Patrons Cataloging? The Role and Quality of Patron Tagging in Item Description

William Lund and Allyson Washburn

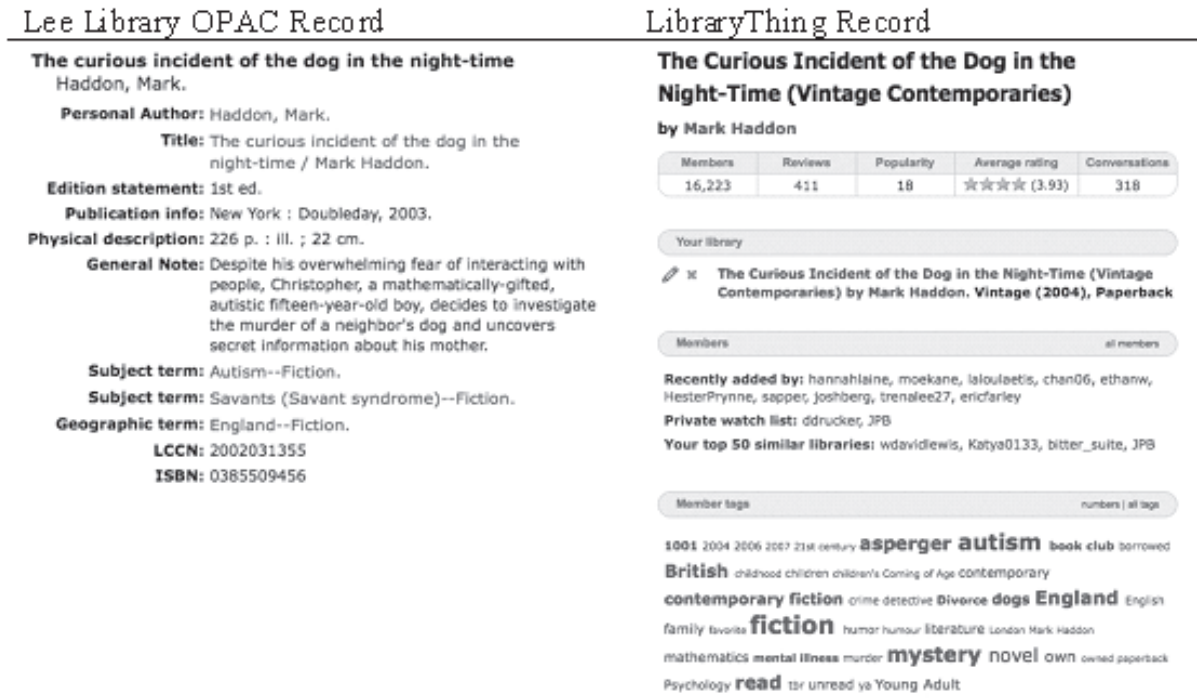
Problem

Consider the case of a searcher wanting to find a recently published mystery regarding a young man with an autism disorder. He searches on the terms "mystery" and "asperger" for Asperger Syndrome, an autism spectrum disorder. Such a book would be *The curious incident of the dog in the night*-time. The search does not retrieve the book in the traditional library online catalog because the Library of Congress Subject Heading (LCSH) is "autism". (See Figure 1.) However, when the searcher refers to the social cataloging site LibraryThing (LT), he finds the book because a contributor tagged the book with the term "asperger." This scenario is typical of many searchers who use current online public access catalogs (OPAC) of integrated library systems (ILS) that do not support patron tagging. Many searchers whether searching an online library catalog, article databases, Google or the internet, tend to use natural language keywords as opposed to Library of Congress subject headings (LCSH) or other controlled vocabularies.[1] Consequently, a large number of searches return no records or records that are not what the user wants[2-6] (Norgard et al. 1992). Additionally, users construct queries based on their expectation of the terms to be found describing the document, rather than an unfamiliar controlled vocabulary[7-10] (Yu and Young, 2004). As early as 1987, Frost[11] reported that "researchers found that, for a majority of users, the library's source of controlled subject headings remains intellectually inaccessible." From the perspective of one of the authors who provides both reference and instruction at the library, this is still largely true. The recent advent of tagging and folksonomies presents an opportunity to supplement library catalogs and improve user search results.

