
This study examines the overlap of user-generated tags from LibraryThing and subject headings from a library catalog for the most circulated monographs in history and political science at Duke University from 2007-2011. By examining the overlap, implications are found that suggests the presence of user-generated metadata can have impacts on circulation rates in academic libraries.

Headings:

- Subject headings
- Metadata
- Subject cataloging
- Library catalogs and users
- Catalogs, Library
IMPACT OF USER-GENERATED METADATA ON CIRCULATION RATES IN ACADEMIC LIBRARIES

by
B. Young

A Master’s paper submitted to the faculty of the School of Information and Library Science of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Science in Library Science.

Chapel Hill, North Carolina
April 2012

Approved by

_______________________________

Robert Losee
# Table of Contents

- Introduction .................................................................................................................. 2
- Literature Review ......................................................................................................... 4
- Methodology ................................................................................................................ 12
- Data and Analysis ....................................................................................................... 15
- Implications ................................................................................................................ 19
- References .................................................................................................................. 21
Introduction

To loosely borrow from Walter Ong, libraries may be considered the storehouses of cultural knowledge as stored in various forms of media. The librarian’s role is to mediate, whether directly through research services or indirectly through technical services, the user’s discovery of these resources. Librarians and library users, alike, often use the catalog as the tool for locating both known-items and unknown-items, such as those required for investigating a particular topic. Users, though, often declaim the catalog as being difficult to use, and subsequently, librarians often conduct instruction sessions intended to, in part, introduce and demonstrate how to navigate the intricacies of the catalog. Digital tools and other technological developments can be leveraged to improve the usability of the catalog from the non-librarian’s perspective.

Library catalogs, especially subject heading access to items in the catalogs, depend on controlled vocabularies, from which catalogers assign descriptors to resources. Controlled vocabularies refer to a defined list that “introduces a measure of control over the terms used” (Foskett, 1996, p. 113). Catalogers depend on the Library of Congress Subject Headings (LCSH) controlled vocabulary when assigning descriptors to resources for subject access. The formalized nature of controlled vocabularies is contrary to the natural language searching, that is, searching for terms as they appear in documents, enabled by modern search engines and accustomed to by the majority of library users.
Uncontrolled vocabularies, unlike controlled vocabularies, are not formalized and are based more on natural language. Tagging systems, such as those found at LibraryThing and Delicious, are examples of uncontrolled vocabularies where users have organically-created classifications. These classifications depend on tags: user-generated annotations assigned to the resources found on the sites (books and websites, respectively, for LibraryThing and Delicious). The subject tags, as opposed to more personal tags such as own or read later, for books at LibraryThing provide libraries the opportunity to compare how professionally-generated subject headings from LSCH compare to the user-generated terms.

Online catalogs remain the gateway to library resources. Hence, searchers’ ability to navigate the catalog to find a resource should have a direct impact on the circulation on that resource. Other researchers have discussed the impact of enriching MARC records, and this study proposes to continue that research by looking at one potential source for uncontrolled keywords to populate the 653 MARC field. Specifically, this study proposes that non-fiction monographs whose MARC records include at least two user-generated subject tags have higher circulation rates at academic libraries than non-fiction monographs that include zero or one user-generated subject tags.
Literature Review

Online library catalogs are tools through which users locate monographs and other materials. To provide results to search queries, catalogs compare the user’s search terms against the available indexed information. Known-item search queries, such as searching by author or title, are typically straightforward for users as they are, generally, factual information with only derivations of names to complicate searching. Most queries, though, result in a more active process occurring between the searchers and the information system. These active searches can require several iterative steps consisting of the searcher querying and re-querying based on the results (Foskett, p. 25, 1996). Though Foskett wrote about searching well before the implementation of modern-day catalogs, the idea may even be more pertinent today due to the keyword searching process often used by library catalog users.

Catalogs are created based on the structure of controlled vocabularies with Library of Congress Subject Headings (LCSH) being the primary source. The establishment of standards and rules is intended to allow for precise cataloging of an item based on its content (Dezelar-Tiedman 2011). While useful for grouping similar items, users may not always approach a topic from the same perspective, or they may use terms different than those available in the controlled vocabulary when conducting their search. Controlled vocabularies, then, are not ideal for keyword searching as they only provide a limited number of words to be searched.
This constraint can result in unsuccessful searches because, as Borgman notes, non-expert users face the difficult challenge of creating search queries to match terms available in the catalog’s vocabulary (Borgman 1996). Belkin, Oddy, and Brooks (1982) question the possibility for searchers to match their query to the indexed versions of the resources. This anomalous state of knowledge requires users to be able to construct a query in such a way that is readable by the information system, rather than the system being able to facilitate the search process. Further, Bates (1986) demonstrated that it is psychologically difficult for individuals to ascribe the same meaning to commons words and to describe straightforward operations in the same manner. Additionally, some entries require scope notes to distinguish nuances between two headings, such as alcoholism and drunkenness (Foskett, p.341, 1996). If disambiguation is required in order to appropriately catalog, then it should be expected that novice users will have difficulty formulating queries to return desired results. Hence, searchers, particularly new library users, such as undergraduates, can have difficulty navigating the vocabulary and searching via keywords.

