

What Do You Call It?

A Comparison of Library-Created and User-Created Tags

Catherine Hall
College of Information Science and
Technology
Drexel University
3141 Chestnut Street
Philadelphia, PA 19104
catherine.hall@ischool.drexel.edu

Michael Zarro
College of Information Science and
Technology
Drexel University
3141 Chestnut Street
Philadelphia, PA 19104
michael.a.zarro@drexel.edu

ABSTRACT

In this paper, we describe an exploratory study comparing the abstracting and indexing practices of a semi-expert LIS community (metadata creators for the digital library, ipl2) and the social tags generated by Delicious users for the same corpus of materials. We find over 88% of the resources in the ipl2 History collection were tagged at least once in Delicious. Overlap between the tags applied to ipl2 resources and indexing shows terms that the two groups are similar enough to be useful, yet dissimilar enough to provide new access points and description.

Categories and Subject Descriptors

H.3.7 [Information Storage and Retrieval]: Digital Libraries—collection, dissemination, user issues.

General Terms

Measurement, Documentation, Design, Economics, Reliability

Keywords

Digital libraries, controlled vocabularies, social tagging, folksonomies.

1. INTRODUCTION & BACKGROUND

Over several years there has been an evolution in the way folksonomies and social tagging activities are viewed by the LIS community. Much of the impassioned early rhetoric pitted social tags as a competitor to traditional abstracting and indexing services [e.g. 11], generating a situation where tagging was derided or largely ignored by the LIS community [10]. Subsequently, academic research and practitioner discussion has been dedicated to discovering if tagging projects and behaviors are reflective of a passing trend or indicative of a sea-change set to sweep through the cultural heritage sector [4]. Now that the dust is somewhat settled, the reality seems to lie somewhere between the two positions. Tagging practices are increasingly common; individuals tag bibliographic materials, photographs and film reviews, and established projects such as Flickr Commons (flickr.com/commons) and Steve (steve.museum) are examples of cultural heritage organizations eager to explore new ways of generating metadata for resource discovery. At the same time, traditional forms of knowledge organization systems (KOS) such

as thesauri, classifications and subject headings have proven value as organization and retrieval tools.

The creation and maintenance of professional metadata is time consuming and expensive. Although cataloging and indexing has traditionally been viewed as a one-time operation [4], new knowledge must be integrated in order to stay relevant and responsive to changes in domains and user expectations [13]. Additionally, the diversity of content in large collections makes it difficult for most institutions to ensure sufficient in-house knowledge for their description [4]. Additionally, web resources, such as those found in the ipl2 collection, are not static, and might change over time.

The aims of professional abstracting and indexing and social tagging are remarkably similar: both attempt to describe, locate and facilitate information retrieval. Unlike traditional KOS which aim to be objective, social tags represent the conceptual understanding or categorization of a resource from a very personal point of view [4]. But with an expanding pool of human annotators acting to fulfill widely varying purposes and in possession of a broad range of expertise [2], at a critical mass the tags should prove useful to a wider community.

The hypothesis that access to large sets of social tags may provide better understanding of user needs and generate new or improve existing metadata has driven a number of research studies. [13] concluded that Flickr sets could be an important source of noun terms and named entity information for both TGM and LCSH. [8] found little overlap between tags from CiteULike and Medical Subject Headings (MeSH) in MEDLINE, concluding that the two types of vocabularies embody largely heterogeneous understanding of items. Previous studies [2, 8] have found that the volume of metadata created by academic social tagging tools remains comparatively low.

2. ipl2 AND DELICIOUS

In order to study this phenomenon further, we looked at the metadata created by two different communities, the ipl2 digital library, and the social tagging system, Delicious.

2.1 ipl2

ipl2 was born as the Internet Public Library in 1995 in a library and information science class taught by Joe Janes at the University of Michigan, with the central motivating question of “what does librarianship have to say to the networked environment and vice-versa?” [5]. After moving to Drexel in 2008, and merging with the Librarian’s Internet Index (LII) in January 2010, ipl2 currently provides subject-categorized collections of more than 40,000 online resources, generated by students, volunteers, and staff members [6].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

JCDL '11, June 13–17, 2011, Ottawa, Canada.

