Exploring User-Contributed Metadata's Potential to Enhance Access to Literary Works

Social Tagging in Academic Library Catalogs

Christine DeZelar-Tiedman

Academic libraries have moved toward providing social networking features, such as tagging, in their library catalogs. To explore whether user tags can enhance access to individual literary works, the author obtained a sample of individual works of English and American literature from the twentieth and twenty-first centuries from a large academic library catalog and searched them in LibraryThing. The author compared match rates, the availability of subject headings and tags across various literary forms, and the terminology used in tags versus controlled-vocabulary headings on a subset of records. In addition, she evaluated the usefulness of available LibraryThing tags for the library catalog records that lacked subject headings. Options for utilizing the subject terms available in sources outside the local catalog also are discussed.

I n recent years, many academic libraries have implemented Web 2.0 or next generation–style catalogs, often characterized by a streamlined search interface, relevancy-ranked search results, faceted browsing displays, and opportunities for more user interaction via tagging, ratings, reviews, and so on. Although interactive features are popular in large commercial sites such as Amazon and iTunes, users seem less interested in using social features in site-specific library catalogs. In an environmental scan of library, archive, and museum websites performed by a task force coordinated by the OCLC Research Libraries Group (RLG) Partnership, social media features seemed more likely to be used if the site served a niche audience or was national or multi-institutional, providing a sense of community or critical mass.[1] Michalko suggested that users do not view a library catalog as a social networking site where many people gather to share their interests and expertise.[2] Instead, they view it as a tool to help them find useful resources for their information needs. Users are more likely to want to interact with like-minded individuals on heavily aggregated sites, such as Amazon (where book reviews are heavily used and read) or Flickr (where thousands of users share

Christine DeZelar-Tiedman (dezel002@ umn.edu) is Archives and Special Collections Catalog Librarian, University of Minnesota Libraries, Minneapolis, Minnesota.

and tag photographs), or on narrowly-focused, discipline-specific websites. The research project reported here sought to compare similarities and identify differences in user tags assigned in a social networking site and Library of Congress Subject Headings (LCSH) assigned in a library catalog.

Background

A number of studies have been published investigating end user tagging of information resources. Folksonomies are the vocabularies that result from the act of users' application of subject terms, or tags, to particular items using their own vocabulary and understanding of, or relationship to, the object.[3] In contrast, LCSH and other controlled vocabularies that traditionally appear in library catalogs are based on carefully developed principles of thesaurus construction and are applied according to established standards and rules so that, ideally, terms are assigned consistently and accurately to aid the search process. Studies of inter-indexer consistency, however, have demonstrated that this ideal is rarely met.[4] Another perceived benefit of controlled vocabularies is the disambiguation of terms, so that words with multiple meanings are understood to refer to a single definition. Unlike tagging, cataloging rules and local practices tend to limit the number of subject headings to be applied to a given item. Broader or narrower headings are applied according to the number of topics covered in a work, and subject headings are typically not assigned for topics comprising less than 20 percent of the text.

Conversely, free-text keywords in the form of tags are based on whatever, and how many, terms an end user feels are appropriate and meaningful for his or her personal use. Terminology in folksonomies may vary widely based on users' personal vocabularies, cultural or geographic backgrounds, levels of expertise, or particular interest in the item. Terms may be applied inconsistently by different users because of misspellings, use of plural or singular forms, or capitalization. Tag terms could be much broader or narrower than the controlled vocabulary terms assigned.[5] Many of these disadvantages could have a positive side for information retrieval purposes in some circumstances. End users may be using language that is more current than the controlled vocabulary, more specialized, or more targeted to the layperson. By providing terminology from many levels of specificity, different users could be helped by different terms.[6] A clear disadvantage of tagging, at least in terms of search enhancement, is that since taggers primarily tag for their personal use, they might use tags that are meaningless to others.[7] These types of tags are often referred to as personal tags; examples include nonstandard abbreviations or codes, information about the location or particular details of the user's own copy, or opinion-based terms, such as "favorite" or "boring."

Even if some users are uninterested in providing their own tags to library catalog records, they might be helped by the presence of tags assigned by other users. If tags are indexed for searching, they can provide additional keyword access to items, using terms that may not appear in a record's description or controlled subject headings. Once a user retrieves a record, viewing a tag cloud (which presents an array of terms associated with an item) might help the user determine whether a resource meets his or her interests or needs. In addition, the presence of tags in a record might influence other users to begin tagging.[8] To this end, libraries may choose to import user tags into their local catalogs from larger data aggregations. Rather than relying on their own users to tag items, a ready-made folksonomy can be supplied to supplement the data in the catalog record. One such source of user tags is LibraryThing for Libraries (LTFL) (www.librarything.com/forlibraries), a commercial product (now owned by Bowker) that was developed by LibraryThing for libraries to use with their existing library systems.

LibraryThing (www.librarything.com) is a social networking site that allows individual users to catalog their own book collections. Members can add tags and reviews to records for books, as well as engage in online discussions and other interactive activities. As of May 2011, 1,343,647 LibraryThing members had cataloged more than 62 million items. The personal copies cataloged by users represent 6,106,556 works, to which 76,008,376 user tags have been added, or an average of 12 tags per record.[9] Some works generate hundreds of tags while others have none or only a few. LTFL is a separate, fee-based service offering some of the social features of LibraryThing to enhance local library catalogs, including tags, reviews, recommendations, and a browsable virtual bookshelf display of book cover images. To resolve the issue of personal tags, "preselected LTFL tags have been approved for usefulness and appropriateness by LibraryThing librarians. Highly personal tags (to read, gift from mom) have been excluded."[10] LTFL (citing March 2009 data) claimed an average 75 percent overlap with titles in the catalogs of its public library clients, but admitted that due to the nature of their collections, academic, foreign and specialized libraries would tend to have a lower overlap rate.[11]

In a pilot study conducted in 2008, the author of the present paper selected a random sample of 383 bibliographic records from a large academic library catalog and searched the titles in LibraryThing.[12] Because many of the works searched were specialized, scholarly, foreign, or older materials (as is typical in a research library), only 21 percent of the titles were found in LibraryThing, which favors newer, more popular works in English as well as canonical literary works. However, the hit rate within the original sample for creative works of literature was 45 percent. Match percentages were even greater when the literature sample was

limited by publication date or language. Fifty-four percent of literary works in English were found in LibraryThing, as was 61 percent of twentieth- or twenty-first-century literature in any language. For twentieth- and twenty-first-century literature in English, records were found for 68 percent of the sample. Because the number of literary works in the sample was relatively small (only 49 titles), further investigation is needed to determine whether this trend is representative of a larger population.