Compounding the problem of unfamiliarity with controlled vocabularies is the advent of modern search engines, namely Google, which emphasize keyword searching and rely on their algorithms to return relevant results. Now, introductory library users have had prior experience searching information systems and maintain those paradigms when interacting with the catalog (Yu and Young 2004). Individuals commonly use these other search interfaces that require no prior understanding to accomplish daily tasks, thus they expect library platforms to provide similar experiences (Sadeh 2007). Preference for keyword searching, though, is not singularly related to recent technological developments
as Larson (1991) showed in his study on search behavior, which demonstrated users’
tendency for keyword searching over subject searching.

Larson also reported that Hildreth (1987) and Hunter (1991) found that subject
searching was more often than not unsuccessful, as searches resulted in non-zero hits
only 37% and 37.9% of the time. Lancaster et. al. (1991) further shows the lack of recall
provided by subject headings. Recently, libraries have begun introducing new catalog
interfaces, commonly called next-generation catalogs or discovery platforms, which are
intended, in part, to mimic online search behavior. Part of this trend is the introduction of
additional elements to the MARC record that enables successful keyword searching.

Next-generations catalogs, such as those powered by Endeca’s Information
Access Platform, are one potential way to overcome the controlled vocabulary dilemma.
In addition to introducing commonly found search features (relevance ranking, stemming,
etc.), these catalogs facilitate access to subject headings through facets (Antelman 2006).
Svenonius discussed the potential role of browsing, via facets, as a method for improving
search recall when conducting subject searches (1983). Facet-based searching offers
users the opportunity to refine their initial search based on pre-defined categories,
including subject headings and subdivisions (Kornegay, p. ix, 2009). Post-coordinated
searching, such as faceted browsing, does not require the user to have prior knowledge of
the underlying structure. Marchionini (1995) and others have extolled the virtues of
browsing as an effective technique for finding items. Fagan (2010) offers an extensive
review of the faceted browsing literature, and in it, notes Hearst’s (2006) conclusions that
facets are a “proven technique for supporting exploration and discovery.” Additionally,
Walsh (2012) finds that the literature on next-generation catalogs generally reports
positive feedback from users during usability studies. Fagan, though, concludes that “user studies about faceted browsing revealed empirical evidence that faceted browsing improves user performance, “[y]et this evidence does not necessarily point directly to user success in faceted library catalogs, which have much more complex databases than those used in the experimental studies” (Fagan 2010).

The literature reveals that catalogs have traditionally been difficult to use and that the underlying controlled vocabulary should not be considered entry-level (Foskett, p. 25, 1996). While an active, iterative process is likely to yield successful results, introductory users, today, are accustomed to more friendly interfaces and their expectations do not match the capabilities of catalogs. Harping (2010), in her book on controlled vocabularies for cultural heritage institutions, refers to the expected difficulties of end users who are not specialists in using controlled vocabularies and may be unable to identify the appropriate term (i.e., which synonym is used). Facets, though, introduce an easier method for exploring the library’s resources, yet they still rely on the user’s initial query for recalling a corpus of records to be browsed. This expanded emphasis on recall and attention to user search behavior, has led some libraries to enhance their MARC records with additional metadata in order to improve the usability of the catalog.

Additional metadata can include many different aspects, such as tables of content, summaries, or alternate descriptors, of the material being described. Morris (2001) describes a report from the Subject Access Project in 1978 as an initial spark for discussing the enhancement of catalogs. Once catalogs became digitized, this became much easier to accomplish, and thus, other researchers have continued to investigate. Morris herself concluded that the presence of tables of content increased the likelihood of
monographs being used by 45%. Similarly, Tosaka and Weng (2010) found that enhanced records (presence of tables of content and/or summaries) for titles published between 1990 and 2004 were used between 30.7% and 45.5% more often than monographs without enhanced records. Moreover, Markey and Calhoun (1987) demonstrated that tables of content and summaries added roughly fifteen unique index terms per bibliographic record. While studies have concluded that enriching catalog records can increase the number of access points and circulation rates for monographs, little research has been conducted regarding the addition of user-generated metadata and the impacts it may have on collection usage.