This paper will compare the user-created tags from the LibraryThing folksonomy with the assigned LC subject headings of the collection at the Harold B. Lee Library at Brigham Young University, a major academic research library with 3.7 million volumes. LibraryThing, found at http://www.librarything.com, is a social networking site for cataloging books and currently contains records of close to 4 million books tagged by over 574,000 contributors. The controversy surrounding the practice of community contributions to catalog records comes from the traditional view that

William Lund is Assistant University Librarian for Information Technology at Harold B. Lee Library, Brigham Young University, e-mail: bill_lund@byu.edu; Allyson Washburn is eLearning and User Assessment Librarian at Harold B. Lee Library, Brigham Young University, e-mail: allyson_washburn@byu.edu

Figure 1. Catalog Record Showing Library of Congress Subject Headings



only professionally trained catalogers are qualified to assign appropriate access points to library materials.[12-13] Proponents of tagging maintain that folksonomies are inclusive, current and facilitate discovery.[14] This study quantified the matches or lack thereof, between tags of LibraryThing and LCSH. Results showed that LibraryThing tags provide more descriptive information and more access points to library monographs than LCSH subject headings.

In the first example above, the searcher was unable to locate a known item because the record used a term that was not assigned to the LCSH headings for the item. Table 1 shows the two sets of descriptive tags for the item. They overlap, but not completely.

TABLE 1
Subject Headings as Found in the Lee Library Catalog as Compared to User Entered Tags in Librarything for "The Curious Incident of the Dog in the Night-time."

LCSH as found in the Lee Library Catalog	LibraryThing User Supplied Tags			
Autism--Fiction Savants (Savant syndrome)--Fiction England--Fiction	1001	children	family	mystery
	2004	children's	favorite	novel
	2006	Coming of Age	fiction	own
	2007	contemporary	humor	owned
	21st century	contemporary fiction	humour	paperback
	asperger	crime	literature	psychology
	autism	detective	London	read
	book club	divorce	Mark Haddon	tbr
	borrowed	dogs	mathematics	unread
	British	England	mental illness	ya
	childhood	English	murder	Young Adult

For example, both systems use the tags "autism," "fiction," and "England," however, only the library catalog record uses the term "Savants." Likewise, the LibraryThing tags include additional terms such as "divorce," "asperger," and "mathematics," not found in the Lee Library catalog entry, but which relate to the work and could be argued provide a broader description. Lastly, note that LibraryThing includes tags apparently used by the individual user to indicate information not relevant to the work itself, for instance "read" and "unread." This research does not compare non-descriptive tags such as those to LCSH.

Methodology

Comparing the descriptive metadata, such as LCSH or LibraryThing tags, between a collection using LCSH and a folksonomy in which there is no authority control required finding two collections where common materials could be compared and evaluated. The library OPAC provides a broad collection of 3.7 million volumes in which items are cataloged with LCSH assigned by professionally trained catalogers. On the folksonomy side, the developers of LibraryThing, a user-driven personal library site, provided a folksonomy of 3.9 million records. A folksonomy differs from the controlled vocabulary of the LCSH in that users provide descriptive terms based on their own understanding and vocabulary.

Linking the two collections required a common unambiguous point to match records. This turned out to be the ISBN, which is provided by both systems. Although there were over 3 million records in each collection, ultimately only about 433,000 matches were discovered based on the ISBN. The authors had hoped to be able to compare at least a million records. The smaller number of matches between the two systems was not ideal, but unavoidable. With some reflection this is not terribly surprising given the different nature of each collection. For example, the Lee Library's collection extends back over 100 years, adding approximately 50,000 new titles each year. LibraryThing, based on personal collections, is likely to favor more recent titles. Looking at the books with the most references in LibraryThing shows that they tend toward trade books, while the Lee Library collection tends toward materials appropriate to undergraduate education and graduate research.

