

For my dissertation research, I investigated photojournalism professionals’ preferences for search and retrieval methods of digital news photographs in online photo archival systems. These professionals rate photographer-supplied keywords (tags) as a highly preferred method, which is the method employed by most of the newspapers I studied. In this case, tagging is necessary: likely only the photographer knows the name of the person in a given photo, the location in which he shot a photo and so on. However, my research uncovered some issues that potentially inhibit successful photo retrieval, such as misspelled tags and various inconsistencies in the photographers’ tagging practices.

I have little doubt that tag quality may be an issue in many contexts. Certainly, if keywords in any form are misapplied or misspelled, they will hinder successful retrieval. However, it may be difficult to determine whether tags are truly incorrect in the user’s context. One possibility for improving tag quality may include a system that suggests accurately spelled tags or provides an “if you liked this tag, you’ll love this other tag” option.

In the end, many online users need and want control over the representation of their own documents. The sheer volume of tags and folksonomies present in today’s online environment (and, we must assume, the success users find in searching on their own tags for later retrieval) provides sufficient evidence that information professionals should not attempt to impede the folksonomy phenomenon. Some folksonomy skeptics argue in favor of more traditional, controlled approaches to online document representation, such as top-down hierarchical taxonomies and faceted schemes. These methods certainly hold value in many contexts, but it seems counterproductive to simply dismiss the value of user-assigned descriptors. It seems more productive for us to work with users in the development of mostly unfettered, but still appropriately controlled, tagging systems. This practice is becoming known in some circles as *collabulary*. 

For Further Reading


Websites Mentioned in the Article

Amazon – www.amazon.com/
AquaBrowser Library - www.medialab.nl/
CiteULike – www.citeulike.org
del.icio.us – http://del.icio.us/
Digg – www.digg.com/
ESPGame – www.espgame.org/
Flickr – www.flickr.com/
LibraryThing - www.librarything.com/
YouTube – www.youtube.com/

Examples of Tagging Applied in Libraries

Ann Arbor District Library Catalog – www.aadl.org/catalog/
Edmonton Public Library Subject Guide Pages tag cloud – www.ep.l.ca/EPLCloudNine.cfm?all=yes
Queens Library AquaBrowser demonstration – http://aqua.queenslibrary.org/

It seems counterproductive to simply dismiss the value of user-assigned descriptors. It seems more productive for us to work with users in the development of mostly unfettered, but still appropriately controlled, tagging systems. This practice is becoming known in some circles as *collabulary*.
So how do we start?

Both the personal information management dimension and the social dimension of folksonomy warrant further thought, research and development. I hope that this special section on folksonomies and image tagging will provide you with ideas for some directions that information professionals can take to improve the “bottom-up” intellectual access to information that is so popular online today.

P. Jason Morrison’s article “Why Are They Tagging, and Why Do We Want Them To?” further defines the utility of folksonomy; its main use is to support successful information retrieval, but he discusses other uses as well. He also explains some of the reasons why users tag. These include retrieving information, gaining exposure and sharing opinions. More reasons can be found in his article.

As users contribute tags and other information to a website, certainly there will be some disagreement about the contributed content. In “Trouble in Paradise: Conflict Management and Resolution in Social Classification Environments,” Chris Landbeck explores the different conflict resolution methods available on three websites that allow tagging: Flickr, Wikipedia and ESPGame.

In keeping with our emphasis on image tagging, Elaine Ménard’s article “Image Indexing: How Can I Find a Nice Pair of Italian Shoes?” explores the special issues associated with images and image users. She compares the usefulness of image tagging with the value of controlled vocabularies used for image indexing. Also, Joan Beaudoin’s “Flickr Image Tagging: Patterns Made Visible” discusses a categorization study of 140 Flickr tags. Based on her model and analysis, the most frequently used categories of tags include names of places, things and people associated with the photograph.

Consolidated List of Websites mentioned in the Special Section on Folksonomies

Amazon – http://www.amazon.com/
AquaBrowser Library – http://www.medialab.nl/
CiteULike – http://www.citeulike.org/
del.icio.us – http://del.icio.us/
Digg – http://www.digg.com/
ESPGame – http://www.espgame.org/
Furl – http://www.furl.net/
Flickr – http://www.flickr.com/
LibraryThing – http://www.librarything.com/
Mealographer – http://www.mealographer.com/
Reddit – http://reddit.com/
Slashdot – http://slashdot.org/faq/tags.shtml
StumbleUpon – http://www.stumbleupon.com/
Technorati – http://www.technorati.com/
YouTube – http://www.youtube.com/
Social tagging systems, which allow large numbers of users to classify items, are a fascinating subject. When I began planning for my master’s thesis in the Information Architecture and Knowledge Management Program at Kent State University, I knew I wanted to study collaborative tagging and folksonomies. The next step, deciding on which systems to study and how to study them, was more difficult.

Just a year or two earlier the task of narrowing down the types of folksonomies to study would have been much easier. Now it seems as though nearly every website invites users to tag articles, photos, videos and merchandise with keywords. Folksonomies are no longer limited to cool Web2.0 startups working out of their garages to change the way we think about enterprise social wiki pet care. Amazon displays user tags on item pages, IBM sells products that employ social tagging and Microsoft’s blogs have tag clouds.

In the course of choosing specific sites to study, I looked at a lot of different tagging systems and folksonomies. While browsing, searching and trying out as many tagging systems as I could get my hands on, I noticed some broader issues that I could not cover directly in my research but definitely merit discussion. Information architects and web developers will be asked more and more to incorporate these systems into websites, and there are a number of open questions about how the systems should work and how they are used.

On the surface, tagging systems and the resulting folksonomies seem almost magical. Is it really possible that the problem of organizing and classifying can be solved as simply as allowing random users to contribute tags? This notion is made even more tempting by the fact that many content management systems now have the functionality built in, or it’s just a matter of installing a plug-in.

I suspect, however, that a folksonomy is most likely to be successful when the goals of the website or information system intersect with the goals and motivations of users. So when considering how to add tagging functionality and how to build and use a folksonomy, we have to ask: What are folksonomies for, and why do users tag?

What are folksonomies for?

Folksonomies are generally used to organize information and support information retrieval (IR). The public face of many folksonomies is often a tag cloud, with tags usually listed alphabetically and weighted by popularity. Users might navigate the folksonomy by following tags they’ve used, related tags, popular tags or recent tags. In many cases folksonomies are also used to support search. A site with a huge database of photos like Flickr, for example, might rely almost entirely on a folksonomy for search since there is no full text to fall back on.

Tags might have some other, more subtle uses as well. Even if users do not use the folksonomy to navigate the site, the presence of descriptive tags may suggest whether they are on the right track. In addition a site’s information architects and user experience designers could make use of a folksonomy themselves. For example, tags could be data mined to provide additions to synonym rings or an insight into the kinds of topics users are thinking about at the moment.

It is also possible that a site might adopt tagging and create a folksonomy with no real interest in how well it supports IR whatsoever. Tagging could
be seen as a small way to get users interacting with a site, to get them to return, to encourage them to sign up for a user account or just to give them a richer experience and a sense of participation. This is the same reason why many sites allow users to rate items, vote in polls and contribute comments.

When thinking about adding tagging to a site, the first question should be: What do we want to get out of this? Does the site need something to improve search results or a new navigational facet to better connect related pages? Is the goal to classify lots of multimedia objects with minimal cost or to get users to interact with the site a little more? The answer could be "all of the above." Of course if the cost of adding the functionality is negligible, the answer might even be "who knows, let’s throw it at the wall and see what sticks."

Once the goals have been decided, decisions about the tagging interface can be made to support them. There are so many open questions at this point that are just begging for further research. Should the tagging system suggest tags to users? If so, should it suggest popular tags from other users or that user’s most-used tags? Should the tagging system attempt to control vocabulary in any way or perhaps just apply spell-check? What should the search system search? When users bookmark a web page with Furl, for example, they are able to contribute topics, keywords, comments, clippings and a rating, and Furl can also save a copy of the page.

In my research on the use of folksonomies to support search I found a great example of how relatively small details can impact the usefulness of the entire system. Time and time again I wondered why so many systems encourage or require single-word tags. On the popular tags page at del.icio.us, tags like howto and rubyonrails illustrate that expressing even simple concepts often requires more than a single word. Users can work around this limitation by concatenating words together, but this likely limits the folksonomy’s IR performance. While users browsing a tag cloud might recognize howto as denoting items that explain “how to” do something, users typing search queries will not have the same luck. Del.icio.us contains thousands of items tagged information, architecture and informationarchitecture. If a user searches for “information architecture,” how is the search system to know which of those three tags are relevant? If one of the goals of the folksonomy is to support search, my guess is it is worth the extra lines of code to support spaces or at least some substitute character.

**Why are users tagging?**

Social bookmarking systems like del.icio.us and Furl worked well for my study because users tagged websites in order to organize their bookmarks and find things later or to share them with others. On photo and video sites users might be motivated to tag for similar reasons but also because without relevant tags the items they are adding to collections may not be searchable at all. When I looked at news sites like Slashdot, however, I began to wonder what utility users get from tagging. Tags on some sites were often inside jokes by regular users or comments on the quality of a story. On some sites when the headline for a story was in the form of a question the most popular tags were invariably “yes” and “no.”

