Date : 14/05/2007

User tagging of library resources: Toward a framework for system evaluation

Jonathan Furner
University of California
Los Angeles, USA

Meeting: **157 Classification and Indexing**
Simultaneous Interpretation: no

WORLD LIBRARY AND INFORMATION CONGRESS: 73RD IFLA GENERAL CONFERENCE AND COUNCIL
19-23 August 2007, Durban, South Africa
http://www.ifla.org/iv/ifla73/index.htm

Abstract

Although user tagging of library resources shows substantial promise as a means of improving the quality of users' access to those resources, several important questions about the level and nature of the warrant for basing retrieval tools on user tagging are yet to receive full consideration by library practitioners and researchers. Among these is the simple evaluative question: What, specifically, are the factors that determine whether or not user-tagging services will be successful? If success is to be defined in terms of the effectiveness with which systems perform the particular functions expected of them (rather than simply in terms of popularity), an understanding is needed both of the multifunctional nature of tagging tools, and of the complex nature of users' mental models of that multifunctionality. In this paper, a conceptual framework is developed for the evaluation of systems that integrate user tagging with more traditional methods of library resource description.

Tagging and user tagging

Tagging is the process by which the resources in a collection are tagged—i.e., assigned tags in the form of words, phrases, codes, or other strings of characters—with the dual intention (i) that the tags individually or collectively represent features of the tagged resources (or of resource–tagger relationships), and (ii) that such representations or descriptions may be exploited by search services that enable people to discover the particular resources that are of interest to them at particular times.

User tagging is tagging done by the "users" of search services—i.e., by those whose participation in the resource discovery or information retrieval process has historically been limited to the expression of information needs and construction of search queries, stopping well short of the

determination and recording of resource metadata. User tagging—also known as collaborative tagging (Golder & Huberman, 2006; Macgregor & McCulloch, 2006), social tagging (Tennis, 2006), and social bookmarking (Hammond et al., 2005)—is a kind of tagging that has attracted considerable attention in the early 2000s, as the technologies that support web-based implementations of user tagging have become more widely understood.

The characteristics of user tagging that distinguish it from other manual methods of textual representation, description, annotation, or categorization of resources (such as conventional subject cataloging, abstracting, indexing, and bibliographic classification) include the following (Furner, Smith, & Winget, 2006):

(a) User tagging is *user-oriented*. Tags for the resources in a given collection are generated by the members of the community of people who have a demonstrated interest in searching that collection, rather than by professional catalogers or indexers who are tasked with tagging as a means to support others' resource discovery.
(b) User tagging is *empowering*. People who might in the past have been accustomed to searching databases by attempting to predict the descriptors used by "experts" are given the opportunity to record their own knowledge about resources.
(c) User tagging is *democratic*. Taggers are not selected for their expertise by collection managers, but are self-selected according to taggers' own interests and goals.
(d) User tagging is *cheap*. Taggers typically volunteer their efforts at low or no cost to collection managers.
(e) User tagging is *collaborative*—if only in the sense that any given record or description of a resource is potentially representative of the work of multiple people.
(f) User tagging is *distributed*. No single person is required to tag all of the resources in a given collection. At the same time, no single resource needs to be tagged by all of the people in a given community.
(g) User tagging is *dynamic*. The description of a given resource may change over time, as different people come to make their own judgments of its nature and importance.
(h) User tagging is *instructive*. The descriptors supplied by taggers may be analyzed with a view to learning about the kinds of aspects of resources that are interesting or significant for the members of the taggers' community.