Traditionally, libraries have paid less attention to providing subject access to literary works than to nonfiction works. Part of this is practical because the "aboutness" of literary works is often more subjective than for nonfiction. Depending on the work, settings, historical periods, or characters, recurring fictional characters, and genres can be relatively simple to identify for works of fiction or drama, but thematic topics can be more elusive and more challenging to determine without reading the entire work.[13] Even then, themes such as alienation, redemption, or betrayal may be open to interpretation and are rarely explicitly stated.[14] By its very nature, poetry in most cases defies easy classification by subject, and poetry collections often lack an overarching theme to which subject terms can be applied. In 1990, the American Library Association (ALA) published *Guidelines on Subject Access to Individual Works of Fiction, Drama, etc.*, providing a thesaurus of genre terms for literary works.[15] In 1991, the OCLC/LC Fiction Project commenced in which the two sponsoring organizations collaborated with six academic and public libraries to add subject and genre headings to bibliographic records for fiction. The OCLC and LC support of the project ended in 1999, although some libraries have continued a policy of assigning subject headings to particular categories of fiction.[16]

ALA published a second edition of the *Guidelines* in 2000.[17] Later, the LC updated instructions in its *Subject Cataloging Manual: Subject Headings* to aid catalogers in providing increased access to individual works of fiction.[18] As of 2001, the provisions were being applied at the LC to current acquisitions for English-language novels.[19] As a result, while subject headings in catalog records for contemporary fiction (particularly children's literature and popular works that might be found in public libraries) are not uncommon, subject headings for older literary works are much less common. Because general library practice does not require subject headings for literary works for full-level cataloging, one can assume that a large proportion of literature records in academic catalogs lack them, depending on local policies. Although both editions of the *Guidelines* include genre terms for forms of literature other than fiction, such as poetry and drama, no high-profile projects have focused on providing subject access to literary forms other than fiction, so fewer records for those forms have any subject access.

While the pilot study cited above found the match rate between the academic library catalog and LibraryThing to be low, the match rate for literature was significantly higher, particularly for twentieth- and twenty-first-century literature in English. The aims of the current study were

- to verify the match rate between the library catalog and LibraryThing for twentieth- and twenty-first-century English language literature on a larger sample of records;
- to assess the accuracy and usability of user tags from LibraryThing by comparing LCSH and user tags from LibraryThing for these works; and
- to determine the frequency of user tags for these works in LibraryThing for various literary genres.

Literature Review

A number of studies have compared controlled vocabulary or professionally created metadata with end user tags, descriptions, and folksonomies. Several articles explore user tags and folksonomies for digital resources in the networked environment. Tonkin and colleagues investigated the collaborative aspects of social tagging in a study to determine whether the number and nature of tags assigned varied depending on whether the tagger was tagging only for him- or herself or with the larger user community in mind.[20] They found that tag use and tagger motivation vary depending on the internal culture of the particular tagging community. Kipp examined keywords assigned to journal articles by three distinct groups: user, author, and intermediary (e.g., librarian).[21] The results indicated that differences in keywords assigned by the three groups were influenced by the discrete contexts in which the indexers approach the material they are tagging. Spiteri looked at the linguistic structure of user-assigned tags from three online bookmarking sites: Del.icio.us (now Delicious), Furl (since defunct), and Technorati.[22] She found that the vast majority of user tags represented things, as opposed to materials, activities, events, properties, disciplines, or measures.

Hidderley and Rafferty proposed the concept of "democratic indexing."[23] In 1997, before the advent of social networking and user tags, they argued that different readers approach fiction in different ways and that each reader's interpretations and responses to fiction have validity. In addition, academic response and interpretation of fiction changes over time, and indexing practices should reflect this. In 2007, Hidderley and Rafferty applied this model to folksonomies, demonstrating that user tags for images in Flickr show an array of interpretations of images.[24] However, they acknowledged that for precision and recall purposes, folksonomies have limitations without some sort of institutional control.

In studying tagging and folksonomies, several researchers have identified patterns that are consistent over different tagging systems. Munk and Mørk proposed that these patterns follow a power law distribution.[25] The power law, also known as Zipf distributions, refers to phenomena where large occurrences are rare and small occurrences are common.[26] In the online environments they studied, Munk and Mørk demonstrated that a small number of tags are the most heavily used within a particular system, as compared to a vast number of tags that are used less frequently. As a user population grows, users imitate other users by reusing the most popular tags. They also found that

> individual websites are often described with an insufficient number of tags and with very general keywords, because amateurs do not necessarily understand or have not experienced the need for a stringent and precise hierarchization of the information. In this sense, folksonomies often do not yield better search results than using the same keywords in search machines such as Google, because the majority of the keywords used are very general.[27]

In an investigation of Del.icio.us, Golder and Huberman also found that "users have a strong bias toward using general tags."[28] Regarding the proportion of use of the most popular tags versus those that are less popular, they found that a pattern emerges after the first one hundred or so bookmarks are applied, in which a repertoire of tags becomes constant.

Cultural institutions such as libraries and museums have begun to actively solicit end user contributions to aid in the description of resources. One high-profile example is the LC Flickr pilot project. The LC posted digital images from two photographic collections to the popular photo-sharing site Flickr.[29] The project's primary objectives were to increase awareness of the LC's photographic collection, to gain an understanding of the mutual benefits of social tagging and end user input to the LC and the community, and to gain experience in engaging the emergent web community. According to the project report, the public response was "overwhelmingly positive and beneficial."[30]As of October 2008, only twenty-five instances of inappropriate (i.e., falling below an acceptable level of civil discourse) user-generated content were found out of 7,166 comments and 67,176 tags. Through the pilot, the LC was able to "collect user-centric, relevant terms that have the potential to increase retrieval of items in the Library's collection."[31] At the same time, in sampling the tags supplied for each collection, 45 percent of the tags of one collection and 23 percent of the tags in the other repeated words and phrases already present in the LC-supplied description.

Another experiment in soliciting user contributions to enhance or complement standardized institutional descriptions is Steve: The Museum Social Tagging Project (www.steve.museum).[32] Users were asked to provide descriptive tags to digital images of museum objects, to get a non-expert, noninstitutional perspective. Eighty-six percent of the tags assigned used vocabulary not found in the museum documentation, and 88 percent of these tags were assessed as useful by museum staff. Also speaking to user motivation, Matthews and colleagues conducted an experiment in which users were asked to tag resources in a database of selected web documents.[33] A survey of the participants indicated that some were more likely to be motivated to provide subject terms for general use if there was some indication that the terms would actually be used by and benefit others.