Although I did not test any news sites, my guess is that in cases like this the results are much weaker for search. Inside jokes might help some long-time users navigate, but it is hard to imagine the utility of a tag cloud filled with yes, no and maybe. Users must be getting something out of these systems, though, and it is important to realize that different users have different motivations for tagging.

**Users tag things in order to find them again later.** Users of social bookmarking sites like del.icio.us and Furl might use those systems to discover websites and share them with others, but the primary goal of bookmarking is tagging an item so that you can find it again. Tags can later be searched or used to organize a large collection into categories in tune with the user’s own idiosyncratic mental model.

This scenario is one we usually have in mind when we talk about folksonomies. One user might tag a given photo with sailboat, while another chooses schooner, a third Miami and a fourth user might choose peaceful. Individually, the users are motivated to provide good keywords, and with enough users the folksonomy should become fairly robust.

Providing good keywords is not, however, the only motivation for users...
to tag. My guess is that this particular case would likely result in folksonomies that are effective in information retrieval, so long as there are enough users participating.

Users tag things to get exposure and traffic. Content producers submitting items to a collection will tag their items so that users browsing through or searching the folksonomy will ultimately see their content. Bloggers, for example, tag stories with an eye toward showing up in Technorati searches.

In this case, tags are not completely different from the meta keywords that were once used by search engines in the early days of the web. Does this similarity necessarily mean tags submitted by authors will inevitably become filled with spam? Not necessarily. It certainly seems logical to prohibit abuse behaviors such as registering large numbers of fake user accounts to push up rankings or adding excessive numbers of keywords. My guess is that most folksonomies are not harmed when creators submit their items and tag them with a sane number of keywords, especially if the items would be virtually invisible otherwise.

They key difference between most folksonomies and the old, abused meta keywords system is that users can mitigate the impact of spam tags by providing their own, more relevant tags. Since most sites will have many users tagging for every content producer, the spam tags should be pushed down over time. Also, abusive tagging can lead to consequences for the spammer. This potential behavior might not be a concern at all if users are trusted, as in an intranet setting.

Users tag things as a way of voicing their opinions. Many social news aggregation sites like Digg and Reddit have voting or rating built directly into their interfaces, but when a system only includes tagging, users may provide tags that are a judgment of the content rather than a description.

Some sites like Slashdot even encourage this practice. Tags like slownewsday and fud regularly appear, meant by users to denote that a story is not very notable or as an attempt to sow fear, uncertainty and doubt, respectively.

It is a little hard to see how a folksonomy built from these tags would be effective in information retrieval. How often do people search for “awesome” or “finally”? In this situation an information architect might recommend adding a rating system, since users are obviously interested in that kind of interaction.

Keep in mind that qualitative tagging is a perfectly valid activity, so long as it aligns with the goals of the site. A humor site might get much more value from a folksonomy generated from collaborative sarcasm than it would from a very accurate system of classification. These tags might even help IR in some ways, perhaps by enhancing the information scent of certain items. A search for “Firefox” might turn up thousands of results, but the results also tagged cool might seem like better paths to pursue than those tagged lame.

Users tag things incidentally as they perform other IR tasks. Some users may be tagging when they do not even know it. As part of an independent study project, I created a website called Mealographer that allows users to track the nutrition in their diets. In the course of evaluating the usability of the site, I found that the text of the USDA database of foods did not provide very good results for users’ full-text searches. I could, however, capture the text of the searches and then associate that text with the food item the user ultimately decided on. Users were tagging items implicitly when they searched and then found an item that satisfied their search.

This method can help in some situations – users searching for “salad” might see salad dressings, potato salads and similar items ranked highest. Once a user realizes that salad is too general a term, a search like “salad with lettuce and carrots” would allow common salad constituents to climb quickly through the ranks as later users searched for “salad” and then chose lettuce off the list.

I did find a number of possible drawbacks to this method. For one, if the item’s current description does not intersect with common search terms at all, it might never come up in search results and then never be tagged. For example, the entry for cola would never show up on searches for “Pepsi” or “Coke.” Over time the entry might rack up tags like “Coca Cola” or “RC Cola” but that still leaves a lot of Pepsi drinkers frustrated. The best technique to mitigate this issue I found was to cheat – to manually tag the items myself.
Users tag things to take advantage of functionality built on top of a folksonomy. This can be the case for either content producers or users. While some blog writers may employ Technorati tags primarily to gain exposure, Technorati also allows bloggers to tag their posts to create links to similar articles on other sites. Some plug-ins to popular blogging software like WordPress employ tags to generate lists of related articles from within the site. In recommendation systems like StumbleUpon, users may be motivated to rate or tag websites in order to get more interesting, targeted suggestions.

As folksonomies become more integrated into the functionality of websites this motivation may become more and more important. If content producers and users are motivated to tag in order to get related items, the resulting folksonomies could support IR quite well. It would be interesting to see if this would result in less varied tags – perhaps bloggers would realize that more posts are tagged with IA than information architecture or vice versa and change their tagging accordingly.

Users tag things to play a game or earn points. Although this application is not very common, it is a very interesting one. One great example is Luis von Ahn’s ESP game. In the game, users are presented with images pulled from the web and asked to guess the same keywords as another user. If the two match, they get points and move on to the next image. Some sites already reward users for forum posts or other user-generated content, and it would be quite possible to do the same for users that tag prodigiously.

The resulting folksonomies may or may not support IR as well as the other scenarios already discussed. Users primarily trying to earn points might submit a lot of tags that are just good enough to not get them banned from the site. Community-driven sites could use additional functionality such as user ratings to reward quality tags and contributions to ameliorate this problem.

My experience in playing the ESP game was that my tagging behaviors changed as I tried to rack up more points in the game. I began using shorter words and simpler concepts to try to get more matches under the time limit, and I found myself not going with my first instinct and instead entering words I thought other players would use. Since I found that others were typing “sky” if the sky was at all visible in a photo, I began tagging the same way, whether or not it seemed like a striking or important part of the image. Of course more study is needed, and these sorts of tags might be exactly what are needed for certain applications. When presented with a large and unorganized collection, games like this may just be the fastest and least expensive way to get a large amount of classification done.

Aligning user motivations with site goals

The discussion of uses for folksonomies and the list of motivations for tagging are far from complete, but the value of taking some time to match up site goals with user goals is already clear. It is also clear that information science researchers have a lot of work to do in the next few years to get an empirical grasp on the strengths and weaknesses of all the variations we have seen and the new ones being created in a dorm room as we speak.

Resources

Websites Mentioned in the Article
Information Architecture and Knowledge Management program at Kent State University – http://iakm.kent.edu/
Slashdot tagging FAQ – http://slashdot.org/faq/tags.shtml
del.icio.us – http://del.icio.us/
Digg – www.digg.com
ESPGame – www.espgame.org/
Flickr – www.flickr.com/
Furl – www.furl.net/
Mealographer – www.mealographer.com/
Reddit – http://reddit.com/
Slashdo – http://slashdot.org/
StumbleUpon – www.stumbleupon.com/
Technorati – www.technorati.com/
Trouble in Paradise: Conflict Management and Resolution in Social Classification Environments

by Chris Landbeck

The advent of personal computers and the Internet has also brought about social indexing groups on a scale never before seen. These folksonomies organize information with little regard for the methods used in library or information sciences. The collaborative nature of folksonomies almost guarantees conflict will arise among some members of the community. A conflict can manifest in several ways:

- It can be an honest disagreement between two community members about the truthfulness of a given comment or whether it is appropriate.
- It can be an unreasonable demand on the community as a whole on the part of one member.
- It can come from a troublemaker, looking to sow the seeds of discontent in the community for nothing more than entertainment purposes.

If these groups and their sponsors wish their communities to persist, they must protect themselves from the potentially negative consequences of such conflicts by providing some method for resolving them.

What conflict resolution methods exist in these communities? What lessons can we take from their examples as both organizations and as collections of data? One might speculate that the advent of new rules in these communities, whether slowly emergent or suddenly imposed, will affect aspects of the organization – how it is perceived by its members, how it operates and how it gathers and organizes information. Such rules either mandate or forbid certain actions by the members of the respective societies, and therefore we might expect them to directly affect their operations. By studying rules and regulations of these groups as they affect communication among members and the information they gather, we might find out how the rules develop, what the intent of the rules is from the sponsor’s perspective and both the effectiveness and pertinence of these rules in successful social classification communities. In this article we will examine certain conflicts that arise among members of a community or between a portion of the community and the people who run it. However, copyright and other legal issues are beyond the scope of this work, as are standards for determining truth or shaping community standards for decency or other related matters.

To this end, we look at three very different organizations: wikipedia.org (an attempt at creating an encyclopedia of general knowledge), flickr.com (a collection of personal images with descriptive tags) and espgame.org (an effort to tag all the images of the Internet). These organizations were chosen because they are representative of large, ongoing, noteworthy and noted social classification efforts. Note that the content of the various collections is not being studied because the data therein are, for the purposes of this work, irrelevant. We are not looking at the content so much as we are looking at how the stuff of the community is organized. For this study, we analyzed the pages from each website speaking to the organization’s efforts to manage conflict among its users and counted them as an indication of the level of development of each conflict resolution system. The content of these pages was also analyzed to determine their nature.