Enhancing MARC records should be seen as complementary to traditional subject searching. Librarians can, and should, still emphasize the importance of subject heading facets during library instruction and reference consultation sessions as these are an invaluable tool for refining searches and eliminating unneeded items. Researchers, however, consistently present keyword and subject searching as a dichotomy. Mann (2005), for example, extols the virtues of subject searching for its ability to aggregate similar items and increase precision, while dismissing keyword searching as being untenable for dedicated researchers. Gross and Taylor (2005) demonstrate the need for subject terms to return search results during keyword searches. Others, such as Shirky (2005), posit that controlled vocabularies are unneeded now that libraries have online catalogs and that controlled vocabularies may even lead to users missing relevant sources. Moreover, Lakoff (1990) discusses the importance of tags being able to capture how non-specialists view items, rather than trained catalogers. These collective classifications, though, also include subjective descriptions not related to aboutness
(good, long) and personal indicators (to be read). The lack of conclusive evidence and myriad opinions among experienced information professionals demonstrates the need for further investigation into the potential benefits of adding and/or linking to additional metadata in the form of uncontrolled terms.

The fundamental differences between controlled and uncontrolled vocabularies are implied by the name of each. Controlled vocabularies depend on consensus as to the appropriate term for a particular concept. Taylor (2004) recognizes the fact that rarely do true synonyms exist; however, individuals often interchange words for concepts that are related and/or similar. As Haykin (p.7, 1951) notes, “a subject catalog must bring together under one heading all the books which deal principally or exclusively with the subject, whatever he terms applied to it by the authors of the books and whatever the varying terms applied to it at different times.” Controlled vocabularies, therefore, enable searchers to feel confident that all resources related to a given concept are grouped under one heading. Moreover, controlled vocabularies enable disambiguation of polysomic words and consistent handling of degrees of specificity (Golder and Huberman 2006). Disambiguation is often performed by the addition of parenthetical statements (Taylor 2004). Researchers have also discussed the merits of controlled vocabularies, in relation to uncontrolled vocabularies, of being able to provide guidelines for appropriate usage of plurality and specificity (Rolla, 2009, Mathes, 2004, Guy and Tonkin, 2006). The structure provided by controlled vocabularies, specifically LCSH, allows users to quickly narrow searches by applying qualifiers, such as geographic terms (- United States) or format (- case studies), to initial queries. For experienced researchers who are
accustomed to the vocabulary, this allows ready access to materials cataloged under these headings.

Uncontrolled vocabularies embrace less formal approaches to subject searching by creating systems geared more towards keyword searching. User predilection to this form of searching warrants investigation into supplementing subject headings with other terms. Haykin (p. 7 1951) notes that the fundamental purpose of catalogs is to facilitate access, which supersedes “conformity to a chosen logical pattern.” Tagging systems, such as those found on websites like LibraryThing, Delicious, and Connotea, embrace uncontrolled vocabularies, depending on user-generated lists that organically form a vocabulary. As already discussed, relying on user-generated terms should increase the likelihood that a user’s query will find an indexed term as these descriptors tend to represent the ways users themselves describe the resources. Foskett (p. 112, 1996), while discussing the need for consistency, even notes the difficulty for indexers to categorize resources consistently due to contextual (background, frame of reference, etc.) differences. Hence, it should not be surprising that users may find it difficult to query an interface with a term that will lead to the correct subject heading used by a controlled vocabulary. Lancaster (p. 23, 1977) discusses the advantages inherent by enabling users to search using the language in which they communicate. Ultimately, libraries must consider whether the benefit of additional descriptors, resulting in an increase in recall, warrants inclusion at the risk of potentially losing precision or increasing the need for modifying online catalog features to allow for these additional descriptors.

Rolla (2009) investigated this potential by analyzing the records on the website LibraryThing. Rolla found that LibraryThing records always offer at least one keyword
that can provide access to concepts not covered by the subject headings assigned to the record. Emphasizing the need for user-generated metadata and subject headings to be used in conjunction with each other, Rolla also found that 55.6% of the bibliographic records contained access points to concepts not covered by user tags. Several studies conclude that users more often tag items broader than librarians classify with LCSH (Munk and Mork, 2007, Golder and Huberman, 2006).

This conclusion has direct impact on catalog searching because it indicates differences in how general users and library catalogers view items. For twentieth and twenty-first century English-language literary works, DeZelar-Tiedman (2011) concluded that while more popular works from this genre had rich collections of usable tags that could supplement the resource’s subject headings, lesser-known works lacked tags of sufficient specificity (items described as poetry, for example). Mendes et. al. (2009) investigated the efficacy of LibraryThing for Libraries (LFTL) and found that “for every new book a user discovers using LCSH headings they will discover four books using LFTL.” They further discuss the potential lack of relevancy associated with these titles and the need for precision in the context of academic research. Faceted searching within the results, however, can aid new researchers in refining their search. Additionally, LFTL does not integrate the tags into the catalog, so each set of terms must be searched independently. None of these studies, however, attempt to correlate the presence of user-generated tags to the circulation rates for monographs in academic libraries.
Methodology

This study will focus on non-fiction monographs contained in Duke University Library’s collection and, for the purposes of this study, non-fiction monographs will be defined as those books that lack the subject heading term fiction in LCSH, as displayed in Duke University’s OPAC. Similar to Tosaka and Yeng’s (2010) study, I will be focusing on monographs from two distinct portions of the LCSH vocabulary. The monographs will be limited to history (call numbers DEF) and political science (call number P=J) classifications due to the heavier reliance on monographs by researchers in these areas than others, such as science and engineering.