Web 2.0, Library 2.0, and OPAC 2.0

For several years, the opportunity to tag resources has been offered to the users of many popular web-based search services. Among these are the "bookmarking" service Del.icio.us,[1] which supports user tagging of websites, and the photo sharing service Flickr[2] (both now owned by Yahoo! Inc.). More recently, cultural institutions of various kinds—such as libraries, archives, and museums—have begun to experiment with online services that similarly offer tagging opportunities to their clienteles (Trant, 2006). The most celebrated example of such a service is probably PennTags,[3] a bookmarking tool that (like Del.icio.us) allows its users at the University of Pennsylvania to assign tags to any web-addressable resource, but that (unlike Del.icio.us) is integrated with its university library's online catalog so that tags may be added by catalog users to the records representing individual resources held by the library. A prominent example of a user-tagging service offered by a public library is John Blyberg's production for the Ann Arbor

District Library,[4] which allows library users to assign tags to individual library resources, to write reviews of resources and to comment on others' reviews, to view lists of most frequently assigned and most recently assigned tags, to view lists of most recently written reviews and comments, to flag useful reviews, and to search the content of tag-sets, reviews, and comments.

Library implementations of user tagging can be viewed as instances of a particular type of library service that draws, for much of its inspiration, on the ideas and technologies of the so-called "Web 2.0."[5] Characterizations of Web 2.0 typically focus on the ways in which web-based services are increasingly being used in support of collaborative activity, social networking, resource sharing, and community building (Maness, 2006). Other types of library service that are commonly considered collectively under the "Library 2.0" heading (and exemplified by services such as Hennepin County Library's Bookspace[6]) are:

- the provision of opportunities for library users to customize and/or personalize various aspects of the content, format, and style of online interactions with library resources;
- the provision of opportunities for library users (i) to share with others their reviews, recommendations, annotations, and lists of library resources, (ii) to comment on others' reviews, etc., (ii) to search the contents of shared reviews, etc., and (iii) to navigate between records that are identified, through analysis of users' recommendations, as "related" records;
- the production and publication, by library staff, of regular blog entries that include references to library resources (e.g., new acquisitions, resources relevant to current events, staff favorites) and links, either to the resources themselves or to records further describing those resources;
- the compilation, from various sources (traditional and/or non-traditional), of catalog records that provide enhanced or enriched data about library resources (e.g., tables of contents, cover art, biographies, summaries, reviews); and
- the distribution of new records, blog entries, reviews, recommendations, annotations, lists, comments, and search results by RSS feeds.

Even more specifically, implementations of user tagging that are designed to enhance the functionality of online public-access catalogs of library resources (OPACs) are the results of developers' consideration of the potential of "OPAC 2.0"—a model for the redesign of catalogs as "social OPACs" that purposefully invite the users of catalogs to participate not only in the exploration and exploitation of catalog records, but also in their creation. One especially promising method of cheaply implementing OPAC 2.0 is to exploit the functionality of open-source blog-management software in the production of blog/OPAC "mash-ups." Casey Bisson's "WPopac"[7] for Plymouth State University's Lamson Library, for example, is an implementation of the library's Innovative Interfaces OPAC within the framework provided by WordPress software, and offers a static, permanent address for each library resource; simple commenting, trackbacking, and tagging facilities; automatic creation of lists of recent searches and comments; and automatic creation of links to "related" resources that share certain characteristics.

Toward a taxonomy of library implementations of user tagging

Library implementations of user tagging may be categorized on several separate dimensions. The type of parent institution (academic, school, public, non-profit, corporate, etc.), the type of user to whom the tagging tool is directed (domain expert, learner, or novice; scholar, professional, or layperson; etc.), the type of resource to which tags may be assigned (monographic and/or serial titles; journal articles and/or conference papers; blog entries; institutional and/or personal websites; etc.), and the type of access offered to the content of resources (full-content, part-content, or surrogate only) are four such dimensions. More interestingly, perhaps, library tagging systems may be distinguished on the basis of (a) the functionality of the tagging service itself, (b) the functionality of the associated search service, (c) the goals of the implementers of the service, and (d) the motivations of its users.