The records from each system were linked through the ISBN and compared using the LC subject headings from the library catalog and the tags from the folksonomy. From the library catalog the comparisons used the MARC fields 650 (Topical Term), 651 (Geographic Name), and 655 (Index Term--Genre/Form). LibraryThing provided individual tags linked to the work ID. Table 2 displays the number of records, LCSH entries and LibraryThing tags used in this study. The LCSH terms were evaluated in two ways:

TABLE 2
Study Data

Records used in the study:	433,416
LCSH Entries (650:)	830,658
LCSH Geographic Entries (651):	134,571
LCSH Genre Entries (655):	75,519
Total LC Subject Headings:	1,040,748
Total LibraryThing tags:	18,783,751

as specified by the Library of Congress and split into individual keywords. For example, the LCSH term "Education--Political aspects--United States" was evaluated for matches in LibraryThing both as indicated above and as individual keywords "Education," "Political aspects," and "United States." Also, to counteract some variance in the folksonomy, all terms from both systems were converted to lower case for comparisons. One of our original thoughts was that there would be very few LibraryThing tags which matched the form of LCSH using the double dash "--". This turned out to be incorrect. We suspect there are librarians using LibraryThing for their own collections.

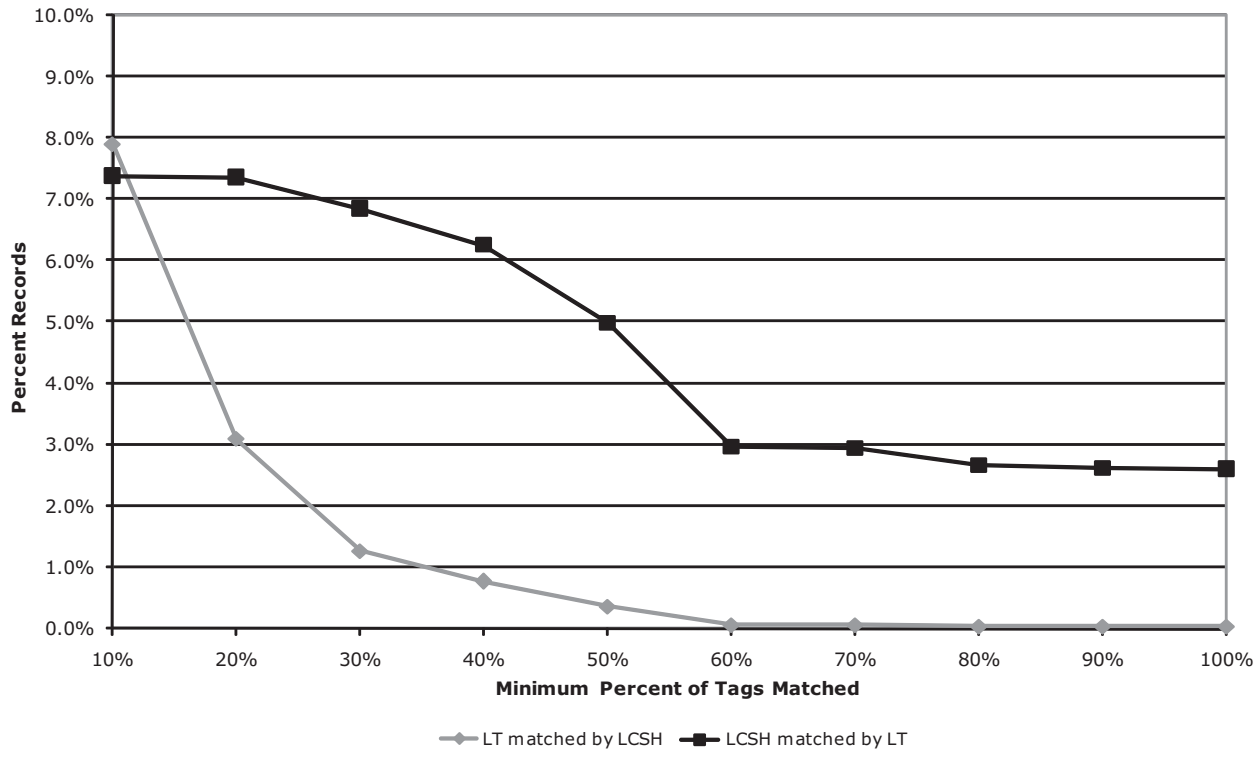
In order to facilitate query processing, the extracted records were housed in a MySQL database, from which SQL queries and Perl code could evaluate and compare the records.