Several articles have been published comparing LCSH with user-supplied keywords or tags. Although different methods were used to determine matching between controlled and uncontrolled terms, the studies universally concluded that the two methods of providing subject access to library resources are complementary, rather than showing that one method is superior. Wetterstrom compared user tags assigned to general collection books from the National Library of New Zealand to the LCSH terms for the same books.[34] Terms were coded according to whether they were an exact match to the LCSH heading, a partial match (matching a cross reference or subdivision, or representing a spelling variation), or no match (differences in specificity, point of view, geographic differences in vocabulary, currency of term, or use of more popular language). Seventy-five percent of the tags did not match LCSH, and the highest instances of the divergence from LCSH included use of more popular language, related terms, and broader or narrower terms.

Strader examined the overlap between author-assigned keywords and LCSH for electronic theses and dissertations.[35] Matching was analyzed according to whether the term was identical to a main heading or cross reference, whether the same words appeared in a different order, matched partially, or other variations. She found that keywords and LCSH were complementary, and both the controlled and uncontrolled terms provided unique terms that were not otherwise present in the record, aiding retrieval by keyword.

Rolla compared LCSH headings to LibraryThing tags for forty-five books.[36] As other studies have indicated, Rolla found that LibraryThing taggers often used broader or narrower terms than supplied by LCSH or used different terminology to identify the same concepts. Additionally, every LibraryThing record included tags for at least one concept not brought out by the LCSH. Rolla also noted that the structure of tag clouds gives more weight to some user tags on the basis of the the aggregation of multiple users' tags. This is more useful for works with greater numbers of tags because inaccurate or misleading tags are less likely to come to the forefront than in records with few tags.

Weaver conducted an experiment to generate user tags for fiction in a more structured environment than free-form tags.[37] Public library users were asked to create tags for the novel *The Da Vinci Code* using an input form that prompted the taggers to create tags according to specific facets: character, plot, subject, setting, and genre. Although some personal tags were elicited, generally Weaver felt that the resulting folksonomy provided a richer description of the book than did the tag cloud for the same title in LibraryThing.

Research Method

Data Gathering

The author used the University of Minnesota online catalog, MNCAT, and LibraryThing as data sources. LibraryThing was selected because of its relative popularity and name recognition among book-oriented social networking sites, and because of the prominence of its tagging feature.

During June 2010, the author obtained a list of records from MNCAT by performing a call number browse search for items in the LC call number ranges PR6001–6126 and PS3500–3626, representing twentieth and twenty-first century English and American literary authors. Because the focus of the study was subject access to imaginative literary works, the author eliminated works of literary criticism found within these call number ranges. To make useful one-to-one comparisons of subject headings and tags for individual literary works, the author also excluded records for publications that collected or compiled works of literary authors that were originally published separately. Of the remaining items, the author manually searched every 125th title, generating a sample of 444 records out of approximately 55,500 titles. According to Krejcie and Morgan, a sample of 381 would be generally representative to a .05 degree of accuracy for a population size of 50,000.[38] This method was chosen for reasons of practicality because of the author's lack of back-end access to the catalog data. The author recognizes that this does not constitute a systematic random sample and therefore compromises the statistical validity of the data, but she hopes that the method employed may be replicated and that the results indicate trends that can be further explored by other researchers with larger samples and in other genre or disciplines. The author created a spreadsheet listing each title in the sample, with columns for author, title, publication date, number of subject headings, literary form, an indication whether or not a record for the work was found in LibraryThing, and the number of LibraryThing tags.

Between June and July 2010, the author viewed each record in the sample in MNCAT to identify its author and title (and uniform title, if applicable) and ensure it met the inclusion criteria as described above. The author counted the number of subject headings, if any, in the record and recorded them in the spreadsheet. For this portion of the study, the author counted each MARC 6XX tag as a single subject heading, so LCSH headings and genre terms in 655 fields, which come from a number of different thesauri, were counted. However, foreign-language subject headings and Medical Subject Headings were not counted. The author then searched LibraryThing to find a matching record for the literary work represented by the catalog record. Unlike library catalogs, LibraryThing typically contains a single record for a given work, representing all editions and translations. Therefore publication details, such as imprint and publication date, were not considered when determining a match. If a match was found, the author recorded the number of user tags, if any, in the spreadsheet. When a given work has many tags in LibraryThing, only the 30 most popular tags are displayed in the tag cloud on the initial view of the record screen. The user then has the option to click to view all the tags. For some works, the number of tags is in the hundreds or thousands. Manually counting each tag in these cases would have been impractical. Therefore, for titles with more than 30 tags, the author listed the number of tags in the results spreadsheet as "30+." All tags (up through 30) were counted and were not evaluated for accuracy or usefulness at this point in the study. The literary form (novel, poetry, drama, short story, essay, memoir, children's literature, mixed form, or undetermined) of the work also was noted in the spreadsheet to assess whether particular forms are more or less likely to have records in LibraryThing and more or less likely to have subject headings and useful user-supplied tags.

Subject Term Analysis

From the full sample of 444, 150 works had records in LibraryThing and had both library-assigned subject terms and LibraryThing tags. Of these, the author chose a subset of 50 that were representative of the full subset of 150 in publication date range and literary form to provide closer analysis of the nature and accuracy of the user tags applied. Personal tags (such as "ToRead" and "at_moms") and tags providing descriptive information (such as the author's name or publication details) were not analyzed. For LibraryThing records with more than 30 tags, only the tags appearing in the initial view of the tag cloud, which represent the 30 most popular tags, were considered. This limitation was applied for reasons of practicality because some records contained hundreds or thousands of tags. The first 30 tags appearing are those that the majority of LibraryThing users found useful and descriptive of the work under consideration.

As opposed to most user tags, LCSH are constructed by combining a main heading with one or more subheadings,

further refining a broad topic by bringing out subtopics: chronological, geographic, or form and genre aspects. This method of precoordinating topics is useful for browsing arrays of subjects in library catalogs but serves less of a purpose for keyword searching. Although in certain contexts the construction of subject heading strings provides a more nuanced method of identifying the meaning and contents of works, the LCSH headings and subheadings appearing in catalog records were considered as separate terms, or facets, for the purposes of comparing with user tags in this study.

Each LibraryThing tag matching the criteria above was compared with the subject headings in the corresponding MNCAT record and placed into one of the following categories:

- Match (M)—the tag exactly matched an LCSH heading or subheading
- Partial Match (PM)—the tag matched a word in a multiword heading or subheading, or varied slightly (e.g., spelling, singular or plural)
- No Match: Specificity (NS)—the tag was more general or more specific than the LCSH term
- No Match: Vocabulary (NV)—the tag represented the same general concept as the LCSH term, but used different vocabulary
- No Match: New (NN)—the tag identified a subject or concept not covered by any of the LCSH terms in the record

LCSH headings that were not represented by any LibraryThing tags also were counted.