(a) The feature-sets of different tagging services vary in many respects, including the following:

- the degree of restriction (if any) placed on the number and/or combination of tags that a tagger may assign to a given resource;
- the degree of restriction (if any) placed on the tagger's choice and form of tags;
- the provision (if any) of context-sensitive suggestions for tags, or for facets that the tagger may wish to consider;
- the provision of access (if any) to structured vocabularies of tags;
- the provision of access (if any) to lists or clouds of most frequently- or recently-assigned tags; and
- the provision of online access to the full content of resources: In some cases, taggers are allowed to view the full content of a given resource, and then to assign, directly to that resource, tags that describe it; in others, taggers are allowed merely to view the record describing a given resource, and then to enhance that record by adding to it tags that further describe the resource.

(b) Similarly, the feature-sets of the search services associated with different tagging services also vary in many respects, including the following:

- the degree of restriction (if any) placed on the number and/or combination of tags that a searcher may use in a given query;
- the degree of restriction (if any) placed on the searcher's choice and form of tags;
- the provision (if any) of context-sensitive suggestions for tags, or for facets that the searcher may wish to consider;
- the provision of access (if any) to structured vocabularies of tags; and
- the provision of access (if any) to lists or clouds of most frequently- or recently-searched tags; and
- the extent to which tag search is integrated into the existing OPAC search: For instance, are tags searched when users do "keyword" searches of the OPAC? Is it possible to search or browse tags from within the OPAC?

(c) The actual goals of a given tagging service's implementers—whether explicitly expressed or left unstated—may potentially include any, or any combination of, the following (among others):

- to engender a sense of community among library users in separate and remote locations;
- to allow library users to identify other individuals with whom they share interests;
- to engender a sense of empowerment among library users who may not otherwise participate in or contribute to library activities;
- to encourage library users to engage with the resources that they tag, and thereby to allow users to come to a deeper understanding of those resources and of the contexts in which they were produced;
- to improve the effectiveness of retrieval of records and discovery of resources, (i) by enhancing the quality of existing records, and/or (ii) by allowing users to discover resources in a new way, viz. by taking account of the recommendations of other users like themselves;
- to improve the effectiveness of personal rediscovery of resources, by offering users the opportunity to bookmark resources that they like and that they predict they will wish to access again;
- to allow library users to determine which kinds of resources and/or topics are currently popular, newsworthy, or receiving attention;
- to improve the entertainment value of, and thereby the level of user satisfaction with, the search experience; and
- to reduce the costs normally incurred in manually cataloging, indexing, or classifying the resources in a collection.

(d) Similarly, the motivations of prospective taggers can be identified as potentially including any of the following (among others):

- to participate in the activities of a community of like-minded people;
- to identify other individuals with whom they share interests;
- to contribute to the activities of the library;
- to engage with the resources being tagged and/or with the records that describe them;
- to contribute to improvements in the effectiveness of other users' searches, (i) by enhancing the quality of existing records, and/or (ii) by recommending resources that are determined to be of potential interest to others;
- to bookmark resources to which repeated personal access is foreseen;
- to determine which kinds of resources and/or topics are currently receiving attention;
- to pass the time in a manner that provides entertainment;
- to share their knowledge of the content of resources with others;
- to demonstrate the extent of their knowledge of the content of resources; and
- to benefit from the receipt of any concrete incentives supplied by the implementing institution in return for tagging efforts.

The great variation in the kinds of motivation prompting libraries to implement user-tagging services, and library patrons to make use of them, is a factor that complicates efforts to evaluate and compare the performance of different services, as we shall see in the next few sections.

Evaluating system success

Although user tagging of library resources shows substantial promise as a means of improving the quality of users' access to those resources, several important questions about the level and nature of the warrant for basing retrieval tools on user tagging are yet to receive full consideration by library practitioners and researchers. Among these is the simple evaluative question: What, specifically, are the factors that determine whether or user-tagging services will be successful?

An essential preliminary to any attempt to answer this question is a definition of "success" in this context. The extent to which a given system succeeds is typically understood as equivalent to the level at which it performs its function—either the function that it was intended by its designers to perform, or the function that it has for its actual users.