Copyright 2010 ACM 978-1-4503-0744-4/11/06...\$10.00.

© ACM, 2011. This is the author's version of the work. It is posted here by permission of ACM for your personal use.

Not for redistribution. The definitive version was published in Proceedings of the 11th Annual ACM/IEEE Joint

Conference on Digital Libraries, JCDL 2011

A collection development manual provides guidance on the creation of metadata: *"In addition to the words and phrases found in the "Main title", "Abstract", and subject heading fields, Hypatia's search function, which also powers the ipl2 website's search, uses words and phrases found in the keyword field to help identify resources. Add at least 2 keywords or phrases not found in the "Main title", "Abstract", and subject heading fields. Include both singular and plural versions of keywords. Separate each keyword or phrase with a comma."*

2.2 Delicious

Delicious assists users with organizing and retrieving information found across the internet. It allows for the creation and application of metadata via single word 'tags'. Users free tag URLs with any words they choose, although Delicious does not recommend tags previously used for that resource. Delicious is a "social tagging" system, and allows users to share tags and activity with the wider community [9]. This tagging system was chosen because ipl2 resources are URLs found on the publicly accessible Web. Other systems like Connotea and CiteULike contain more academic resources that are not generally found in the ipl2 collection.

2.3 Tags and Tagging Motivations

Users tag for personal and social reasons, and to organize their resources or communicate information to others [1]. Users of tagging systems have been described as "categorizers" and "describers." Categorizers attempt to form a hierarchical organization scheme using a small set of tags, while describers use a large set of tags describing their resources [7]. Motivations are not mutually exclusive; tags for personal use can be useful for social purposes and vice-versa. They are freely applied keywords with minimal vocabulary control imposed by the system and can be ambiguous, misspelled, contain symbols or consist of compound words [3]. The uncontrolled nature of tags will affect their usefulness to librarians, but the wide variety of text processing tools can help mitigate these issues.

Despite the issues presented by the nature of tags and tagging systems, they offer a window into patron's conceptualization and labeling of ipl2 resources. The labels created by Delicious users have the potential to enhance ipl2 resource discovery and organization. To investigate this, we look for overlap between existing cataloging and Delicious tags, and investigate the nature of the Delicious tags themselves.

3. RESEARCH QUESTIONS

This study sought to examine how social tags might be used in conjunction with traditional LIS cataloging to improve the discovery and retrieval of digital library resources.

Using ipl2 and Delicious data, this study aimed to explore the following particular questions:

- How different from, or similar to, the controlled vocabularies and terms generated by ipl2 staff are the tags in Delicious?
- What new access points for ipl2 can be provided by social tags?
- In what ways can tags be used to supplement traditional abstracting and indexing activity?

4. METHODS

Records were extracted from the ipl2 Hypatia database by ipl2 staff. 732 records, representing the total records indexed as

"history," were extracted in the winter of 2011. Comparing URLs for unique values resulted in a total of 720 records. Duplicated records were combined into one record for that URL. Metadata for the records included the unique identifier, URL, page title, abstract, subject headings and free-text keywords. Although we extracted records from the History subject collection, examination of the records revealed some resources span multiple subjects. For example, the website "Wolfram Alpha" appears in many different subject collections. There were 6449 total keywords applied to the ipl2 resources during indexing, for an average of 9.6 terms per record.

Each of the 720 unique URLs was used to find Delicious tags on the webpage <http://www.delicious.com/url/>. Using a PHP script written by one of the authors, up to the top 30 tags as measured by frequency (the maximum number exposed by Delicious) were "screen scraped" from the resulting webpage. Screen scraping uses a computer script to mimic a browser loading the webpage, which allows us to collect and process the structured text. We inserted the results of this process into a MySQL database, allowing use to generate reports by running SQL queries on the data.