In addition, the tags for the 191 records from the sample that had LibraryThing tags but no LCSH in MNCAT were examined to assess how useful they might be in providing subject keyword terms for users or to library catalogers in assigning controlled subject headings. Personal tags, tags duplicating descriptive elements, or those not among the 30 most popular for a given work were not analyzed. Each tag was placed into the following categories, all but the first corresponding to the types of terms typically assigned by catalogers to individual works of imaginative literature:

- Broad Form/Genre Term (e.g., novel, literature, classic, poems)
- Specific Form/Genre Term (e.g., thriller, mystery, romance, African American fiction)
- Geographic Term
- Topical Term
- Chronological Term
- Character Name

The frequency of the types of terms was compared in the aggregate as well as by literary form (novel, drama, poetry, short story, and other (a catch-all category for forms with one or few occurrences, which included memoirs, essays, children's literature, mixed forms, and one undetermined)).

Results

Record Matching

The match rate for records in LibraryThing and MNCAT was notably higher than in the 2008 pilot (see table 1). Of the sample of 444 MNCAT records, 367 (82.7 percent) had matching work records in LibraryThing. The sample consisted of 244 novels, 96 poetry collections or individual poems by a single author, 45 plays, 38 short story collections by a single author, and 21 works in miscellaneous forms (essay, memoir, children's books, mixed forms, or, in one case, undetermined). Fiction was better represented in LibraryThing than the other literary forms, with 89.8 of the novels and 89.5 percent of the short story collections having LibraryThing records as compared to a 68.8 percent match rate for poetry and a 68.9 match rate for drama. Eighty-one percent of the works in other forms had matches in LibraryThing.

No subject headings were present in 271 (61 percent) of the 444 library catalog records in the sample. Figure 1 graphically presents a comparison of LibraryThing's tags to LCSH assigned to the sample set of records. Of the works

Table 1. Match Rate between MNCAT and LibraryThing

Genre	Occurrences in MNCAT	% of Sample (N = 444)	Occurrences in LibraryThing	% of Sample (N = 367)	Match Rate between MNCAT and LibraryThing
Novel	244	55.0	219	59.7	89.8%
Poetry	96	21.6	66	18.0	68.8%
Drama	45	10.1	31	8.4	68.9%
Short Stories	38	8.6	34	9.3	89.5%
Other	21	4.7	17	4.6	81.0%
Total	444	100	367	100	82.7%
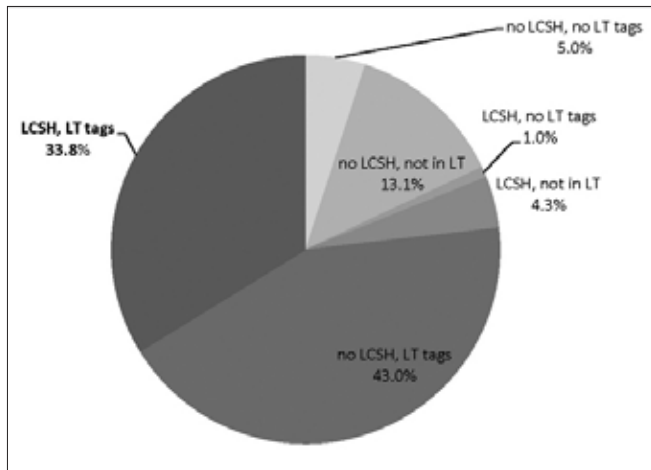
Figure 1. Comparison of LibraryThing Tags and Assigned LCSH (N = 444)

Figure 2. Percent of Records with Subject Headings or Tags by Literary Form (LCSH: N = 444; LibraryThing: N = 367)

without subject headings, 58 were not in LibraryThing, and 22 were in LibraryThing but had no user tags in the record. However, 191 of the MNCAT titles lacking subject access had tags in LibraryThing. Additionally, 150 of the MNCAT records with LCSH also had tags in LibraryThing. Assuming no one-to-one comparison between the LCSH vocabulary and the end user supplied tags, 341 (76.8 percent) of all MNCAT records in the sample show potential for being enhanced by LibraryThing tags.

When comparing whether records have subject headings by literary form, some differences between the library catalog and LibraryThing are apparent (see figure 2). In MNCAT, 45.9 percent of the 244 novels in the sample had subject headings, as did 40 percent of the 45 plays and 42.9 percent of the 21 works in the "other" category. Conversely, only 34.2 percent of the 38 short story collections and 21.9 percent of the 96 poetry collections had subject headings. In LibraryThing, fiction in either the long or short form was more likely to have tags than other literary forms, with 84.8 percent of the 219 novels and 81.6 percent of the 34 records for short stories having tags. Tags were present in records for 63.5 percent of the 66 poetry works and 62.2 percent of the 31 plays. No matter what the form, the assignment of tags in LibraryThing for literature is considerably higher than the rate of subject analysis for the corresponding works in the library catalog.

LCSH and Tag Comparison

By looking at the results described above, the potential for data from network-level social networking sites dedicated to books and reading, such as LibraryThing, to enhance subject access to twentieth- and twenty-first-century literary works in English appears to be high. But before an academic
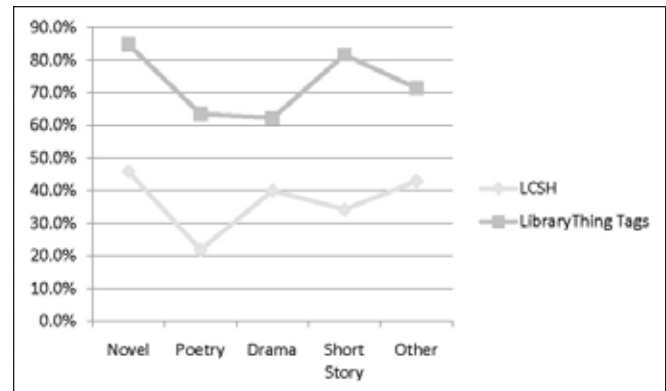
library chooses to further invest in strategies to take advantage social tagging data, a closer look should be taken at the quality and accuracy of the alternative tags available from these sources. To begin investigating this issue, 50 records that had both LCSH and LibraryThing tags were selected from the sample. This smaller sample is proportionately representative of the larger population of 150 records with both tags and subject headings as broken down by publication date and literary form.

In the selected 50-record sample, 114 library-assigned subject headings were found, an average of 2.28 per record. The same 50-item set in LibraryThing contained 684 usable LibraryThing tags, an average of 13.68 per record. These 684 tags were compared to the LCSH assigned to the same items (table 2). Sixty-one (8.9 percent) of the tags exactly matched an LC subject heading or subheading in the corresponding catalog record. Sixty-seven (9.8 percent) partially matched a heading or subheading, i.e., matched a single word in a multiword heading or subheading, or varied by number or case. Seventy-five tags (11 percent) were either broader or narrower than a subject heading term. LibraryThing tags were more likely to be broader than LCSH than narrower, but the latter did occur in some cases. Tags identified the same concept as an LC heading or subheading but used different terminology in 112 (16.4 percent) of the cases. Slightly more than half of the tags, 369 (53.9 percent), identified terms or concepts not covered by the LCSH in the catalog record. Conversely, 56 library subject headings, an average of 1.12 per record, did not have a corresponding tag in the LibraryThing record.