Wikipedia (49 official policies: 41 governing behavior including 7 governing conflict)

Wikipedia concerns itself with building a free online encyclopedia that anyone can contribute to or comment on. The company that runs Wikipedia does not resell the data contained in the community, nor does it profit by the

Chris Landbeck is a doctoral student in the College of Information, Florida State University. He can be reached by email at clandbeck-at-fsu.edu
work of its members: Wikipedia runs on donations. Facts are established by consensus, and any and all subjects of interest are allowed in this community. It is easily the largest social classification community extant today and has the most developed system of conflict resolution. It has 49 separate official policies that speak to the operation of the community, by which is meant the guidelines that spell out how the various essential components of the community function – how articles should be written, how corrections should be made, how members should deal with one another, how citations should be made and so forth. Of these 49 policies, 41 deal with what is and what is not appropriate behavior – either while acting alone (writing or correcting articles) or when interacting with others. Of those 41 policies, seven deal exclusively with what to do when conflicts arise, whether it is between two sincere people or when dealing with some less-than-scrupulous actor within the community. This social classification environment has a linear, hierarchical progression of steps to take for dealing with disagreement and personnel designated as resources for parties that find themselves in a dispute.

More than any of the other organizations studied here Wikipedia.org also has a highly developed vocabulary for those who “game the system” – those who work against the stated principles and intent of the community. Wikipedia is vulnerable to such behavior and, therefore, experiences this type of conflict, for which there is no easy solution. Because of the way the community is set up, one side in this conflict can, for instance, create “sock puppets,” fictional user accounts that one side uses to deluge the other with responses in the survey phase of a resolution exercise. If real people engage in this behavior (say, a boss to his employees) they are then called “meat puppets.” “Trolls” are those who are disruptive to the point of being a nuisance, while those who have been banned from adding to Wikipedia and have somehow circumvented the block are said to be “reincarnated.”

These distinctions point to a system of conflict resolution, one based on the collective experience of the community in question, which is more highly developed than most. Indeed, the progression of steps recommended to community members is clearly spelled out, is linear in nature and includes step-by-step instructions for the members to follow (see Figure 1).

At all levels, these recommendations consistently point to following the editorial guidelines such as neutral point-of-view, references and leaving one’s ego at the door as the first and best way to avoid conflict. Compared to the other communities studied here, Wikipedia is easily the most complete and fully formed system of resolution. Furthermore, there is some anecdotal evidence in Wikipedia itself that suggests that the rules evolved as they were needed, that they were not the product of forethought but rather of necessity.

When we consider the objective of Wikipedia, we see that it invites discussion and argument as a means of building consensus on any and all contentious issues. Anyone can initiate a disagreement on any subject, at which time a resolution hierarchy should (but does not always) come into play. Wikipedia seems to be seeking what might be termed “truth” or “fact” or “reality,” recognizing only the authority of consensus, which might explain the nature of the conflicts found in that community. According to the record, the nature of the information to be collected was the first and greatest concern, and the conflict resolution methods that have developed over time reflect that continued emphasis. For Wikipedia, it appears that the chicken of “what should be done” came before the egg of “how it should be accomplished.” However, what Wikipedia says about itself is only one source of information, and we should be aware that it might have its biases.
Flickr (25 official policies: 3 governing behavior including 1 governing conflict)

Flickr’s approach to conflict resolution was built directly into the system from the outset, although we cannot be certain if this was a matter of planning or of serendipity. Flickr allows community members to post their images to the web while both the owner of the image and other members tag the images in question with free-form descriptors. In this way, Yahoo (the owner of Flickr) gathers several tags per image, providing an online image-sharing community in exchange. Conflict resolution comes into play when two members disagree on what tags should be applied to the images. In every case, the owner of the image – the person who posted the image for community tagging – is the ultimate and final arbiter of who may apply tags to their images, what tags will be allowed, who sees the images and who might comment on them in sentence or paragraph format. In the rare cases where someone gains alternate identities to circumvent being blocked and simply will not leave an image owner alone, Flickr will get involved, identifying and blocking the offender using more sophisticated methods than those available to the image owner. While conflict is possible in Flickr, it is far more difficult to sustain than in Wikipedia because, in the end, the members can pick and choose whom they associate with.

In Flickr, the person who posts an image is the owner of it, which gives that person the right to veto others’ descriptions of their images. In that organization, the owner reigns supreme in any conflicts that may arise. Flickr seems to place equal value on social indexing and community building, realizing a profit from the gathering of images and maintaining the community as a way to ensure the continuous flow of tags into the system. With Flickr, no information about its development from concept to reality was found on the site. One might speculate that both the nature of the data to be collected and the definition of authority were conceived and implemented at the same time, differing from Wikipedia in that the chicken and the egg arrived simultaneously. If this hypothesis is true, the founders of Flickr must have known that the pre-existing conditions of authority would necessarily affect the nature and quality of the data to be collected and have made a decision that this structure would meet the needs of both the sponsoring organization and the community members themselves.

ESPGame: (Five sections, one governing behavior and conflict)

When playing the ESP Game, two users view the same image for a set amount of time while both give their impressions of what tags best apply to the image. When players offer the same tag, points are awarded – the faster a common term is found, the more points are given. In this milieu, there is very little chance for conflict. It is not possible to communicate directly within the game, and there is no opportunity to exchange information at all as this would defeat the purpose of the game. The structure of the community is such that conflict is precluded, yet ESPGame remains a social classification community in that two people are brought together to examine the aboutness of the images in question.

At ESPGame, no conflict is possible (regardless of any disagreements), as the two people randomly paired to blindly assign mutual tags to images are both anonymous and isolated from one another. ESPGame exchanges simple entertainment for image tags. Indeed, in this arena, the website holds the data to be far more important than any potential community, even though it hopes that the community will value the game enough to keep coming back, supplying more tags for more images. While we might again speculate about
the whys and wherefores of its creation, there is even less to go on here than there was for Flickr, which has a simple but complete explanation of what it will do with the data it collects and is reasonably transparent about its creation and implementation. ESPGame has none of this and indeed very little at all about its purpose. Here, it seems, there is no chicken, only the egg of data collection.

Applying These Examples in the Workplace

How might we, as information professionals, seek to use these examples as models in the workplace? What lessons can we take from those who have pioneered the realm of social classification?

Looking at the Wikipedia model, let us suppose that there is a local historical society with ongoing efforts to catalog the landmark buildings in the area. Toward this end, they might seek to collect and electronically publish legal records about the history of ownership of a particular building, newspaper articles from the archives of local papers, construction records and oral histories from residents and former residents of the buildings being written about. The members of the community would seek to exchange ideas about where information might be found, difficulties or successes in forwarding the effort of the group and other details about what has been found and things that might yet be found.

As this community seeks to create and maintain a record of its activities and discoveries, the local library might be asked to administer and host a wiki for them, as the wiki technology and approach would seem to be the best method for this application, since it allows community members to easily record, store, publish and update records and to discuss the accuracy or appropriateness of records within the community. But the very capabilities that are needed would allow conflict to enter into the equation – the tools that allow the work of the group to go forward can sometimes be used to pull it back. Generally, the members will take care of honest conflicts themselves – serious discussion of issues of mutual interest are the bread and butter of wikis, and amiable (or at least civil) disagreements are actively sought in such environments. However, as the administrator of such a community, the librarian might be asked to mediate conflicts between members when they reach an impasse, either as an impartial mediator or as an information expert. This librarian might also be called on to deal with graffiti, with malicious members of the wiki or with other incidental problems that come with the territory that members might not wish to deal with. It is sometimes easy to forget that wikipedia.org runs on donations and that those donations go toward the salaries of those who run it. This fact is a reminder that, while running the wiki for this notional historical society might not by any means be a fulltime job, it will need someone both to settle disputes and to attend to the workaday issues that arise from any community-based effort.

Another implementation might draw on the Flickr model. The library might be asked to establish a virtual scrapbook of a trip that members of a civic group took as part of its relief or charitable efforts. Here, the members could contribute their pictures to a central location that would allow those who went on the trip to post comments, stories and descriptors of what each picture is about. In this application, the librarian has a different set of concerns. As the administrator of the historical website, the librarian was concerned with developing records; now she is being asked to help manage the records of others. She may be asked to assume an active role, serving both as administrator of all things technical and perhaps as the point-of-contact for the scrapbook.

Adopting the Flickr approach up front – that the person contributing each photograph can accept or reject the comments on it – could provide an appropriate rule for settling disagreements. However, the librarian, as the administrator, might also be asked to serve as the de facto help desk for such an effort, answering how-to questions from less tech-savvy users. She might also be sought in her traditional role of librarian to help pin down facts about the pictures. How often have we ourselves taken a picture then forgotten where it was taken and in what way the thing in the picture is important? While the nature of the possible conflict in this social classification environment is different, it is no less a conflict that can disrupt an otherwise happy community.
Finally, in a different vein, a librarian might think it would encourage library use to set up something like the ESP Game, not for images but for, say, children’s books. She might encourage two patrons who have checked out a book to provide terms that describe the book. She might then gather together these terms, rank their frequency and put the staff in a better position to recommend (or not) books based on what others have said about them, possibly even integrating this data into an online catalog. As noted, there is little chance for conflict to arise here, as the way the system is built prevents it. But such an effort might provide a new channel for old conflicts about the appropriateness of certain materials.