Collected with the tag text was the tag frequency, the total number of times Delicious members tagged that URL with that term. An example of this form is "government (1669)" for the URL "<http://www.whitehouse.gov/>" indicating the term government was applied to the White House website by 1669 Delicious users. 89% of the 720 URLs received at least one Delicious tag (see table 1). We do not use this frequency reported by Delicious in this study as even one application of the tag could provide a valuable access point, although this could be an important signal of tag usefulness for future work.

4.1 Limitations

We compared the text of the tag to the title, abstract, and keywords stored in the record for the ipl2 resource. The only preprocessing performed was to lowercase all text used in the analysis in both the tags and ipl2 records. This means that misspellings and other errors have not been corrected, and plural forms of words will not match singular. Because Delicious tags are one word, users may resort to techniques which give them the appearance of two or more words in their tags. This can include using the underscore or hyphen to join words, or the use of so-called "camelCase" where capitalization is used to visually define words. An example of this could be "historyResources", "history-resources" and "history_resources" used as a tag with the same two words joined in three separate ways. Although it is likely the user intends these all to mean "history resources" we do not make that assumption in this paper. Techniques have been proposed to process these sorts of tags that include using the Google Spell Checker [8].

5. FINDINGS

Previous studies have suffered from a lack of social tags. For example, [8] found that only 3.8% of articles had 20 tags or more. In comparison, we found that 42% of our URLs had at least 30 Delicious tags (Delicious only exposes up to the top 30 tags for a given resource). 12,770 total tags were collected, of which 4600 tags were unique.

Tags that appear to apply to a wide variety of resources not surprisingly appear more frequently in Delicious. The most popular tag, "history" was applied to 452 resources, followed by "reference" (262), "research" (212), "education" (144), and "culture" (143). Tags we collected also included the infamous

“toread” tag, and other personal tags. While perhaps not useful as indexing terms, these could be used to indicate the popularity or “attractiveness” of an ipl2 resource.

Table 1: Delicious tag count

Tag Count	# of URLs	% of URLs
At least 1	637	88.5
At least 2	609	84.6
At least 3	588	81.7
At least 4	564	78.3
At least 5	542	75.3
At least 10	456	63.3
At least 30	305	42.3

Table 2. Associated tags for ipl2 resource. Note, misspellings and variant forms as found in original tags

Title	African-American Mosaic
Ipl2 Abstract	Resource guide to the Library of Congress's African-American Collection. "Covering the nearly 500 years of the black experience in the Western hemisphere, the Mosaic surveys the full range size, and variety of the Library's collections, including books, periodicals, prints, photographs, music, film, and recorded sound."
Ipl2 Subject Headings	Arts & Humanities--History--African History; Arts & Humanities--History--History--by Region--North American History--United States History--African-American History; Social Sciences--Ethnicity, Culture, and Race--African/African-American
Ipl2 Keywords	resource guide
Delicious Tags	history, slavery, african-american, black_history, africanamerican, culture, reference, blackhistory, research, africanamericanhistory, education, black, resources, libraryofcongress, loc, libraries, race, african.american, american, african, african_american, black%2Bhistory, primary_sources, african_american_history, multicultural, us_history, civilrights, civil_rights, african-americans, africanamericans,

5.1. Relational quality of ipl2 terms and Delicious tags

For all but 143 records, there was at least one match between a Delicious tag and the ipl2 page title, abstract, or keywords. Eighty-three of these 143 records had no tags assigned. The remaining records have at least one tag that could potentially be used as an access point or description of the resource.

Keywords stored in the ipl2 database can be in the form of a phrase. As discussed above, Delicious has limitations that prevent users from storing phrases, thus a one to one comparison may be difficult. Even with the limitation, 204 records (33%) showed a match between at least one tag and one keyword. More common were matches between title and tag (71%), and abstract and tag (76%). These results suggest there is a match between ipl2 indexer and Delicious user. This is important because too little overlap might suggest that our communities had a too heterogeneous view of the items for Delicious tags to be useful.