Looking more closely at the subject headings and tags assigned to individual works, some judgments can be made regarding the appropriateness, accuracy, and thoroughness of the terms assigned by catalogers and end users alike. Users sometimes can provide more detailed and accurate information about the subject content of a work compared to a cataloger because the former has often read the book in

Table 2. LibraryThing Tags Compared to LCSH

Category of Match	No. of LibraryThing Tags	% of LibraryThing Tags
Exact LCSH Match	61	8.9
Partial LCSH Match	67	9.8
No Match: Specificity	75	11.0
No Match: Vocabulary	112	16.4
No Match: New Concept	369	53.9
Total	684	100.0

question, whereas a cataloger typically must rely on publisher-supplied information or a quick perusal of the book. On the other hand, the terminology employed by end users is often inaccurate or imprecise compared to that used by catalogers. A few examples from the sample can help illustrate this. The LibraryThing record for *Field of Honor* by Donn Byrne has the tag "Biography." This is technically incorrect, because the book is a work of biographical fiction. However, to an end user this distinction may be less critical, depending on whether they are reading the book for research or recreational purposes. Other examples of tags misidentifying the literary form of a work in LibraryThing are a bit more puzzling: *The Rear Column* by Simon Gray labeled as "Fiction" rather than "Drama," or John Berryman's *Homage to Mistress Bradstreet* tagged as "Play" instead of "Long-form poem." However, the application of LCSH in the library catalog also sometimes falls short. In the record for *The Affair* by Ronald Millar, the form subdivision "Fiction" is used instead of "Drama."

In many cases, the library subject headings and LibraryThing tags are complementary. The catalog record for *Bread Givers* by Anzia Yezierksa has the following subject headings:

- Fathers and daughters—United States—Fiction
- Children of immigrants—United States—Fiction

While the LibraryThing record for the same title includes tags about immigrants and fathers and daughters, the tags also specify that the immigrant characters are Jewish and the specific location is the Lower East Side of New York City.

The catalog record for *Perdido Street Station* by China Miéville has the following subject headings:

- Dissenters—Fiction
- City and town life—Fiction

The LibraryThing record provides the following tags, among others: "Dark fantasy," "Dystopia," "New weird," "Speculative fiction," "Steampunk," and "Urban fantasy."

In these examples, the subject headings and tags together provide a more complete view of the nature and thematic elements of the work than either source does alone.

Availability of Tags

One aim of this paper was to explore whether LibraryThing can provide useful subject tags to literary works that lack subject headings in library catalogs. LibraryThing tags were available for 191 catalog records in the sample that had no subject headings. On closer examination of the LibraryThing records, 5 were found to have only personal tags (i.e., tags that would not be useful to anyone other than the tagger), so these records were removed from the sample.

For the remaining 186 records, the existing nonpersonal or nondescriptive tags were categorized by type. As with the sample of records comparing LCSH with tags, only the tags appearing in the initial view of the tag cloud were examined for those records having more than 30 tags. Each tag was identified as being a either a broad form or genre term (broader than would typically be assigned to an individual literary work by a cataloger), a specific form or genre term, a geographic term, a topical term, a chronological term, or a character name. Frequency of various types of tags also was compared according to literary form.

Table 3 presents details about the 2,304 usable tags found in these 186 LibraryThing records. Although this averages 12.4 per record, the number of tags per record varied greatly, with some records having only a few tags and others having tags numbering into the hundreds. The reliability and accuracy of the tags grows for more heavily tagged works because multiple users affirm and verify tags by using them on their own copies of the records. In the sample, 107 records were for novels, 40 for poetry, 16 each for drama and short story collections, and 7 for other types of works. The average number of tags per record was the highest for novels, at 15.4. Poetry and drama typically had the fewest tags, averaging 7.0 and 6.9 per record, respectively.

The highest percentage of 2,304 LibraryThing tags in records for the 186 titles that lacked subject headings in the library catalog were for very broad form and genre tags. Tags such as "Literature," "Fiction," "Novel," "Classic," "Poetry," and "Plays" were very common across the sample, and they were often among the most popular tags (indicated in a tag cloud by larger and bolder type) for a given title. Tags of this type accounted for 981 (42.6 percent) of all the tags examined. More specific form and genre terms, more analogous to the types of terms a cataloger might put in a 650 or 655 MARC field, accounted for 464 (20.1 percent) of the tags. Examples include "African American fiction," "Horror," "Science fiction," and "Crime and mystery." Topical terms made up 487 (21.1 percent) of the tags. Less common were geographic terms, of which there were

212, many of them very general, such as "America," "USA," or "UK." Chronological terms were supplied 151 times, sometimes specific to a decade or a historical period, such as World War II, other times as general as "20th century." Whether the tagger was indicating the time period depicted in the work or the time period in which the work was published was not always clear, but for consistency's sake the former was assumed for all records. Only 9 character names were supplied as tags in the 186 records, accounting for less than 1 percent of all tags. A categorization of LibraryThings tags assigned to records that had no LCSH in MNCAT is presented in figure 3.

Thirty-two (17.2 percent) of the records in the sample of 186 titles had only broad form or genre tags in the LibraryThing record. This situation was most common for poetry and drama—32.5 percent of poetry records and 37.5 percent of records for plays had only very general tags, nothing that could aid users in determining the nature, theme, or subject matter of the work at hand. Only 10.3 percent of novels and 12.5 percent of short story records lacked specific tags in addition to broad form and genre terms.

As expected, some variation exists in the proportion of different types of tags across the different literary forms. However, since the sample size is small for some of the forms, particularly drama, short stories, and the miscellaneous forms, a reliable conclusion cannot be confidently drawn.

As shown in figure 4, broad form and genre terms are by far the most common type of LibraryThing tag for all of the literary forms in the sample of 186 items. This is particularly true for drama, with 68.5 percent of all tags being general terms such as "Drama," "Plays," or "Theatre." Specific form and genre terms and topical terms are the next most common, with a slight variation according to literary form of the work. Geographic and chronological terms make up less than 10 percent of tags for all literary forms, with the exception of geographic terms for novels, which account for 10.8 percent of the tags. The least common tag type was character name, with only 8 character tags being added to records for novels, and one for a work of poetry.