Findings and Discussion

Exact Matches

The first and most obvious question concerns the matches of LCSH and the tags found in LibraryThing (LT). One would think that the number would be quite small, and that is true, but surprisingly there are 149 records, which exactly match between the Lee Library and LibraryThing. Figure 2 shows the minimum exact match of tags between the LCSH found in the MARC record and LT tags. Reading the graph, 7.9 percent of the LT records have at least 10 percent of their tags ex-

Figure 2. Cumulative Matching
Exact Match Between LCSH and LibraryThing Tags



actly matching those from the LCSH of the associated Lee Library MARC records, while 7.4 percent of the LCSH from the MARC records exactly match the tags from LT. Similarly, 3.0 percent of the LT records have at least 60% of their tags exactly matching subject headings from LCSH, while 0.1 percent of the Lee Library records have 60 percent of the subject headings exactly matched by LT tags. Finally, looking at 100 percent matching, 2.6 percent of the LCSH from the library catalog MARC records were an exact match with tags from LibraryThing. Inspecting those exact matches, it appears that they occur exclusively where the LC subject headings are simple (e.g. "acting," "child development," or "democracy") without any subheadings.

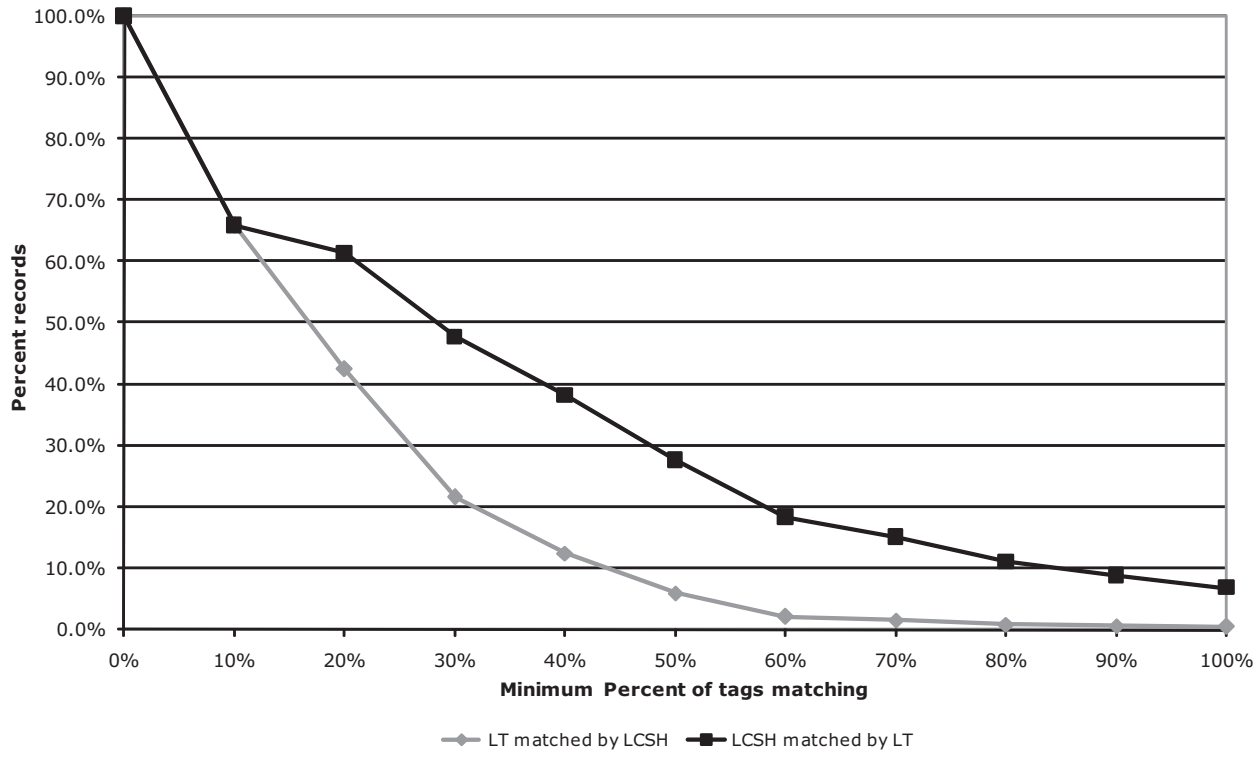
With the exception of the first data point at 10 percent, there were always more instances of LibraryThing tags matching the LCSH entries in the library catalog than the reverse. In any case, given that only exact matches between full Library of Congress subject headings and LibraryThing tags were considered, the total number of matches is still quite small. The vast majority of the records in both the library catalog and LibraryThing had no matches when considering only exact matches of the formal Library of Congress subject headings.