Conclusions

In considering which model to adopt, three general questions should be considered:

1. What do you want your community to be able to do?
2. What kind of data should be collected?
3. What kind of time and resources can be given to the project?

We see in the examples above a wide range of possible activities that users might need or want to engage in. Wikipedia keeps a long history of article revisions and allows for a great deal of interaction, depending on the goodness of its user’s souls to keep things aboveboard and honest. Flickr has placed a great amount of somewhat authoritarian control in the hands of its image owners, again depending on the beneficence of its user base, while ESPGame has taken away the forum for conflict, stopping it in its tracks. Each has different possibilities for interaction among users in their respective communities. What level would you like to see in yours? What is needed? And how do these approaches help or hinder your users in their needs and preferences?

In a similar vein, Wikipedia deals only with full thoughts expressed in (what is hoped to be) clear, concise and complete paragraphs as its finished product; Flickr only deals with images and the associated words or phrases; and ESPGame deals only with words as tags for images. Each of these is a different kind of data to collect, and each has its own ethical and social concerns to deal with. What kind of data is necessary for your community? What kind of issues can be dealt with in advance (and how will they be dealt with)? And, most importantly, who owns the data, and is the owner different from the keeper?

The thread that runs through all of this discussion is the most burning question to ask before implementing any folksonomy: What are you willing to do to make it work for your user community? What kind of skills can you and your organization bring to the task? What kinds of hassles and conflicts are you willing to settle? And who in your organization is best suited to oversee such an endeavor? In each of the examples presented here there is a way to contact the ones who run the community, hear complaints, listen to suggestions or provide help to those in need. While there is necessarily a greater (Wikipedia) or lesser (ESPGame) need for supervision, there is nonetheless someone in charge. Who that is for your organization is essential to the success of the community.

Resources


Image Indexing:  
How Can I Find a Nice Pair of Italian Shoes?  
by Elaine Ménard

From time immemorial the image has been a communication tool. Images have a multifaceted and very real value. Images play a double role. They serve not only as sources of information but, with the development of more accurate visualization techniques, they also enhance the understanding of that information. In recent years the diffusion of images has increased, mainly because of the development of digital technologies and the unprecedented growth of the World Wide Web. It would be impossible to enumerate all the types of images we encounter on the web; however, we can mention visual collections (paintings, prints, engravings, illustrations), drawings, charts, postcards, photographs (historical, botanical, police, medical, documentary, personal and familial, artistic), as well as images generated by computers, among many examples. Actually, the digital image is an integral part of our daily reality. We now scrutinize image search engines in the same way we consulted encyclopedias or illustrated dictionaries a few years ago. There is nothing easier than to discover a close up of Paris Hilton’s latest arrest or a picture of her dog Tinkerbell on the web. Students will search image databases to illustrate a research paper, potential travelers will check travel resorts to have a better idea of the destination of their dreams, sportswriters will enrich their articles with evocative pictures of Maria Sharapova and so on. Google claims in 2007 that their users now have full access to more than 2 billion images.

Confronted with this profusion, individuals now speculate on how to retrieve images with efficacy and efficiency. In general two categories of queries are used to retrieve images on the web: graphic queries or textual queries. In the first category, the individual submits a graphic query (using an image or a drawing), and the system tries to retrieve a similar image by using certain physical characteristics of the image such as color, shape or texture. However, since such content-based-image retrieval systems (CBIR systems) have many limitations, the majority of image searches on the web still use textual queries, and the retrieval’s success depends on the match between the query terms and the text (ancillary text or indexing terms) associated with the images.

Consequently, to retrieve an image the individual must first translate into words what he is looking for. This conversion is the first challenge. The second obstacle comes from the “language” of the image. By their very nature, images are considered to be language-independent resources. Nevertheless, the text associated with the images gives the image a linguistic status similar to any other textual document, which can significantly affect its retrieval. And given the great linguistic diversity existing on the web, we must expect that the text associated with images exists in many different languages. For example, if a user formulates a query in English, and the images to be retrieved are associated with English text, the cross-lingual problem does not arise. However, if an English query is used, and the associated text is Italian (or any other language different from the query language), the retrieval will not be possible unless the retrieval system includes a cross-language information retrieval (CLIR) mechanism which allows cross-language mapping between the query terms and the associated text.

Elaine Ménard is a Ph.D. candidate in the École de bibliothéconomie et des sciences de l'information at Université de Montréal; C.P. 6128, succ. Centre-ville, Montréal (QC) H3C 3J7, Canada. She can be reached by email at elaine.menard<at>umontreal.ca; or at her website at www.gin-ebsi.umontreal.ca/menarde/
Image Indexing: The Case of the Ordinary Image

For many years a great deal of work has been devoted to visual resources such as the image. We usually distinguish three main categories of images: the artistic image, the documentary image and the ordinary image. The artistic image, which is defined as the representation of an artistic expression, is characterized by an elaborate documentary process that includes several levels of significance (pre-iconographic, iconographic and iconological). The documentary image is mainly found in historical files, news services and media files. Family photographs also belong to this second category. Processing this type of image consists primarily of adding legends or information whose objective is to identify the image within a specific collection. Finally, the ordinary image is the image generally being used for commercial purposes or as an illustration. When processing this particular type of image, we generally take little account of such descriptive or analytical metadata as the title or the author, but will rather consider the visual content of the image, that is, its subject. In other words, the ordinary image does not require in-depth processing. A categorization by class such as “animal” or “landscape” could be enough. Nevertheless the advent of the web has highlighted the pressing need to acquire suitable tools for describing ordinary images, since we find them to be the majority of the resources available on the web – personal pages, blogs, virtual libraries, museum collections, services and product catalogues, and governmental information all often fall into this category.

Recently the needs and behaviors of image searchers have changed considerably. We actually perceive an evolution in the manner of formulating the queries for image retrieval. For example, queries containing a single term are less frequent than they were when web searching began. Individuals tend to use more sophisticated search strategies while including more and more search elements such as proper names, trademarks or colors, in their queries. Image searchers also develop queries containing the relationships between these elements. Consequently, if the manner of searching images is evolving gradually, maybe it is time to consider whether the image indexing methods, and more particularly the controlled vocabulary traditionally employed for the indexing process, is still well adapted to the real and current needs and behaviors of image searchers.

The Traditional Way

When users pose textual queries, the success of the retrieval largely depends on the correspondence between the query and the text associated with the images. Since images do not always include a caption or any kind of ancillary text, the indexing process remains crucial. Image indexing has, so far, been divided between two camps: those who concentrate on controlled vocabulary and those who concentrate on uncontrolled vocabulary. The former focuses on assigning index terms extracted from thesauri, classification schemes or subject heading lists, while the latter focuses on terms drawn from natural language.

Over the years several excellent controlled vocabularies have been developed, varying in scope from the general to the very specific. The main purposes of these vocabularies are to help cataloguers find the right term to describe an object, to categorize concepts into broad topics and to generally improve the retrieval. For example, the Art and Architecture Thesaurus (AAT) is a controlled vocabulary intended for indexing physical objects. The AAT provides preferred terms for concepts related to art, architecture, decorative arts, material culture and archival materials. This vocabulary is largely used in museums, libraries, archives, visual resource collections and conservation agencies. Most terms included in the AAT are in English, but terms coming from other languages are progressively being incorporated in that vocabulary.

Two other examples of thesauri used for image description are the Thesaurus for Graphic Materials I and the Thesaurus for Graphic Materials II (TGM I and TGM II) created by the Library of Congress Prints and Photographs Division to support both cataloguing and retrieval needs. The TGM I is mostly used for subject indexing of graphical materials, including historical photographs, architectural drawings and artwork. Available only in English, the TGM I does not include proper names of people, organizations, events or geographic places. As for the TGM II (also available in English), it was created as a complement to the TGM I. This thesaurus provides headings
for categories of material by genre, vantage point, representation method, production technique, marking, shape or size, purpose, characteristics of the image’s creator or publication status.

Another popular vocabulary for image description is ICONCLASS, which provides a different perspective. This subject-specific classification scheme is designed for the description and classification of visual resources collections. ICONCLASS includes a hierarchically ordered collection of definitions of objects, persons, events and abstract ideas that can be the subject of an image represented in various media such as paintings, drawings or photographs. Primary users of that classification scheme include art historians, researchers and museum curators. For the moment this classification scheme only exists in English, but translations in other languages are currently in progress.

The use of controlled vocabularies for image indexing offers many advantages for retrieval, browsing and interoperability. The control offered by these vocabularies is manifold. It manages the use of synonyms, homonyms, lexical anomalies and so on. However, one of the main disadvantages of controlled vocabularies is that they quickly become outdated. For example, neologisms will often take a long time to appear in controlled vocabularies. As a result, the search will be less accurate because controlled vocabulary will sometimes not allow a specific search. Furthermore, we must consider that the development and the management of controlled vocabularies involve significant costs. But the main difficulties associated with the use of controlled vocabularies for image indexing can be summarized as follows:

1. These vocabularies are not suitable for all image types and certainly not for the majority of ordinary images.
2. The use of the majority of controlled vocabularies is beyond the capacity of the non-expert or less trained professional.
3. Most of the controlled vocabularies exist only in one language (in most cases, in English), which implies that it will not be a great help in all linguistic contexts.