Discussion

The results have demonstrated that the number of user tags in LibraryThing is high for the types of twentieth- and twenty-first-century English-language literary works found in a large academic library—the overlap of matching work

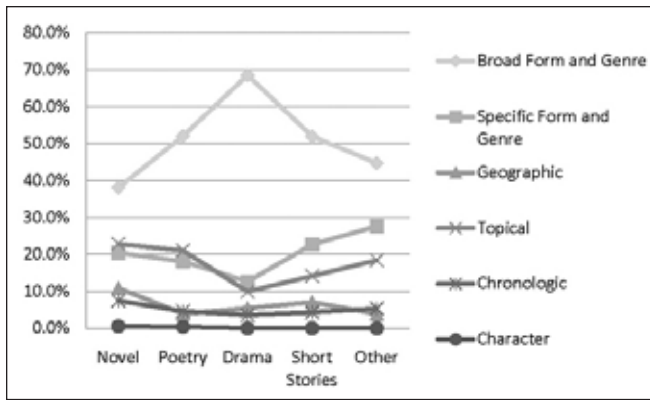**Table 3.** LibraryThing Tags by Literary Form

Literary Form	Records in Sample	% of Sample	Total LibraryThing Tags Assigned	Average LibraryThing Tags per Record
Novel	107	57.5	1651	15.4
Poetry	40	21.5	281	7.0
Drama	16	8.6	111	6.9
Short Stories	16	8.6	185	11.6
Other	7	3.8	76	10.9
Total	186	100.0	2304	12.4
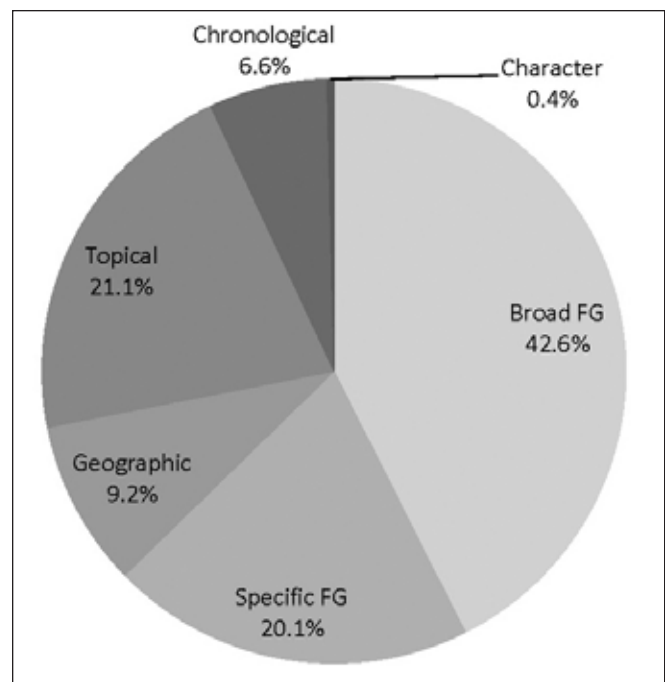
Figure 3. LibraryThing Tags by Type for Items with no LCSH in MNCAT (N = 2,304)

records between the library catalog and LibraryThing is 82.7 percent. However, the availability of records in LibraryThing for fiction is notably higher than for other literary forms, such as drama and poetry.

The record sampling also showed the lack of subject access for many twenty- and twenty-first-century literary works in academic library catalogs; 61 percent of the catalog records in the sample had no subject headings. Conversely, only 7.1 percent of the 367 work records in LibraryThing had no tags. Records with tags were found in LibraryThing for 70.4 percent of the catalog records lacking subject headings, while another 150 records had subject headings in the catalog as well as LibraryThing records with tags. This high rate of tag availability (76.8 percent of the full sample of 444

Figure 4. LibraryThing Tag Type by Literary Form (N = 186)

records) indicates a strong potential for user tags to enhance the subject access to twentieth- and twenty-first-century English-language literary works in academic library catalogs.

However, a closer look at the LibraryThing tags reduces this potential somewhat. While personal tags are heavily used in LibraryThing, these are eliminated in LibraryThing for Libraries (LTFL), and their prevalence was less problematic than the author anticipated when gathering the data for this study. Only 5 records examined had no usable tags for comparison. What was more discouraging for library purposes was the prevalence of general terms such as "American literature," "Fiction," and "Poetry." This seems to indicate that a lower percentage of the tags than originally thought might be useful to librarians and end users, even when personal tags are eliminated. This is especially true with drama and poetry, where a third or more of all the records examined contained only broad terms, and fewer total tags were assigned per record. Conversely, many records for novels in LibraryThing contained a rich tag cloud of terms. While broad, general tags are often the most popular and heavily used by LibraryThing users, consistent with tag distribution found in other online tagging communities, additional terms bringing out topical and thematic elements are abundant for some works.

A library wishing to utilize end user tags from LibraryThing or another source has several options for incorporating the tags into the library's public catalog interface. A number of next-generation catalog interfaces, such as Innovative Interfaces's Encore or Ex Libris' Primo, can show tag clouds as part of the record display. Depending on the technical capabilities of the particular discovery software, the tags may be searchable as keywords or usable for browsing. For example, a user viewing a record that displays the tag "Jewish immigrants" might be able to click on that tag to find other library resources on that topic. Next-generation interfaces typically encourage end user interaction, including tagging. While evidence shows that user tagging is more

popular in national or multi-institutional sites than in site-specific library catalogs, seeding the catalog with tags from a source like LibraryThing might motivate more catalog users to begin tagging materials in the catalog for their own use or to aid others in finding and identifying materials of interest.[39]

When tags are used in the discovery layer of a library's catalog, they are not in any way part of the catalog database that underlies the user interface. For several reasons, including accuracy, appropriateness, or privacy, a library may not wish to have user data as part of catalog records, but one disadvantage to including user data is that the data will not easily migrate if the library converts to a new library catalog system. Advances in cloud computing and network-level data sharing make this less of an issue than it might have been in the past, but libraries should be aware that migrating discovery layer data is a separate consideration from converting catalog record data from system to system.

If a library does wish to use tags to enhance catalog record data directly, it might accomplish this in several ways. The MARC field 653 is for an uncontrolled index term. More fully defined, it is intended for an "index term added entry that is not constructed by standard subject heading/thesaurus-building conventions."[40] User tags could be imported into 653 fields to provide keyword search access. Another option is the 69X fields, which are "reserved for local subject use and local definition."[41]

Some libraries might be reluctant to incorporate user tags directly into their catalog records, especially if they have concerns about accuracy of terminology or appropriate level of specificity. But for titles that have rich, detailed tag clouds, user tags could aid catalogers in creating LCSH. Tags might identify thematic elements that are not readily apparent from publisher-supplied information on dust jackets or back cover copy, especially because a LibraryThing tagger is more likely to have read the entire book, as opposed to a cataloger downloading or creating a record at the time of a book's receipt by the library.[42] If an appropriate level of confidence in the accuracy of the tags can be reached, tag terms could be converted to equivalent LCSH terms and placed in 65X fields. In addition, if particular terms seem prevalent among users across a number of titles, a cataloger could consider submitting a proposal to the Subject Authority Cooperative Program (SACO) (www.loc.gov/catdir/pcc/saco) to add the term to LCSH, add it as a cross-reference to an existing heading, or modify an existing heading.