Keyword Matches

As indicated in Villen-Rueda[15] most library searchers do not use LCSH, but select keywords instead. Like most integrated library systems, the one at the Lee Library will in fact take a keyword search and return results based on matching a portion of a subject heading. For instance a search on "autism" will return the work *The curious incident of the dog in the night-time* in which the LCSH "Autism—Fiction" occurs. An exact match with the keyword in the search was not necessary. In LibraryThing, that same work has a number of tags, one of which is "autism."

Based on this, the next step in the study separated all subject headings into their component parts, i.e., splitting "Autism—Fiction" into two tags "autism" and "fiction" for evaluation purposes. This would appear to be closer to how the user and the system would interact. Figure 3 shows these results. Immediately, it is apparent that there is a better match between LCSH and the LT tags. For instance 65 percent of the re-

Figure 3: Acceptability of Electronic Journals by Discipline



cords of both the library catalog and LT have at least a 10 percent match and 27.7 percent of the catalog records have 50 percent match with the LibraryThing tags. Consider the following examples.

For the work *The imperfect panacea: American faith in education, 1865-1990* shown in Table 3 there are only two LC subject headings, all parts of which are matched by tags from LibraryThing. LT

TABLE 3
Comparing Lcsh and Librarything Metadata. The Imperfect Panacea : American Faith in Education, 1865-1990

Library Catalog Record		LibraryThing Record	
Library of Congress Subject Headings:	Evaluated as	LibraryThing User Supplied Tags	Evaluated as
Education--United States--History. Education--Philosophy.	philosophy history united states education	#edu 370.97320 american education American History education education in america education in the U.S. education in the united states non-fiction History history of education philosophy of education united states education	education nonfiction fiction in 370.97320 history america #edu states of american united the philosophy us

TABLE 4
Comparing LCSH and LibraryThing Metadata.
Integrative Health Promotion : Conceptual Bases for Nursing Practice

Library Catalog Record		LibraryThing Record	
Library of Congress Subject Headings	Evaluated as	LibraryThing User Supplied Tags	Evaluated as
Holistic nursing. Health promotion. Alternative medicine. Health Promotion--methods. Health Behavior. Holistic Health. Nursing Theory.	holistic methods theory medicine nursing alternative behavior health promotion	health promotion holistic	holistic promotion health

users have created tags which fully matched all of the individual components of the subject headings. Full matches occur in only 6.8 percent of the library catalog records.