5.2. New access points for ipl2 resources

Having established that the ipl2 community is similar enough to the Delicious community to potentially benefit from user-generated metadata, at this point we provide an example ipl2 cataloging record to demonstrate some common observations.

Looking at this record we can see that Delicious tags could provide ipl2 with a number of additional metadata and access points for search and browsing.

- *new subject concepts:* **slavery** (tag: slavery), **civil rights** (tags: civilrights/civil_rights), **multiculturalism** (tag: multicultural)
- *new synonyms:* **Black History** (tag: blackhistory) - synonym for African-American History, **US History** (tag :us_history) - synonym for United States History, **LOC** (tag: loc) synonym for Library of Congress
- *new types and purposes for resources:* **reference** (tag: reference), **research** (tag: research), **education** (tag: education), **resources** (tag: resources), **primary sources** (tag: primary_sources) new authorship information: **library** (tag: libraries)

5.3. How can tags supplement traditional A&I activity

We believe that the diversity of user tags is of direct benefit to access and retrieval. Many user tags represent concepts and terms that the LIS expert have not included because they feel it does not adequately explain the resource (e.g. the potentially ambiguous tags of *research* and *reference*). Like [12] we see that the ambiguity and tension exhibited in user tags represent the reality of diverse perspectives towards information objectives and the variety of information activities and needs. In addition to new concept and subject terms, we have found that Delicious tags can also be a rich source of proper nouns and synonym terms.

Social tags also highlight and respond to societal trends and changes over time. For example, *Obama* was the third most frequent Delicious tag for the whitehouse.gov website (ahead of *USA* and *President*) but does not appear in any ipl2 metadata for the same record.

6. CONCLUSIONS AND FUTURE WORK

This work shows that user-contributed tags from Delicious have the potential to be used as additional access points for ipl2 digital

library resources. ipl2 resources received on average 17.7 Delicious tags per URL, suggesting a high level of interest among users. 305 resources received over 30 Delicious tags, accounting for 42.5% of the History collection. In many cases the tag used does not match any existing term for the record, giving us potentially new metadata. The methods in this paper use open source technologies that are freely accessible. The scripting and SQL queries used do not require a high level of programming aptitude. Many libraries likely have IT staff possessing the skills needed to perform similar data collection and analysis. Future analysis will include processing of the tag text using additional open source applications, allowing us to characterize the type of tag submitted and perform more sophisticated comparison with ipl2 records.

Other types of user-contributed metadata might also be useful for further examination and term extraction. User-contributed notes and comments in the Flickr Commons were found to offer new and potentially valuable metadata for Library of Congress photo resources [14]. Techniques developed for processing tags might also be used to extract terms and descriptions from titles, comments, and other user submissions.

A key difference between our work and previous efforts at collecting user tags for library services is that we are utilizing tags created by users on a site, Delicious.com, that is completely independent from the ipl2. The tagger may have never visited the ipl2 website, yet we are still able to benefit from his or her “wisdom.” Libraries and cultural institutions are currently piloting several projects to attract interest from patrons and collect tagging data. Projects like the Flickr Commons, and LibraryThing.com rely on users that visit their websites, reflecting the nature of the materials in their collections, while our collection of URLs has no such constraint.

Future work will include creating prototype interfaces enhanced with the addition of Delicious tags. Such an interface could be compared to a non-enhanced interface allowing us to measure the effect of Delicious tags on information retrieval in the ipl2. These efforts are already under development and will be reported in future publications.

In the present study, 33% of ipl2 resources had at least one tag to keyword match, 71% matched between title and tag, and 76% matched between abstract and tag. Clearly there is some overlap between taggers and the semi-expert indexer, but there is also a wide range of new tags which can be used for enhanced resource description and discovery. A primary challenge is to discover the best way of combining the metadata created by ipl2 staff with that generated by Delicious taggers (and potentially other tagging systems), and how we should present it to our users in order to facilitate information retrieval.

7. ACKNOWLEDGMENTS

The authors are supported by IMLS Laura Bush 21st Century Librarian Fellowships. We thank Mike Galloway of the ipl2 staff for his technical assistance.