Researchers could further explore a number of areas. For library catalogs that are already using LTFL or other imported tag data, the actual usage of the tags could be investigated. Does the presence of tag clouds increase tagging behavior by users? Do the tags provide keywords not found elsewhere in the catalog record? How does retrieval by tag terms compare with keyword retrieval from other

types of record enhancements, such as tables of contents, summaries, or reviews?

Researchers could investigate other social networking sites devoted to books and reading, such as GoodReads (www.goodreads.com) and Shelfari (www.shelfari.com). Hit rates on user supplied data could be compared with library catalogs or LibraryThing to determine relative hit rates and quality of user-supplied content.

While tagging in individual library catalogs has not been overwhelmingly popular, researchers could further explore user motivation. Why have some cultural heritage tagging projects been highly successful, such as Flickr Commons (www.flickr.commons) and Steve: The Museum Social Tagging Project (www.steve.museum)? Something in the presentation of the opportunity for end users to help these cultural institutions has inspired an altruistic spirit in unexpected numbers of members of the general public. Could academic libraries find a way to encourage users to tag by appealing to a feeling of community around the institution's history or traditions? Or is the typical university student or faculty member too focused on his or her own research and concerns to wish to allocate time toward an unproven common good?

Research suggests that fewer library users are starting their search in a local library catalog.[43] Many users begin their search with network-level resources, such as Google, Amazon, or WorldCat. How does this affect the need for enriching data in individual catalogs? Should libraries invest instead on centralizing library catalog data and providing links to local information, such as availability?

Conclusion

The purpose of this research was to determine whether user-supplied tags for twentieth- and twenty-first-century literary works in English could enhance or complement the controlled subject headings for the same works in an academic library catalog, and to explore the quality and accuracy of the user tags compared to controlled vocabularies. The author compared a sample of records from a large academic library catalog with corresponding work records in a popular social networking site, LibraryThing. Records were found in LibraryThing for a large number of the works selected (82.7 percent). Nearly 90 percent of fiction was found in LibraryThing, while match rates for other literary forms ranged from 68 to 81 percent. Because of longstanding library practices and traditions, 61 percent of the library catalog records lacked subject headings. Of this subset of works in LibraryThing, 70.5 percent had tags. In addition, 150 works in the full sample had both tags and subject headings.

A comparison of types of tags assigned by end users and library-assigned subject headings found that the most commonly used tags tended to be broader than the controlled vocabulary terms in the catalog records. Very popular or classic works of fiction were more likely to have rich, extensive lists of tags, including many specific terms not found in the subject headings. In many of these cases, the library catalog record and the LibraryThing record provided complementary subject access. Less well-known works, however, and those in literary genres such as poetry or drama, tended to have few if any tags in LibraryThing, and the tags assigned were often very general.

For works that had LibraryThing tags but no LCSH, 42.5 percent of the tags assigned were broad genre terms, and many of the chronologic and geographic tags assigned also were very broad (e.g., "U.S.," "20th century"). Approximately a third of drama and poetry records had only broad tags, while a majority of fiction works had more specific terms as well as broad terms assigned.

These findings seem to confirm previous studies of end user bookmarking and tagging. General terms may be adequate for personal use or for small, personal book collections. For a large academic library collection, however, broad terms such as "Novel" or "Poems" have less utility in refining searches across many thousands of hits. An academic library wishing to utilize a service such as LTFL might be better served by targeting a smaller subset of records, such as a popular reading collection.

As discovery systems and user behaviors continue to evolve, those working with catalog data must continue to explore ways to enhance access to information resources of all kinds and to improve the user experience. Capitalizing on aggregated end user data about resources is only one way to achieve that aim.

References

1. Karen Smith-Yoshimura, "Social Metadata for Libraries, Archives, and Museums" (presentation, DLF Fall Forum, Palo Alto, California, 2010), www.clir.org/dlf/forums/fall2010/SocialMetadataforLAMs.pdf (accessed Mar. 4, 2011).

2. Jim Michalko, "Things That Happen Elsewhere—User Studies Say," online posting, June 5, 2009, hangingtogether.org, http://hangingtogether.org/?p=702 (accessed Aug. 30, 2010).

3. Thomas Vander Wal, "Folksonomy Coinage and Definition," www.vanderwal.net/folksonomy.html (accessed Dec. 29, 2010).

4. Lois Mai Chan, "Interindexer Consistency in Subject Cataloging," *Information Technology & Libraries* 8, no. 4 (1989): 349–58; Jarmo Saarti, "Consistency of Subject Indexing of Novels by Public Library Professionals and Patrons," *Journal of Documentation* 58, no. 1 (2002): 49–65.

5. Markus Heckner, Suzanne Mühlbacher, and Christian Wolff, "Tagging Tagging: A Classification Model for User Keywords in Scientific Bibliography Management Systems" (presentation, Networked Knowledge Organization Systems and Services, 6th European Networked Knowledge Organization

Systems (NKOS) Workshop at the 11th ECDL Conference, Budapest, Hungary, 2007), www.comp.glam.ac.uk/pages/research/hypermedia/nkos/nkos2007/papers/heckner.pdf (accessed Aug. 30, 2010).

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

7. Scott A. Golder and Bernardo H. Huberman, "Usage Patterns of Collaborative Tagging Systems," *Journal of Information Science* 32, no. 2 (2006): 198–208.

8. Joseph B. Dalton, "Can Structured Metadata Play Nice with Tagging Systems? Parsing New Meanings from Classification-Based Descriptions on Flickr Commons," in J. Trant and D. Bearman, eds., *Museums and the Web 2010: Proceedings*. (Toronto: Archives and Museum Informatics, 2010), www.archimuse.com/mw2010/papers/dalton/dalton.html (accessed Aug. 30, 2010).

9. LibraryThing, Zeitgeist Overview, www.librarything.com/zeitgeist (accessed May 22, 2011).

10. LibraryThing for Libraries, FAQs: General, www.librarything.com/forlibraries/about (accessed Dec. 29, 2010).

11. Ibid.

12. Christine DeZelar-Tiedman. "Doing the LibraryThing™ in an Academic Library Catalog," poster abstract in *Metadata for Semantic and Social Applications, DC-2008, Berlin, Proceedings of the International Conference on Dublin Core and Metadata Application, 22–26 September 2008*, ed. Jane Greenberg and Wolfgang Lkas (Singapore: Dublin Core Metadata Initiative; Gottingen: Universitatsverlag Gottingen, 2008): 211, http://webdoc.sub.gwdg.de/univerlag/2008/DC_proceedings.pdf (accessed Nov. 21, 2010).