So what can we do to overcome the limitations of controlled vocabularies? Well, it seems that web users have created their personal solutions by using their own methods to index images as they do with collaborative tagging, the latest trend in image indexing.

**The Latest Trend**

Collaborative tagging has recently become very popular on many web services such as del.icio.us or CiteULike. Collaborative tagging began with any user assigning his or her own keywords to textual documents, but the same phenomenon was quickly implemented for image resources. Collaborative tagging is now the pillar on which photo sharing sites like Flickr rest. These sites allow massive image storage and web diffusion. In these systems, users upload their own images and index them using their own terms (tags). It is also possible to make these images public; that is, the images can be seen by all users or by a group of people chosen by the system user, thus forming a vast and communal image database. In a Flickr-style system, the user who uploads images can thus determine who will have access to these images by stating certain rules of access control. In parallel other users of the system have the possibility to update the image indexing by adding other keywords or comments to any image they have access to. These annotations assigned by the uploader or by any other user of the Flickr system constitute a form of free indexing. It is this free indexing that is called “collaborative tagging.” Obviously, this kind of indexing supposes that the individuals use their own words to describe images. Of course, they could choose to index with a controlled vocabulary, but why bother? Consequently, tags assigned by collaborative indexing generally contain a single term (for example, *house*, *Christmas*, *Lassie*). However, tags sometimes tend to be more descriptive (for example, “covered cat litter box,” “SportRack bicycle rack,” “black down filled jacket”). Instinctively, the users seem to include what they think is significant and imperative to employ if someone else needs to retrieve an image.

Collaborative tagging may therefore seem very seductive because of its close relationship with the real users and the way they see and describe things. Moreover, neologisms and all forms of newly created terms are quickly
integrated in collaborative indexing. Compared to controlled vocabularies and especially for new topics, collaborative indexing is likely to win hands down since the same words could take months, maybe years, before they are even considered for inclusion in a controlled vocabulary. Besides, in these image tagging systems, the indexing can be done in one language or in a combination several languages, which can ease the user’s retrieval problem. However, despite its growing popularity and much like indexing with controlled vocabulary, collaborative tagging also presents several gaps. For example, some ambiguities emerge because the same keyword is often employed by several individuals, but in various contexts. In the same vein, the lack of synonym control results in the use of many different keywords to describe the same concept. Consequently, free indexing is often considered to be of poor quality.

To illustrate the difference between controlled and uncontrolled vocabularies used for image indexing, a database of ordinary images including 3,950 ordinary images drawn from the eight sections of an online commercial catalogue was created for an in-progress study. Each image was indexed in four different ways (see Figure 1) – with controlled vocabulary (French and English) and with uncontrolled vocabulary (French and English). The indexing process using French controlled vocabulary was carried out by a French-speaking indexer, while the indexing using English uncontrolled vocabulary was carried out by an English-speaking indexer. In order to reproduce the conditions of collaborative tagging, no fixed directive was given to the indexers concerning the number or the form of indexing terms to use for the image’s description. For the controlled vocabulary indexing, a bilingual indexer used the *Nouveau dictionnaire visuel multilingue*. This dictionary contains appropriate terms (French and English) for the type of images contained in the database and offers a form of standardization of the terms that allows a clear and precise identification of the objects while exerting maximum control on word variations.

Following the indexing process, the assigned indexing terms were examined. The objective of this analysis was to identify the specific characteristics of each indexing approach. A grid analysis was developed and applied to the complete set of indexing terms. Three levels of analysis were carried out: terminological, perceptual and interpretative. Preliminary results of this analysis revealed several similarities and differences between the two kinds of indexing.

First, we notice that in most cases, the indexers assigned only one indexing term to the particular type of image we found in this database. Second, among these indexing terms, we observed that uncontrolled vocabularies tended to use indexing terms containing multiple words (84% for French and 94% for English) where controlled vocabularies have a propensity to use uniterms (55% for French and 54% for English). However, perhaps the biggest difference emerging from this partial analysis is that uncontrolled indexing terms often include words referring to size, color, texture, gender or trademarks, contrary to controlled vocabularies which have a tendency to be less graphic and, in many cases, less detailed or descriptive. An extensive examination of the whole database is still in progress and should uncover other significant features differentiating the indexing terms assigned from controlled and uncontrolled vocabularies.

**Conclusion**

The preceding analysis could be a good indication of how ordinary images really need to be retrieved. Furthermore, we must take into account that individuals may be interested in finding images indexed in languages other than English. For example, web users may be interested in buying objects from different parts of the world. And before they do so it is probable that they will want to see an image of the object. But they will have to retrieve the image first using some description of the object. As a result, image indexing remains a problem.
Whatever its virtues, however, there is no reason to consider collaborative tagging as a replacement solution to traditional indexing with controlled vocabulary. The reaction of some veteran information professionals to collaborative tagging is interesting. When this “new” indexing reality is mentioned, some look as if they had just ingested very bad medicine. Of course when you think about it, collaborative tagging is somewhat threatening for information specialists. In fact, there is not really another field that provides an example of how they should approach this usurpation of an important part of their work. We do not encounter this kind of phenomenon in other domains where anybody can walk in and say, “Hey, we think we can do a better job than you do!” It is difficult to imagine a lawyer or a physician confronted with a group of individuals coming up with their own solutions to solve a legal case or a medical diagnosis. But willingly or not, information specialists are now confronted by this kind of rivalry. There is certainly a valuable lesson to learn from collaborative tagging. According to circumstances, these two approaches may co-exist and be very helpful. In the near future we could see more and more information systems allowing the co-existence of controlled vocabularies and uncontrolled vocabularies resulting from collaborative image tagging. But would it be effective, especially if we consider all the images that are now available to the image searcher? We will probably find out soon enough.

---

**Resources**

**Websites Mentioned in the Article**
- CiteULike – www.citeulike.org/
- del.icio.us – http://del.icio.us/
- Flickr – www.flickr.com/

**Indexing Tools Mentioned in the Article**
- Art and Architecture Thesaurus –
  www.getty.edu/research/conducting_research/vocabularies/aat/
- ICONCLASS – www.iconclass.nl/
- Nouveau dictionnaire visuel multilingue –
  www.ledictionnairevisuel.com/langues/index.htm
- Thesaurus for Graphic Materials I – www.loc.gov/rr/print/tgm1/
- Thesaurus for Graphic Materials II – www.loc.gov/rr/print/tgm2/

---

**Websites Mentioned in the Article**

- CiteULike – www.citeulike.org/
- del.icio.us – http://del.icio.us/
- Flickr – www.flickr.com/
The development and subsequent popularity of image tagging at sites such as Flickr (www.flickr.org) has been a phenomenon to receive considerable attention over the course of the last few years. After spending more than a decade cataloging and providing access to images in an academic setting, I, too, felt compelled to take a look at what this Web 2.0 image-sharing site had to offer. There were two interconnected ideas at play when I began thinking about performing a study of Flickr tagging. The first of these ideas had to do with looking for an underlying pattern for the image tags. It seemed likely that some commonalities would occur among what at first glance appeared to be the chaos of personally applied image tags. Finding these common patterns among the tags would clarify what types of information people typically associate with their images. The second idea concerned the effectiveness of image tagging. If patterns were discovered among the image tags, I believed these could be used to alleviate some of the problems associated with tagging. So early in 2006 with these two ideas fresh in mind, I carried out a small study of the image tags used at Flickr.com.

To conduct the study I gathered the top 10 image tags of 14 randomly chosen Flickr users and downloaded them through the site’s open APIs. In order to discern if there were patterns to be found in the application of the image tags, I applied conceptual labels to each of the 140 image tags. This labeling process was iterative, and after several passes through the entire set of image tags a model consisting of 18 categories (Table 1) emerged. During this process it became apparent that image tags could have multiple meanings and as a result, I allowed some to be assigned to several categories. Thus, a tag such as cross could simultaneously be considered a verb, a thing, an emotion and an adjective.

In order to evaluate the usefulness of the model to represent the various image tags, I gave the files of the extracted Flickr image tags and the categories with definitions to four people, who then categorized the image tags in a way that made most sense to them. Their categorizations of the image tags were combined with mine and the occurrences were then tallied to determine the overall category agreement (Table 2) and patterns of category usage (Table 3).