Figure 4. LCSH Subject Headings Per Work

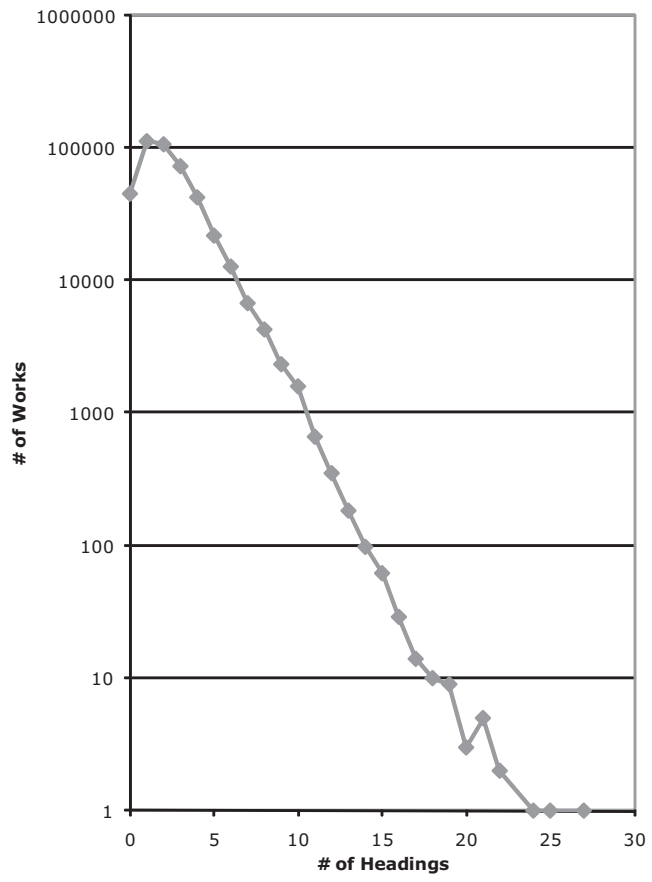
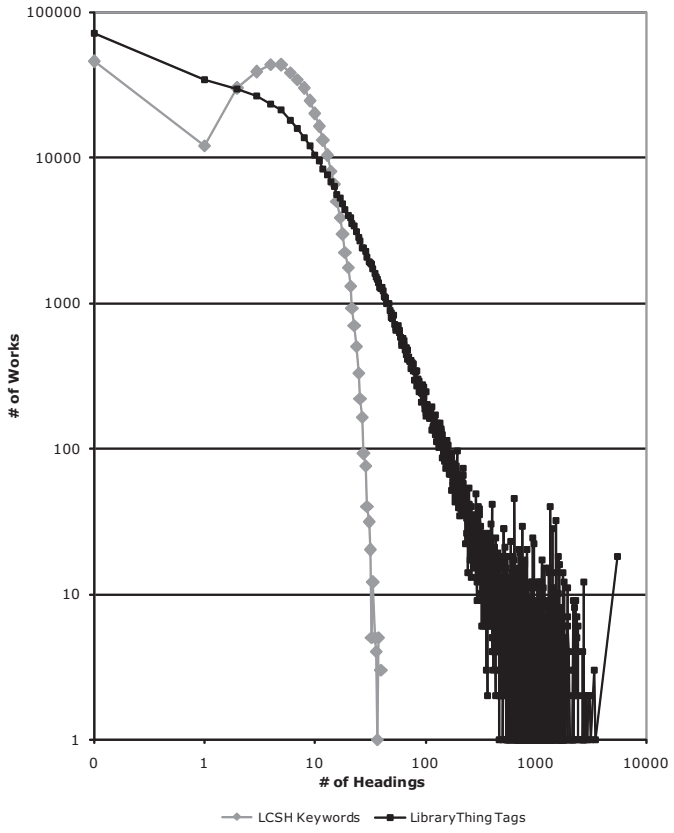


Table 4 illustrates the converse, for the work *Integrative health promotion: Conceptual bases for nursing practice* where the LCSH completely match all of the LT tags. However, referring back to Figure 3 it can be seen that the subject headings from the catalog match the LT tags much less frequently, in only 0.6 percent of the records. One possible explanation for this in the case of the work in Table 4 is that there are only four LT users who have included this work in their collection.

This is the source of much of the observed difference between the library catalog and the LT records. Whereas the library catalog records, are for the most part static once they have been created, the LT records continue to grow as users add tags, which for them are meaningful descriptions of a work. The cataloging standard for creating a subject heading is that when a subtopic comprises 20 percent of a work[16] it should be added as a heading. However, the LT tags are created based on user utility rather than a specific degree of topic coverage. An extreme example of this in LibraryThing is the work *Narrative of the life of Frederick Douglass* for which there are 504 distinct tags in LT and only two subject headings in the library catalog. This work is held by 1,873 users of LT, which is one explanation of the proliferation of tags.