It is often difficult, when attempting to identify the ultimate intentions of systems designers, to separate (i) the desire to create a "quality" product—i.e., one that allows its users to get *their* jobs done effectively, quickly, cheaply, and easily—from (ii) the wish to create a product that can be sold (literally or metaphorically) to as many users as possible. One simple way of measuring the success of a system is to consider the level of its designers' satisfaction with its sales or usage figures. Of course, it is reasonable to assume that, in practice, the "salability" of a given system at least partly depends on the extent to which it is perceived by its users (potential and actual) to work well. It is important to recognize, nevertheless, that the popularity and quality of any system are only ever contingently related. It thus remains useful, at least in some cases, to attribute to designers a genuine desire to build systems that will optimally meet the needs of their users, rather than to build systems that manage to exceed their potential users' threshold of perceived value to whatever minimal extent is necessary to attract mere usage. In these cases, the level at which a system performs the function that it has for its designers can be considered equivalent to the level at which the system performs the function that it has for its users.

Even if popularity is discounted as a measure, the measurement of system success remains problematic in at least two ways. In the first place, a system is typically used by multiple groups of people, each group characteristically engaged in different kinds of task and motivated to interact with the system in different ways, for different purposes, and for different reasons. Moreover, a system typically has multiple functions, even for a single individual interacting with the system at different times, and even for a single individual at a single time. In the second place, the quality or "goodness" of a system's performance of any given function may be measured on the basis of any (or any combination) of a number of different criteria.

Multiple user groups, multiple functions

Library implementations of user tagging are far from unique in that they are services that are used by multiple groups of people and for multiple kinds of function. The "users" of a tagging service can be subdivided into two primary subsets: the implementers of the service (i.e., the collection managers), and the "end-users" who are interested in the contents of the collections to which the service provides access. As we have seen, the motivations and goals of individual members of each of these groups can be complex, amorphous, and difficult to identify or express, even for the

members themselves. The situation is further complicated by the likelihood that different individuals have different mental models of the ways in which use of tagging services can potentially help them to pursue their goals.

Nevertheless, it seems useful to distinguish at a high level (a) among kinds of primary motivation, (b) among kinds of primary usage of the service (as conceived by users), and (c) among kinds of ultimate goal (again, as conceived by users).

(a) Kinds of primary motivation. It is possible to categorize a tagging-service user's primary motivation either as *individualist* or as *social*. An individualist motivation is one that pushes the user to focus on his/her own personal goals; a social motivation is the result of a desire to help others achieve their objectives. In reality, of course, the distinction is blurred: one's wish to help others may itself be the result of a calculation that helping others is likely to be the most effective way of improving one's own circumstances. Skeptics may wonder if social motivations can ever be pure or authentic in that respect.

(b) Kinds of primary usage. End-users of tagging services may conceive their primary usage either as tagging—the creation of descriptions of resources—or as searching—the exploitation of descriptions in order to find resources. In other words, tagging may be viewed either as an end in itself, or as a means to an end. End-users may be characterized, in this respect, either as (tagging-) *intrinsicalists* or as (tagging-) *instrumentalists*.

(c) Kinds of ultimate goal. End-users of tagging services may conceive their ultimate goal in any of a variety of ways. For some, the ultimate goal may be to engage in some deeper form of interaction with the resources in a collection, for the sake of the benefits that the very experience of interaction supplies. For others, the ultimate goal may simply be to complete a particular task, external to their usage of the tagging service itself, and maybe one that is less a matter of personal interest and more a matter of professional or scholarly duty. Again, end-users may be characterized, in this respect, either as (resource-) *intrinsicalists* or as (resource-) *instrumentalists*.

Among resource-intrinsicalists, we may also distinguish between (i) those for whom the primary benefits of interaction with resources include an improved state of *knowledge* about, or *understanding* of, those resources, the contexts in which the resources were produced, and the user's own values and attitudes towards those resources and contexts, and (ii) those for whom the primary benefits of interaction with resources include the *entertainment* value that is derived during or after that experience.