8. REFERENCES

- [1] Ames, M., & Naaman, M. (2007). Why we tag. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '07* (p. 971). Presented at the SIGCHI conference, San Jose, California, USA. doi:10.1145/1240624.1240772
- [2] Good, B. M., Tennis, J.T. & Wilkinson, M.D. (2009). Social tagging in the life sciences: characterizing a new metadata resource for bioinformatics. *BMC Bioinformatics*, 10: 313. Retrieved from: <http://www.biomedcentral.com/1471-2105/1-471-2105/10/313>
- [3] Guy, M., & Tonkin, E. (2006). Folksonomies: Tidying up tags? *D-Lib Magazine*, 12(1). Retrieved from <http://www.dlib.org/dlib/january06/guy/01guy.html>
- [4] van Hooland, S. (2006). From spectator to annotator: possibilities offered by user-generated metadata for digital cultural collections. Presented at the CILIP Cataloguing Indexing Group Annual Conference, University of East Anglia, UK.
- [5] Janes, J. (1998). The internet public library: an intellectual history. *Library Hi Tech*, 16 (2), 55-68. doi:<http://dx.doi.org/10.1108/07378839810303983>
- [6] Khoo, M., & Hall, C. (2010). Merging metadata: a study of crosswalking and interoperability. In *Proceedings of the 10th annual joint conference on Digital libraries*. JCDL '10 (pp. 361-364). Gold Coast, QLD, Australia: ACM. doi: 10.1145/1816123.1816180
- [7] Körner, C., Benz, D., Hotho, A., Strohmaier, M., & Gerd, S. (2010). Stop thinking, start tagging: tag semantics emerge from collaborative verbosity. In *Proceedings of the 19th international conference on World wide web*, WWW '10 (pp. 521–530). New York, NY, USA: ACM. doi:<http://doi.acm.org.ezproxy2.library.drexel.edu/10.1145/1772690.1772744>
- [8] Lee, D. H., & Schleyer, T. (2010). A comparison of MeSH terms and CiteULike social tags as metadata for the same items. In *Proceedings of the ACM international conference on Health informatics - IHI '10* (p. 445). Presented at the ACM international conference, Arlington, Virginia, USA. doi:10.1145/1882992.1883060
- [9] Marlow, C., Naaman, M., Boyd, D., & Davis, M. (2006). HT06, tagging paper, taxonomy, Flickr, academic article, to read. In *Proceedings of the seventeenth conference on Hypertext and hypermedia*, HYPERTEXT '06 (pp. 31–40). New York, NY, USA: ACM. doi:<http://doi.acm.org/10.1145/1149941.1149949>
- [10] Macgregor, G, & McCulloch, E. (2006). Collaborative tagging as a knowledge organisation and resource discovery tool. *Library Review*, 55(5), 291-300.
- [11] Shirky, C. (n.d.). Ontology is overrated: Categories, links and tags. Retrieved from: http://www.shirky.com/writings/ontology_overrated.html
- [12] Srinivasan, R., Boast, R., Becvar, K.M., & Furner, J. (2009). Blobjects: Digital museum catalogs and diverse user communities. *Journal of the American Society for Information Science and Technology*, 60, 666–678. doi:<http://dx.doi.org/10.1002/asi.21027>
- [13] Stvilia, B., & Jörgensen, C. (2010). Member activities and quality of tags in a collection of historical photographs in Flickr. *Journal of the American Society for Information Science and Technology*, 61, 2477–2489. doi:<http://dx.doi.org/10.1002/asi.v61:12>
- [14] Zarro, M., & Allen, R. (2010). User-Contributed Descriptive Metadata for Libraries and Cultural Institutions. In M. Lalmas, J. Jose, A. Rauber, F. Sebastiani, & I. Frommholz (Eds.), *Research and Advanced Technology for Digital Libraries*, Lecture Notes in Computer Science (Vol. 6273, pp. 46-54). Springer Berlin / Heidelberg. Retrieved from http://dx.doi.org/10.1007/978-3-642-15464-5_7