In general, there are far more metadata entries and keywords in LibraryThing than LCSH in the library records. Figure 4 shows the number of works that have the given number of LCSH tags. For example, 21,932 works have five LC

Figure 5: LCSH Keywords (Split Subject Headings) and LibraryThing Tags per Work



subject headings in the OPAC. In fact 93% of the works in this study had five or fewer LC subject headings.

Dividing the LCSH into keywords and comparing this to the number of tags found in LT, we see in Figure 5 that LT tends to have more tags than the same work has keywords derived from splitting apart the subject headings. Specifically, 10 percent of the works had more than 12 keywords (as derived from the LCSH of the MARC record) while 32 percent of the LibraryThing records had more than 12 keywords (or tags). As can be seen from the graph, the maximum number of LCSH derived keywords is 40 and the maximum number of LibraryThing keywords is 5,561.

Quality of Patron Contributions

The findings of this study with regard to the number of tags contributed by users were similar to those reported by Trant[17] in a "proof of concept" study for the Metropolitan Museum of Art in 2005. Thirty-nine non-professional participants, i.e. museum administrative staff and volunteers, viewed thirty works and assigned terms to the works. "88 percent of the terms supplied by participants . . .were not found in the basic museum descriptions" (p.97). Further, the Museum Subject Cataloging committee reviewed these terms and judged more than three-fourths of the terms as valid. Combined with the findings of this study, it appears that user supplied terms for the identical work, match or provide more descriptive terms, thus increasing successful searching and access to collections.

A recent masters thesis in Computer Science at Brigham Young University titled *Improving Library Searches Using Word-Correlation Factors and Folksonomies*[18] further explores this concept by creating an enhanced library catalog using both Library of Congress Subject Headings as well as tags to create a relevancy rated retrieval of results. The author states that "experimental results show that [the system] (i) significantly reduces the amount of queries that retrieve no results, (ii) obtains high precision in retrieving and accuracy in ranking relevant results, and (iii) achieves a processing time comparable to existing library catalog search engines."[19] Experimentally, the author of this thesis found that 16.2 percent of the queries to the OPAC returned no results compared to 1.0 percent of the queries using the LibraryThing folksonomy. Further, 61 percent of the time the first returned result from the OPAC was deemed relevant by human reviewers compared to 84 percent of the time for the results using the enhanced library such with the LibraryThing folksonomy. Clearly these results are an improvement over the OPAC alone.

Conclusions

Graham[20] states "When we suspect that a significant proportion of users' needs may not be adequately met, it is legitimate to consider alternatives to current cataloging practices and policies in order to serve catalog users better" Just as Graham had some success by adding user search terms as cross-references to catalog records, tagging can serve the same function of reducing the number of no-hits searches in OPACs in general. The results of these studies indicate that in general the LibraryThing folksonomy does a better

job of representing what is in the MARC record than the converse. There is value in considering the contributions of informal folksonomies to describe library materials, providing additional access points beyond the formal LC subject headings.

Further Research

Areas of interest for future research may include stemming the LT folksonomy to include variations of a term (e.g. plurals, alternate spellings, etc.) and the exclusion of idosyncratic tags. To illustrate, many tags in LT are meaningless outside of the individual user's context. The tag "Box 1" obviously refers to the location of a work in a user's collection, but does nothing to describe a work. Likewise "read," "unread," provide only personal user information. Future research could explore a way to exclude tags, that are not descriptive in nature, perhaps based on frequency across works to indicate that they are meaningless as descriptive information.

Acknowledgments

The authors are grateful to the catalogers of the Lee Library and contributing libraries for the LC subject headings found in the MARC records, which were extracted for this work. We are also very grateful to Tim Spalding and Abby Blachy of LibraryThing and to the contributors to LibraryThing for the folksonomy provided for use in this study. Lastly, we acknowledge the contributions of the Library Information Technology division of the Lee Library for their assistance.

References

1. Patricia M. Wallace, "How Do Patrons Search the Online Catalog When No One's Looking? Transaction Log Analysis and Implications for Bibliographic Instruction and System Design". *RQ* 33, no.2 (1993): 239(14 p).