Criteria for evaluation

On what criteria should our judgment or evaluation of the quality of a tagging service's performance be based? In established frameworks for the evaluation of information retrieval (IR) systems, it is common to distinguish between criteria such as *effectiveness*, *efficiency*, *cost-effectiveness*, and *usability*; and then also to distinguish between methods of obtaining "objective" measurements on these dimensions, and methods of measuring the degree to which

users say they are satisfied with systems' performance on these dimensions (Harter & Hert, 1997).

In the context of IR evaluation, "effectiveness" is understood to refer specifically to the success with which a system is able to help a user both to identify those resources that turn out to be relevant to that user at the time of searching, and, at the same time, to avoid those resources that turn out to be irrelevant. Measures of effectiveness include recall (the proportion of relevant resources that are retrieved) and precision (the proportion of retrieved resources that are relevant). In the same context, "efficiency" is usually used to refer explicitly to the speed with which searches may be carried out. The four criteria of effectiveness, efficiency, cost-effectiveness, and usability correspond, in common parlance, to measures of how well, how fast, how cheaply, and how easily system users can get their jobs done. Of course, it is possible either to canvass users' opinions on these matters (and thus to measure user satisfaction), or to track the usage that is made of systems and to measure effectiveness, etc., directly by analyzing the observations, traces, or results of that usage.[8]

Even after almost fifty years of IR tests designed in part to clarify the nature of the relationship between (i) the quality of the results of any process of resource description (such as indexing, classification, or tagging) and (ii) the level of effectiveness of retrieval from collections of indexed resources, there are many aspects of the resource-description process that are contested matters. Should the terms used to describe the subjects or topics covered by the content of a resource be generated automatically or manually? Should those terms be derived directly from resources' contents, or assigned from other sources? How specific in meaning should each term be? How exhaustive of the topics covered in a resource should a set of assigned terms be? Should the choice and form of terms be subject to the control provided by a vocabulary of candidate terms? How useful to indexers and searchers is access to a vocabulary in which the semantic relationships among terms are represented by a faceted and hierarchical structure?

There appears to be broad consensus, nonetheless, that *indexer–searcher consistency*—the degree to which indexers and searchers agree on the subjects and concepts that given resources are considered to be "about," and on the combinations of terms that are used to express given subjects and concepts—is a fairly robust indicator of retrieval effectiveness. The assumption is that, if indexers are able successfully to predict those terms that will be used by searchers in future queries to which the resources being indexed are relevant, then levels of retrieval effectiveness will be correspondingly high. Historically, observations of the correlation between indexer–searcher consistency and retrieval effectiveness have been used as evidence in support of the provision, both for indexers and searchers, of access to structured and controlled vocabularies of various kinds.

User tagging is commonly associated, both in the popular and in the professional literature, with the emergence of *folksonomies*—vocabularies that, instead of being deliberately constructed by domain experts for the purpose of controlling indexers' choices of resource descriptors, are seen to emerge as a byproduct of large numbers of taggers' uncontrolled choices, and thus to reflect more accurately the actual consensus of opinion of end-users of tagged resources (Guy & Tonkin, 2006; Spiteri, 2006). Although a system designer's decision not to provide vocabulary control is independent of the decision to implement user tagging, most implementations of user tagging are

based on a folksonomic model. Tagger–searcher consistency in concept identification and term selection is not considered to be adversely affected by the lack of vocabulary control, since it is assumed that taggers are drawn from the same population from which searchers are drawn, and hence (so the argument goes) high levels of tagger–searcher consistency are assured.