---

**Table 1. Category model for image tags found in Flickr.**

<table>
<thead>
<tr>
<th>Category</th>
<th>Definition</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjective</td>
<td>All adjectives</td>
<td>cold, wet, bright</td>
</tr>
<tr>
<td>Compound</td>
<td>Terms with two or more words combined</td>
<td>Newyorkcity, Mydog</td>
</tr>
<tr>
<td>Emotion</td>
<td>Identification of emotional state</td>
<td>happy, depressed</td>
</tr>
<tr>
<td>Event</td>
<td>Pertaining to holidays, happenings or news occurrences</td>
<td>wedding, Easter, assassination</td>
</tr>
<tr>
<td>Humor</td>
<td>Terms used for humorous reasons</td>
<td>Ithinkbobbyisgreat</td>
</tr>
<tr>
<td>Language</td>
<td>Terms in any language beyond English</td>
<td>eau, gefühle, madrina</td>
</tr>
<tr>
<td>Living Thing</td>
<td>Living, non-human creatures and plants</td>
<td>bird, rose, tree, dog</td>
</tr>
<tr>
<td>Number</td>
<td>Terms composed of numbers</td>
<td>64325, 1+111, 2000</td>
</tr>
<tr>
<td>Person</td>
<td>Named (common and proper) individuals and groups</td>
<td>baby, Elvis Costello, Girl Scouts, woman</td>
</tr>
<tr>
<td>Photographic</td>
<td>Terms relating to imaging/photographic devices and,</td>
<td>Cannon, SLR, II100</td>
</tr>
<tr>
<td></td>
<td>or processes</td>
<td></td>
</tr>
<tr>
<td>Place-general</td>
<td>Places identified with their common names</td>
<td>beach, field, bedroom</td>
</tr>
<tr>
<td>Place-name</td>
<td>Places identified with their proper names</td>
<td>Amsterdam, Seoul</td>
</tr>
<tr>
<td>Poetic</td>
<td>Terms that are poetic in nature</td>
<td>heavenly mirage, daydream</td>
</tr>
<tr>
<td>Rating</td>
<td>Terms which evaluate images</td>
<td>topten, tag1, taggedout</td>
</tr>
<tr>
<td>Thing</td>
<td>Non-living objects</td>
<td>house, car, rock, water</td>
</tr>
<tr>
<td>Time</td>
<td>Terms with chronological meaning</td>
<td>June, 2006, night</td>
</tr>
<tr>
<td>Unknown</td>
<td>Unidentifiable terms</td>
<td>Sha78, Pp73</td>
</tr>
<tr>
<td>Verb</td>
<td>All verbs</td>
<td>running, look, crying</td>
</tr>
</tbody>
</table>

Joan Beaudoin is a doctoral student at Drexel University. She can be reached by email at jeg56<at>drexel.edu.
The overall category agreement of the model suggests that the model is modestly effective in describing the concepts of the tags as assigned by Flickr users. Five people agreed on 11 of the 18 categories more than 50% of the time. Several of the lower performing categories (photographic, language, rating) seem to have been the result of a lack of specialized knowledge on the part of several of the participants. For instance, language tags (Weg, tio) were sometimes placed in the category Unknown. One category, number, seems to have had an unclear meaning for several people, and several tended to use the category time for tags consisting of digits, which could account for the lowered agreement in this case. The categories of humor and poetic also performed poorly, but the reason for this situation is unclear. It seems likely that these categories of tags were the most open to individual interpretation, however. The model’s performance could be improved through modification of those categories with low percentages of agreement across participants and by providing some further instruction and examples of typical tags to be found in each of these categories. An additional means of testing the model would be to instruct the participants to choose a single, best category for each tag.

**Tag Categories – Frequency of Use**

Looking at the percentages of tag category usage it is clear there are preferences for tagging among the Flickr users. The most frequently used categories of tags by this group of Flickr users were named geographical locations (New York City, China). Compound tags (white dish, rock star glasses) were the second most frequently occurring category of tags among these Flickr users. However, unlike the other highly occurring tag categories in the model, which are descriptive in nature, the compound category is applied for reasons of form alone. With these tags the user applies a tag that combines two or more terms. The high percentage of compound tags suggests that Flickr users find single word tags to be inadequate for describing their images. The next most used categories were inanimate things (water, bottle) and people (Debbie, woman).

After these high performing categories of image tags follows a group that saw only modest usage. These tags each accounted for approximately 1-5% of the total tag use by the Flickr users. Even though their usage appears to be limited, their overall importance should not be underestimated. For example, the category event (party, wedding) appears in each of the users’ top 10 most frequently employed tags. This usage illustrates the value people place on identifying special occasions within their images. The relatively modest use of this category among the tags explored here I believe speaks more to the rarity of these events than it does to a lack of interest in tagging images with these terms. One only needs to look at the tag cloud of Flickr’s most popular tags to see the frequency with which event tags are used. While the list of most popular tags reflects those terms that are most frequently applied by the entire Flickr community it is suggestive of how common these terms are as image tags.

At the lowest end of the spectrum of tag usage, all occurring at below 1%, are the categories humor, poetic, number and emotion. These categories plainly see little usage among the top 10 tags of these Flickr users. However, it needs to be mentioned here that the full range of tags being assigned by users was not investigated. It may be that these types of tags are being employed at a
higher frequency but that they are not well represented in the users’ top 10 tag lists. Whatever the cause of this situation, the phenomenon is unclear and warrants further research.

Problems of Meaning

A further issue needing additional study is the difficulty of discerning the meaning of some image tags. Even when the participants could apply ambiguous tags to several categories, they could not characterize a high percentage (nearly 5%) of them. In some cases these unknown tags appear to have a personal meaning for the Flickr users who applied them (for example, “tg78”) and so the participants performing the categorization could not readily understand them. In other cases the unknown tags illustrate how important contextual knowledge is to the categorization of the tags. This factor was most clearly seen in the tags relating to Flickr groups and photographic devices, which were unfamiliar to several of the participants. Foreign languages and unknown geographic locations posed recognition difficulties for some participants as well, so a number of these tags also found their way into the unknown category.

Individual Differences

The tag categories discussed above give an overview of the way Flickr users tag their images as a group, but their application by single users showed some degree of variation. For example, several users had a single instance of place-name in their top 10 tag lists, while others had top 10 tag lists consisting nearly exclusively of this category of tag. The categories time and photographic also saw highly individualized usage. While tags with photographic or imaging concepts behind them were used by a majority of Flickr users at least once in their top 10 tags, for one individual this type of tag accounted for half of the top ten tags applied. The frequency of this tag suggests that most Flickr users are interested in recording the processes and devices used in the creation of their images, but for some users it was a far more important form of information. The use of time showed the most variation among the Flickr users that were studied. Eleven of the 14 users did not have a single instance of time among their top 10 tags, but three users found it to be a useful categorization, and for one user it accounted for six of the 10 top tags. The limited use of time by Flickr users is an interesting discovery since it is a frequently used organizational principle associated with the management of images. This situation is possibly due to the fact that image files commonly receive a date stamp when they are created. Therefore, some Flickr users would see this information as redundant.

Flickr Developments

Interestingly, over the course of the year that has passed since I performed the tagging investigation Flickr has implemented several features that support many of the categories of information this study found to be important to users. When files are uploaded to the site the date stamp (partially representing the information stored in the category time) now automatically displays alongside the image. In addition, the device used (partially representing the information stored in the category photographic) to create the image is also automatically recorded as the file is uploaded to the site. Place-name, the category that saw the highest frequency of use, is now partially accommodated through the use of geotagging, and as a result a named geographic location is now displayed alongside the image. To take advantage of this feature users place their images on a map manually, or they use an application written for camera phones which records the geographic coordinates of the location of the cell phone’s tower for photographs uploaded to Flickr. Photographs that have been geotagged using either of these methods can be viewed on a map launched through Flickr that can also display the locations for other nearby photographs that have been uploaded to the site. Three other categories that were highly used, thing, person and event, haven’t seen similar support from Flickr. However, the tag category compound has seen the development of a technique that now allows for multipart tags to display as separate words in the user’s tag list. By inserting double quotation marks around the entire tag, users can cause compound tags to display as they are entered (for example, Father Time instead of fathertime). Each of these improvements to the Flickr site offers additional...
means by which to record information about images and in several cases these require little or no effort on the part of the user. Not all Flickr users apply tags to their images, and so in these cases this automatically recorded information may be the only means of access beyond visual browsing.

Implications for Users

The model revealed several important aspects about the tagging behaviors within the Flickr community. The identification of the types of information being recorded for images is obviously a useful step in helping to develop more effective methods of tagging. Flickr has been hard at work in this regard, as was noted above. In addition to streamlining the tagging process, the automatic ingest of information that happens with image uploads to Flickr deepens the information pool and removes human error. Further developments could be implemented to decrease the effort involved in tagging. The most basic of these is reducing the cognitive load associated with the task. A basic schema developed from the model could be employed to prompt individuals to enter tags rather than trying to choose the “right” words to represent the image. Associating tags with conceptual headings would also prove useful for clarifying meaning (that is, does the tag *cross* signify the verb, the thing or the adjective). Tags using conceptual headings could have thesauri associated with them to facilitate the choice of additional tags. This area is one I think would see heavy use by Flickr taggers. One of the most interesting aspects of tagging witnessed at Flickr is the care some users take in applying as many tags as possible for a single concept. For example, an image may be tagged as cat, kitten, feline, or *felis silvestris catus* in addition to adding tags in multiple languages: chat, gato, gatto, Katze. In addition, each of these tags is often entered again in its plural form. An automated tool to accommodate the plural or singular form of a tag would obviously be useful in these cases. Combining tagging with a basic model enhanced with tools like those discussed above would serve to strengthen the natural language of the users’ efforts and help to increase the retrieval of relevant images.