2. Rhonda N. Hunter, "Successes and Failures of Patrons Searching the Online Catalog at a Large Academic Library". *Reference and User Services Quarterly* 30 (1991): 395-402.

3. Ray R. Larson, "Classification, Clustering, Probabilistic Information Retrieval and the Online Catalog". *The Library Quarterly* 61, no.2 (1991): 133-173.

4. Barbara Norgard, Michael G. Berger, Michael K. Buckland and Christian Plaunt. 1993. The Online Catalog: From Technical Services to Access Service. In *Advances in Librarianship*, 111-148. New York: Academic Press.

5. Deborah D. Blecic, Nirmala S. Bangalore, Josephine L. Dorsch, Cynthia L. Henderson, Melissa H. Keonig, and Ann C. Weller, "Using Transaction Log Analysis to Improve OPAC Retrieval Results". *College & Research Libraries* 59 (January 1998): 39-50.

6. William Lund, "Unicorn Search Results from August 2006 to December 2006". Unpublished internal report of Harold B. Lee Library, Brigham Young University, Provo, UT.

7. Allyson Carlyle, "Matching LCSH and User Vocabulary in the Library Catalog". *Cataloging and Classification Quarterly* 10, no. 1/2 (1989): 37-63.

8. Holly Yu and Margo Young, "The Impact of Web Search Engines on Subject Searching in OPAC". *Information Technology and Libraries*, 23, no.4(2004): 168-180.

9. Karen Markey, *Subject Searching in Library Catalogs Before and After the Introduction of Online Catalogs.* Dublin, OH: OCLC Online Computer Library Center, Inc., 1984.

10. Karen M. Drabenstoff and Diane Vizine-Goetz, *Using Subject Headings for Online Retrieval: Theory, Practice, and Potential.* San Diego, CA: Academic Press, 1994.

11. Carolyn O. Frost, "Subject Searching in an Online Catalog". *Information Technology and Libraries* 6 (March 1987): 60-63.

12. Adam Mathes, "Folksonomies - Cooperative Classification and Communication Through Shared Metadata," 2004. 1-13. Retrieved from: http://adammathes.com/academic/computer-mediated-communication/folksonomies.html

13. Elaine Peterson, "Parallel Systems: The Co-Existence of Subject Cataloging and Folksonomy". *Library Philosophy and Practice,* April 2008: 5 pages.

14. Ellen Kroski, posting to Infotangle blog, December 11, 2005, http://infotangle.blogsome.com/2005/12/07/the-hive-mind-folksonomies-and-user-based-tagging/.

15. Luis Villen-Rueda, Jose A. Senso, and Felix de Moya-Anegon, "The Use of OPAC in Large Academic Library: A Transactional Log Analysis Study of Subject Searching". The Journal of Academic Librarianship 33 no. 3(2007): 327-337.

16. This standard is explained by the following reference: *General topic and subtopic; principle vs. specific case.* If a work discusses a general topic with emphasis on a particular subtopic, or presents a principle and illustrates the principle with a specific case or example, assign headings for both the general topic or principle and for the subtopic or specific case or example, provided that the treatment of the latter forms at least 20% of the work. *Example: Title: Revolutions yesterday and today.* [A survey of revo-

lutions with emphasis on the Cuban Revolution of 1959] 650 #0 $a Revolutions $x History.#0 $a Cuba $x History $y Revolution, 1959. Retrieved from: http://desktop.loc.gov/nxt/gateway.dll/Subject%20Cataloging%20Manuals/scmshm/3/10?f=templates&fn=document-frame-htm.htm&an=&3.0&q=

17. J. Trant, "Exploring the Potential for Social Tagging and Folksonomy in Art Museums: Proof of Concept". *New Review of Hypermedia and Multimedia*12, no. 1 (2006): 83-105.

18. Sole Pera, "*Improving Library Searches Using Word-Correlation Factors and Folksonomies."* Master's thesis, Brigham Young University, 2008.

19. Ibid, 5

20. Rumi Y. Graham, "Subject No-Hits Searches in an Academic Library Online Catalog: An Exploration of Two Potential Ameliorations." *College & Research Libraries,* 65, no.1(2004): 36-54.