Conclusion

Very little of the existing literature on user-tagging services has reported on empirical evaluations of the effectiveness of retrieval from collections of tagged resources, and few tests have been carried out that would allow for confirmation of the validity of the assumption that taggers tend to use the same terms to tag resources as searchers use to look for those resources. On the other hand, it remains far from clear that such a directed evaluation—focusing on retrieval effectiveness as the most important criterion against which tagging services should be judged—would be appropriate, since so much of the usage of tagging services can be construed as the result of motivations far removed from a simple desire to improve future retrieval.

It is hoped that the presentation in this paper of a conceptual framework for evaluation will encourage further discussion of this topic. We have seen how different kinds of user have different motivations for making use of tagging services and correspondingly different perceptions of the functions of those services, and how assessment of the success with which systems perform any of those functions may be based on a variety of different criteria. It appears that it will be important, in the course of future evaluations of user-tagging services in libraries, that clarification and justification be provided of research-design choices of several kinds.

Footnotes

1. See http://del.icio.us/.
2. See http://www.flickr.com/.
3. See http://tags.library.upenn.edu/.
4. See http://www.aadl.org/.
5. The phrase “Web 2.0”—itself implicitly referencing the title of the magazine *Business 2.0*, launched in 1998—seems to have been used for the first time in the title of a conference sponsored by O’Reilly Media and held in San Francisco in October 2004. “Library 2.0” was coined in late 2005 by Michael Casey in his blog LibraryCrunch: see, e.g., http://www.librarycrunch.com/2005/10/working_towards_a_definition_o.html.
6. See http://www.hclib.org/pub/bookspace/.
7. See http://www.plymouth.edu/library/opac/.
8. A fairly common finding is that observed levels of user satisfaction are independent of levels of effectiveness, and are in fact to be explained by the combined influence of many otherwise unrelated factors—among them, the entertainment value derived from passing the time by using the system.

References

Furner, Jonathan, Martha Smith, and Megan Winget. 2006. Collaborative indexing of cultural resources: Some outstanding issues. In *Digital Humanities 2006: The First ADHO International Conference* (Paris, France, July 5–9, 2006), ed. Chengan Sun, Sabrina Menasri, and Jérémy Ventura, 69–71. Paris: Université Paris–Sorbonne.

Golder, Scott A., and Bernardo A. Huberman. 2006. Usage patterns of collaborative tagging systems. *Journal of Information Science* 32, no. 2: 198–208.

Guy, Marieke, and Emma Tonkin. 2006. Folksonomies: Tidying up tags? *D-Lib Magazine* 12, no. 1. Available online at http://www.dlib.org/dlib/january06/guy/01guy.html.

Hammond, Tony, Timo Hannay, Ben Lund, and Joanna Scott. 2005. Social bookmarking tools (I): A general review. *D-Lib Magazine* 11, no. 4. Available online at http://www.dlib.org/dlib/april05/hammond/04hammond.html.

Harter, Stephen P., and Carol A. Hert. 1997. Evaluation of information retrieval systems: Approaches, issues, methods. *Annual Review of Information Science and Technology* 32: 3–94.

Macgregor, George, and Emma McCulloch. 2006. Collaborative tagging as a knowledge organisation and resource discovery tool. *Library Review* 55, no. 5: 291–300.

Maness, Jack M. 2006. Library 2.0 theory: Web 2.0 and its implications for libraries. *Webology* 3, no. 2. Available online at http://www.webology.ir/2006/v3n2/a25.html.

Spiteri, Louise F. 2006. The use of folksonomies in public library catalogues. *The Serials Librarian* 51, no. 2: 75–89.

Tennis, Joseph T. 2006. Social tagging and the next steps for indexing. In *Advances in classification research, Vol. 17: Proceedings of the 17th ASIS&T SIG/CR Classification Research Workshop* (Austin, TX, November 4, 2006), ed. Jonathan Furner and Joseph T. Tennis. Available online at http://dlist.sir.arizona.edu/1726/.

Trant, Jennifer. 2006. Exploring the potential for social tagging and folksonomy in art museums: Proof of concept. *New Review of Hypermedia and Multimedia* 12, no. 1: 83–105.