Implications for IS Professionals

The popularity of image tagging is a testament to the effort people are willing to expend in describing their images. As information professionals we need to develop new methods and techniques to assist people in tagging and retrieving their ever-growing body of visual materials. This information will in turn inform our own image indexing practices. With a better sense of how individuals are tagging their personal images it would behoove us to offer a similar kind of information for the visual materials we provide access to in our own collections. If we are not recording similarly detailed, descriptive information concerning the places, people, things and events in our image collections we are probably not reaching the broad audience we all hope to serve. Additionally, this study should clarify for information professionals just how highly personalized image information can be among individuals. Although a general pattern of image tagging was discerned across the Flickr users, information an individual was interested in might focus on a single aspect. This discovery is indicative of just how similar visual information research needs are to text-based investigations. Visual information is as richly complex as text-based materials, but at present text is more accessible. Concurrently with the technological developments over the past decade, there has been considerable progress in image retrieval, but a great deal remains to be done. One way that we can make these improvements is by continuing to look closely at how individuals categorize their own images and to assist them in their efforts.
In the classic science fiction film *Them!* [1] the forces of good (sturdy law enforcement and geeky science) seek out and destroy the forces of evil (gigantic mutant ants) that have resulted from foolish tampering in 1945 by the forces of good with the forces of nature (atomic energy) in the desert. In a key scene, the lone survivor of an ant attack is shocked out of her comatose state via a whiff of formic acid (essence of ant) and shrieks “Them! Them! Them!” Given the criticism of Google by some in our field it is quite easy to imagine librarians as the comatose victims shrieking in terror as they relive their encounter with a gigantic monster on the information frontier. In the movie, the forces of good come to realize that, while the ants may be mutant monsters, their natural instinct to band together, propagate, spread out and forage for food remains unchanged. They have no plan to replace humans – they simply do what they do. Despite this insight, it is clear the ants must be destroyed for sake of humankind’s survival.

It is a bit extreme to paint librarians and Google locked in the same battle, but the questionable characterization by some librarians of Google as a mutant technology lamentably exists alongside those who admire it for its many innovations. From all accounts, Google has no interest in replacing libraries. They express [2] a great appreciation of libraries. “Even before we started Google, we dreamed of making the incredible breadth of information that librarians so lovingly organize searchable online,” said Larry Page, Google co-founder and president of products. However, their awareness of the library and information science (LIS) beyond that of supplying people to stock and staff book storehouses is vague at best. Their notice of the field manifests itself much like the ants, who only bother with the humans when they are presented as a possible food source.

In order to understand this ambivalence it may help to contrast Google’s search engine with traditional library classification and to understand the original intentions of Google’s creators, Brin and Page. There is a limit to what we can find out about Google’s search engine by delving into the resources about the company and its technology, but what is there provides interesting food for thought.

“Google,” John Battelle [3, p. 37] writes, “is currently our culture’s grandest declaration of the power of search – but it is by no means the first.” In his description of the development of search engines, Battelle relates the motivations and actions behind the rise, and sometimes downfall, of well-known search engines such as Google, Alta Vista, Excite and Yahoo!. In particular, he lovingly expounds on the “Database of Intentions,” which he declares to be “…the most lasting, ponderous and significant culture artifact in the history of humankind….the aggregate results of every search ever entered, every result list ever tendered, and every path taken as a result.” (p. 6) The overall focus of Battelle’s book is search (that is, web traffic), and in particular, paid search, which some librarians may see as an uncomfortable contradiction to, say, the *ALA Library Bill of Rights* [4]. We don’t sell access to our users, although we may all be secretly envious of the perks enjoyed by those who do (for instance, Google reported revenues of $3.87 billion for the quarter ending on June 30, 2007 [5]). According to the Google Book Search Help Center FAQ on the Google website [6], “There are currently no ads on books from the Library Project.” However, there are links to online books stores, but no evidence of linkage for donations to library funding.

Shawne D. Miksa is assistant professor in the School of Library and Information Sciences at University of North Texas. She can be reached by email at smiksa<at>unt.edu
This simple division of purposes is quite central to the argument of the pros and cons of Google. Sandler [7] calls Google a “disruptive technology” whose “catalyzing changes are already underway.” (p. 21) He asks, “How can we expect libraries to maintain relevance in a changed world?” (p. 21) Anderson [8] echoes this theme as he chides librarians for not taking more seriously the effect Google’s services have on the role of the traditional library. Brewster Kahle of the Open Content Alliance (OCA) contends that the “idea of making all books accessible online in new and different ways is all good news. But if you do this in a way that the materials that have been housed in libraries for centuries are made available only through one corporate interface, this is an Orwellian future.” [9] Jeanneney [10, p.71] cautions that “accessibility to everything without Ariadne’s thread to guide our curiosity may cause us to lose our way.”

Motivations Driving Google’s Work

There are plenty of resources recounting the story of Google’s origins, including many studies of the accuracy and efficiency of its growing stable of information tools (for example, comparing Google Scholar with Web of Science is currently quite popular). An interesting fact is that Brin and Page used the very simple premise of citation analysis (with the proper accolades to Eugene Garfield and citation indexing) as the basis for PageRank, the primary Google algorithm that employs the hyper-linking structure of the web to calculate the ranking of websites. (Depending on what corner in LIS one inhabits this origin may be common knowledge or a surprisingly unknown fact.) There still lingers, however, the disconnect between LIS and computer science. Langville and Meyer [11] preface their exploration of the science of search engines, including PageRank, with a short history of information retrieval and compile the following list:

Notable artifacts that belong in our information retrieval museum are a few lists of individual library holdings, sorted by title and also author, as well as examples of the Dewey decimal system (1872), the card catalog (early 1900s), microfilm (1930s), and the MARC (Machine Readable Cataloging) system (1960s). (p. 2)

An exploration of the About Google web pages, in addition to external literature written about Google, reveals that perhaps the search empire is still conflicted (or ignorant) as to whether its technology provides meaningful searches with context based on ranking or just ranking. Battelle reports “the Google service made no pretensions of actually reading a particular site, or of understanding its content. It simply laid bare the often ugly truth of how well connected a site happened to be.” (p. 80). Google’s Company Overview page [12], however, states that “Google’s mission is to organize the world’s information and make it universally accessible and useful.” We can add an exception – it is universally accessible except for where censored by certain countries that use Google (for instance, China), or when Google censors itself [13].

Ranking does provide a type of organization, but the question remains as to whether PageRank also gives a sense of context and relevance before users obtain the material or if just the fact that they obtain it is enough of a justification to consider a search successful. Brin and Page [14] write that, “our notion of ‘relevant’ is to include the very best documents since there may be tens of thousands of slightly relevant documents. This very high precision is important even at the expense of recall.” (p. 108-109) However, Milner [15] suggests

If we had a computable theory of meaning then it should be possible to get a document, such as an email, summarized by software. Currently, it cannot be done. Even Google has to do the next best thing and show you passages from the document, which may match your search. (p.193)

Library classification is one process of information organization that does not work based on the concept of popularity as employed by Google. Most classification systems are based on a concept of subject, in which like subjects are grouped together, while unlike subjects are not. Assignment of classification numbers is based on subject analysis, or content analysis, powered by human judgment as to what something is about and where in a finite collection of information objects it should be placed in relation to everything else in the collection. A classifier would thus assign a classification number to a web page based on analysis of its content.
By contrast PageRank’s main thesis, as described by Langville and Meyer, is that “a page is important if it is pointed to by other important pages.” (p.31) In addition, Google tells us [16] the following:

The engine then conducts hypertext-matching analysis to determine which pages are relevant to the specific search being conducted. By combining overall importance and query-specific relevance, Google is able to put the most relevant and reliable results first.

Google’s purpose and many projects have inspired both praise and criticism, some of which carries equal ambivalence. One of the most recent, mentioned above, is a work entitled Google and the Myth of Universal Knowledge [10] by Jean-Noël Jeanneney, president of the Bibliothèque nationale de France. Jeanneney writes that “behind its majesty, Google is hiding frailties, like any company founded on a (currently, at least) single, albeit profuse, activity, without any parent corporation or other major body to lean on.” (p. 62) He is especially critical of the Google Book Search because of potential biases toward English-only literature, as well as biases in ranking (what he terms the “gondola end”). Jeanneney asserts the following:

Due to the criteria of frequency and density of links, the pages most often recognized by the engine will in turn be more easily called up by the other users clicking on the links, and we can be sure (thanks to the principle of lending only to the rich) that the pages that are already overwhelmingly “selected” will continue to be so. (p. 45)

As a countermeasure, Jeanneney suggests that a European algorithm be defined. Google Book Search is now available in nine languages, according to the Google website. Despite that statement, there isn’t a “language” search feature visible on the Book Search interface, although searching with non-English terms will net some non-English books.

In their 2005 “Letter from the Founders,” [13] Brin and Page discussed the negative reaction to their Book Search. They write, “We believe one of the greatest services we can provide to users around the world is to increase people’s access to human knowledge” and that Google respects copyright and does not have permission to break copyright on any work. Their reasoning for the criticism from authors and publishers is that the “…transition to the online world is a huge change, and one they understandably view with some trepidation.” In the 2006 letter [17], the founders state that “much of the highest quality information in the world may be found in tens of millions of books tucked away in libraries and on publishers’ shelves. These books can be tremendous assets – but only if people know that they exist.” From a librarian’s perspective there is an underlying tone of disregard for the purpose of libraries prior to Internet search engines and the work in bibliographic control that is currently practiced, hence the resentment and concerns over competition between libraries and Google. In actuality Google [18] states, “We consider our primary competitors to be Microsoft and Yahoo.”