13. Christine DeZelar-Tiedman, "Subject Access to Fiction: An Application of the *Guidelines*," *Library Resources & Technical Services* 40, no. 3 (1996): 203–10.

14. Susan M. Hayes, "Use of Popular and Literary Criticism in Providing Subject Access to Imaginative Literature," *Cataloging & Classification Quarterly* 32, no. 4 (2002): 71–97.

15. American Library Association, Resources and Technical Services Division, Cataloging and Classification Section, Subject Analysis Committee; Subcommittee on Subject Access to Individual Works of Fiction, Drama, etc., *Guidelines on Subject Access to Individual Works of Fiction, Drama, etc.* (Chicago: ALA, 1990).

16. Mary Dabney Wilson et al., "The Relationship between Subject Headings for Works of Fiction and Circulation in an Academic Library," *Library Collections, Acquisitions & Technical Services* 24, no. 4 (2000): 459–65.

17. Association for Library Collections & Technical Services, Cataloging and Classification Section; Subject Analysis Committee, Subcommittee on the Revision of the Guidelines on Subject Access to Individual Works of Fiction, Drama, etc., *Guidelines on Subject Access to Individual Works of Fiction, Drama, etc.*, 2nd ed. (Chicago: ALA, 2000).

18. Library of Congress, Cataloging Policy and Support Office, *Subject Cataloging Manual: Subject Headings*, 5th ed. (Washington, D.C.: Library of Congress, 1996–).

19. Hayes, "Use of Popular and Literary Criticism in Providing Subject Access to Imaginative Literature," 95.

20. Emma Tonkin et al., "Collaborative and Social Tagging Networks," *Ariadne* 54 (Jan. 2008), www.ariadne.ac.uk/issue54/tonkin-et-al (accessed Aug. 10, 2010).

21. Margaret E. I. Kipp, "Complementary or Discrete Contexts in Online Indexing: A Comparison of User, Creator, and Intermediary Keywords," *Canadian Journal of Information & Library Science* 29, no.4 (Dec. 2005): 419–36.

22. Louise F. Spiteri, "The Structure and Form of Folksonomy Tags: The Road to the Public Library Catalog," *Information Technology & Libraries* 26, no.3 (Sept. 2007): 13–25.

23. Rob Hidderley and Pauline Rafferty, "Democratic Indexing: An Approach to the Retrieval of Fiction," *Information Services & Use* 17, no. 2/3 (1997): 103.

24. Pauline Rafferty and Rob Hidderley, "Flickr and Democratic Indexing: Dialogic Approaches to Indexing," *Aslib Proceedings: New Information Perspectives* 59, no. 4/5 (2007): 397–410.

25. Timme Bisgaard Munk and Kristian Mørk, "Folksonomy, the Power Law and the Significance of the Least Effort," *Knowledge Organization* 34, no. 1 (2007): 19.

26. Lada A. Adamic, "Zipf, Power Laws, and Pareto: A Ranking Tutorial" (Palo Alto, Calif.: Information Dynamics Lab, 2002), www.hpl.hp.com/research/idl/papers/ranking/ranking.html (accessed Dec. 29, 2010)

27. Munk and Mørk, "Folksonomy," 28–29.

28. Golder and Huberman, "Usage Patterns of Collaborative Tagging Systems," 204.

29. Michelle Springer et al., *For the Common Good: The Library of Congress Flickr Pilot Project* (Washington, D.C.: Library of Congress, 2008), www.loc.gov/rr/print/flickr_report_final.pdf (accessed Aug. 10, 2010).

30. Ibid., iv.

31. Ibid., 2.

32. Trant, "Exploring the Potential for Social Tagging and Folksonomy in Art Museums"; Jennifer Trant, *Tagging, Folksonomy and Art Museums: Results of steve.museum's Research* (Toronto, Ontario: Archives & Museum Informatics, 2009).

33. Brian Matthews et al., "An Evaluation of Enhancing Social Tagging with a Knowledge Organization System," *Aslib Proceedings: New Information Perspectives* 62, no. 4/5 (2010): 447–65.

34. Mikael Wetterstrom, "The Complementarity of Tags and LCSH: A Tagging Experiment and Investigation into Added Value in a New Zealand Library Context," *New Zealand Library & Information Management Journal, Ng Prongo* 50, no. 4 (2008): 296–310.

35. C. Rockelle Strader, "Author-Assigned Keywords versus Library of Congress Subject Headings: Implications for the Cataloging of Electronic Theses and Dissertations," *Library Resources & Technical Services* 53, no. 4 (Oct. 2009): 243–50.

36. Peter J. Rolla, "User Tags versus Subject Headings: Can User-Supplied Data Improve Subject Access to Library Collections?" *Library Resources & Technical Services* 53, no. 3 (July 2009): 174–84.

37. Matt Weaver, "Contextual Metadata: Faceted Schemas in Virtual Library Communities," *Library Hi Tech* 25, no. 4 (2007): 579–94.

38. Robert V. Krejcie and Daryle W. Morgan, "Determining Sample Size for Research Activities," *Educational & Psychological Measurement* 30 (Autumn 1970): 607–10.

39. Smith-Yoshimura, "Social Metadata for Libraries, Archives, and Museums," 1.

40. Library of Congress, MARC 21 Format for Bibliographic Data, 653—Index Term-Uncontrolled (R) (2/26/2008), www.loc.gov/marc/bibliographic/bd653.html (accessed Aug. 27, 2010).

41. Library of Congress, MARC 21 Format for Bibliographic Data, 69X—Local Subject Access Fields (R), www.loc.gov/marc/bibliographic/bd69x.html (accessed Aug. 27, 2010).

42. Margaret Beecher Maurer, "Social Tagging, Folksonomies and Controlled Vocabularies—Can't They Just be Friends?" *TechKNOW* 15, no. 1 (June 2009): 9–13.

43. OCLC, *OCLC White Paper on the Information Habits of College Students: How Academic Librarians Can Influence Students' Web-Based Information Choices*, June 2002, www5.oclc.org/downloads/community/informationhabits.pdf (accessed Mary 24, 2011); Karl V. Fast and D. Grant Campbell, "'I Still Like Google': University Student Perceptions of Searching OPACs and the Web," *Proceedings of the 67th ASIS&T Annual Meeting* 41 (2004): 138–46; Cathy DeRosa et al., *Perceptions of Libraries, 2010: Context and Community: A Report to the OCLC Membership* (Dublin, Ohio: OCLC, 2011), www.oclc.org/US/EN/reports/2010perceptions/2010perceptions_all.pdf (accessed May 24, 2011).