If libraries were to suddenly develop a business plan for making money then we might one day make it to the competitors list. Until then, it would be beneficial to continue analysis of Google’s information tools and provide the type of guidance to information users that has been the mainstay of our field.
Resources, continued


We are seeing considerable advances in the processes of communicating information in all its forms. As a consequence, these historical and technological transformations reflect, in a specific way, in the field of informational activities. They carry with them threats and doubts and the need to redesign characteristics of our undertakings in many different areas of human knowledge and human activities.

Information science has been characterized by the diversity of its approaches, seeking to define its essence through different studies and theories. In these studies different authors have presented numerous concepts based on their particular visions. Each of them adopts a distinct way of observing and explaining reality. Some emphasize the communicational aspects of information science, while others highlight its social function, and still others point out its strong connection to information technologies with their focus on innovation and continuous change. This close intimacy with technology has been subverting fundamental concepts of the area, while, at the same time, amplifying the field by posing new research problems.

Today, information science maintains a balance among the communication, technological and social approaches. Advances in any area of knowledge occur on the borders of each discipline, in the confluence with other sciences. These intersections with other areas are only possible through joint research, which are already occurring in information science, since activities of organization, analysis and availability of information increasingly are considered the work of interdisciplinary teams. However, in these partnerships, the role of the information scientist is still tenuous.

Information science in Brazil has come a long way since the opening of the first M.Sc. course at IBICT – the Brazilian Institute of Information Science and Technology – in 1970. It developed in close concert with library science. Such a partnership between information science and library science still exists in Brazil, even though it is sometimes not an easy one. The decades of the 1970s and 1980s saw the founding of other Ph.D. and M.Sc programs and courses in information science, as well as the beginnings of scientific journals in the field. These activities lead to the creation in 1989 of ANCIB – Brazilian Association of Information Science. ANCIB can be considered a landmark in the scientific establishment of information science as an independent area in Brazil. But if we compare its development to that of other areas such as computer science, information science in our country still lags other fields.

Currently, information science in Brazil has strengthened its interdisciplinary relationships with areas such as communication, linguistics, psychology and information...
technology. Even if we have in our information science Ph.D. and M.Sc. programs an increasing number of professors of various areas of knowledge (library science, engineering, linguistics, history, administration and informatics, among others) multidisciplinary research is still wanting. But an interdisciplinary approach is necessary if we are to develop models, redefine concepts and create new approaches that incorporate cultural, historical and social aspects to face problems caused by the new roles of knowledge in today’s society. Only then, will information science be fulfilling its mission.

We are currently experiencing a new stage of development, where information exchange by digital networks is permitting the emergence of a new paradigm – the paradigm of cooperative sharing of knowledge. Knowledge transmission to those who need it is a social responsibility.

However, in what measure does information science participate in this exchange? How can our science, which deals with information use, transform the individual and the society?

A field can be built only if it understands the specifics of the processes that cross its path. The information society manifests itself by means of contradictions and conflicts. The continuous production of information and the accumulation of knowledge contribute to the strengthening of policies of concentration and exclusion. However, the information partnership also contributes to the appearance of intellectual technologies that transform cognitive capabilities. So, the sharing of wisdom and knowledge is a reality of the informational society. Understanding the complexity of the information society demands a plural approach through interdisciplinary studies which will only be possible in Brazil if Brazilian information scientists join researchers with other areas of knowledge.

In a world of social, economic and political differences, the simple possibility of access to information does not imply its effective use in a way that can be transformed into knowledge. More than that, it is necessary that the individual who receives it be able to process it and transform it to his benefit and to the benefit of society. One of the fundamental roles of information science is surely to contribute to information and knowledge sharing and toward the reduction of digital exclusion.

In order to intervene in social life and to encourage development, information must be accepted, assimilated and transformed. But this is not an easy task, since the reality in which we have to intervene is diversified, multifaceted and dynamic. So an efficient and specific information transfer depends on the suitability of that information to the conditions and the reality of the receiver. Considering the social disparities in the Brazilian population, one of the missions of information scientists in our country, therefore, is the effective socialization of knowledge, which can be fulfilled through actions of information literacy development.

In this sense, through the idea of socialization of knowledge, it seems to us that the mission of information science is fulfilled. By active participation in an information society, abilities are exchanged in a way that diminishes the information gap and contributes to a fairer world.
When asked whom she wanted to invite to her 100th birthday party Brooke Astor answered without hesitation, “100 librarians.” Librarians are the ultimate information highways. They gently guide us to frame our information need so that we find exactly what we want or need. Librarians don’t know where everything is but have a sense of shared meaning. Search engines know where everything is but know nothing about what it means. Given a choice, who would want to celebrate a birthday with a search engine?

Organic search engine optimization (SEO) is a process of constructing and amending sites to map to the functionality of search engine technology so that sites can appear higher in the results set for client search terms and phrases. Good SEO does not have to interfere with or dictate to IA. It is a composite of actions and best practices that can ensure maximum visibility with peaceful coexistence.

Search engines are attractive technology because they put wayfinding directly into the customer’s hands. Results are immediate, revelatory (showing multiple results at once) and efficient (for known item searches). Google made it all look so easy, so now our clients expect efficient, quick and plentiful results all of the time. Close enough is often good enough. Except that PageRank has nothing to do with the client query. It is a pre-query value calculated at indexing and recalculated periodically.

As the hardware gets cheaper (Google’s index is reputed to live on cheap PC servers running open source software… many, many cheap PCs are needed to host a 9 billion page inverted index and cache of all pages), the software becomes more sophisticated with asynchronous as well as synchronous query calculations taking place. This brute force computational power seeks to derive contextual meaning through rapid and multiple comparisons.

Search technology now uses computational methods to apply a growing ontology based on a vector space comparison. Examples are found in the Hilltop algorithm that sees the web divided into categories with designated expert or authority pages and in the similar Hypertext Induced Topic Selection (HITS) algorithm. Both of these algorithms are performed on a small subset of the corpus that best represents the nature of the whole web. Results are ranked according to the number of non-affiliated “experts” that point to a particular site through links.

The beauty of Hilltop and HITS is that unlike PageRank, they are query-specific and reinforce the relationship between the authority and the user’s query. You don’t have to be big or have a thousand links from auto parts sites to be an authority. Once designated as an authority, links from these pages are afforded more weight.

So, the good news for wayfinding is that the technology is again looking to a human-mediated classification schema for web content and search results ranking. The not-so-good news is those information architects/interaction...
designers/experience designers and others with expertise in human behavior toward information and information spaces are nowhere to be found in these developments.

On the search engine event horizon is a thesaurus-like search engine that has combined its index with human mediated categories. A query is matched to the categories. One search company wants to take the trouble out of selecting a search engine by brokering queries and directing them to the engine with the best results. Both MSN and Google are working on behavior search applications that collect information such as how the client segments information on the hard drive, what they search for, what they click on, how long they stay on a page, what they have in their bookmarks, how those bookmarks are arranged or how they iterate their searches. This information is then synthesized and used to sort web search results.

The future doesn’t look so bright for IA/ID/UX involvement unless we engage quickly and directly. We can continue to disengage entirely. We can cooperate by developing best practices based on our knowledge of user needs and our abilities to structure information spaces. We can initiate changes in the direction of search technology to facilitate understanding of spatial relationships through the modeling of information spaces. Obviously, this approach is the most complex requiring changes in thinking and practice.

Navigation blindness, navigational fatigue, Mark Hurst’s the Page Paradigm and Danielson’s transitional volatility tell us it is time to change our wayfinding strategies and tools. And, if you look at the analytics on your site, we know this also. Considering the sometimes hundreds of sites folks visit in a given session, it is understandable that they get tired of figuring out the navigation schemas for each one. It is so much easier to go to the search box or pogo back to the search results.

I believe that our path back to full participation in our customer’s information is enhanced by the following new tools:

- **Page code strategy:** ensuring that rich media applications are coded for maximum visibility to search engines – spiders that do not have thumbs to press buttons or eyes to see that beautiful Flash or AJAX.

- **Metadata strategy** that sees us using all means to describe the content to the spider with keyword rich titles and to our customers with rich description text that disambiguates the destination. Like Jessica Rabbit, keywords are not bad – they were just drawn that way for a bit. In today’s topical, categorical, ontological web, keywords may experience a comeback.

- **Content strategy** that publishes deep, rich topic focused authoritative content and hubs that provide navigation to other authority resources. The fold is dead. Customers no longer pay for web access by the hour. They will scroll, and they will print out when they do not want to read on the screen.

- **Linking strategy** to make semantic relationship between content items. The new algorithmic sophistication rewards us when we “think about the link” and so we should. A good linking strategy includes requests to similar websites for reciprocal linking to capitalize on vertical search engines, online bookmarking sites, blogs and newsgroups.

I am not suggesting that we do away with all of our traditional tools. Many of them have and will continue to
withstand the test of time. I am suggesting that we acknowledge the prominence of search tools in wayfinding. Only then we can bring the locus of attention back to the person and away from the tool, the processing speed and the complex math required to build predictive engines that will never fully understand meaning.

As I see it, we can either cooperate with the technology so that we can initiate positive directions for our customers. Or we can capitulate and let the technology develop around and eventually ahead of us. For me, the choice is obvious. We must integrate our human nature into the system design. We can work with search technologists to make our work more visible. It is not too late for us to regain our place at the information finding table.