From these call numbers, I selected the 125 most circulated titles within each call number, resulting in a data set of 250 titles. Circulation rankings were based on statistics extracted from Duke’s local ALEPH ILS for the five years between January 1, 2007 and December 31, 2011 for bibliographic records in the aforementioned classification ranges. Further, checkouts will not include renewals or interlibrary loan fulfillments because this study is only interested in checkouts that may be directly influenced by the online catalog.

This study defines tags as user-generated annotations attributed to an item. In this study those items are books cataloged using LibraryThing, “an online service to help people catalog their books easily” (LibraryThing). The tags for each title will be manually collected by accessing each resource’s respective LibraryThing page and then
extracting the information found under the *Tags* section. Additionally, this study will only focus on subject tags—those tags that describe the aboutness of the resource. More specifically, subject tags will not include personal descriptions (i.e., *read*, *to be read*, etc.) or terms that describe the resource itself and not its content, such as title, publisher, or author. Mainly, this is to ensure that when resources are documented as overlapping, it is due to terms related to subject access, and no other access points, like title and author, which are more useful when conducting known-item searches.

Related to the idea of aboutness, comparisons will be made between top-level subject headings and LibraryThing tags. Limiting to only top-level subject headings, such as only looking at *civil rights* for the subject heading *civil rights -- history*, reflects the search process of academic library searchers who are less likely to use broad terms such as general subject areas when first conducting a search. Moreover, this study counts partial matches, such as a tag *politics* compared to subject heading *world politics*, because a keyword search would return the record. However, the reversal, *world politics* as the tag, does not count as the keyword search for world politics would not return the record.

Once the list of subject tags for each monograph is compiled, they will be compared to the MARC records for their respective monograph. Only specific words contained in both the compiled tag list and designated MARC records will be documented as meeting the hypothesis criteria. As discussed earlier, many instances can occur where user-generated tags are variations of indexed subject headings. Specific examples of instances that do not qualify as successful matches are: acronyms (SCUBA and Self-Contained Underwater Breathing Apparatus), abbreviations (US and United
States), and singular/plural forms (library vs. libraries or horse vs. horses). From these comparisons, each monograph can then be categorized as either containing or failing to contain two or more user-assigned subject tags. For each resource, the number of overlapping terms will also be documented in addition to denoting a simple yes/no.

Once data has been collected, each monograph will be coded as either yes or no depending on whether two or more user-generated subject tags are present. If they are, the resource will be coded yes. After coding, I will be able to determine if those books whose MARC records have two or more overlapping terms have higher circulation rates by reviewing statistical measures, such as mean and median circulations, for both those monographs that have two or more overlapping terms and those that do not. Regression analysis will also be performed to create a model for predicting circulation rates based on number of tags.
Data and Analysis

Before performing any analysis, some records were excluded in order to present an accurate depiction of how overlaps between subject headings and tags can impact circulation. Monographs with either fewer than two subject headings or less than ten LibraryThing “owners” were excluded from the study. The former were excluded due to, by default, failing to meet the assigned criteria and the latter were excluded due to the lack of tags with which to compare to subject headings.

<table>
<thead>
<tr>
<th>Subject Area</th>
<th>Items with 2 or more overlapping terms</th>
<th>Items without 2 or more overlapping terms</th>
<th>Items with less than two subject headings</th>
<th>Items with less than 10 LibraryThing members</th>
</tr>
</thead>
<tbody>
<tr>
<td>History</td>
<td>72</td>
<td>31</td>
<td>5</td>
<td>17</td>
</tr>
<tr>
<td>Political Science</td>
<td>47</td>
<td>45</td>
<td>17</td>
<td>16</td>
</tr>
<tr>
<td>Total</td>
<td>119</td>
<td>76</td>
<td>22</td>
<td>33</td>
</tr>
</tbody>
</table>

Table 1: Breakdown of data

For the remaining 195 items, the majority (61%) had at least two or more user-generated tags that overlapped with cataloger-assigned subject headings. History monographs in the study were 33% more likely than political science monographs to have user-generated tags as subject headings and to have at least two of them as subject headings. The discrepancy between history and political science monographs, though, is
much closer when looking at the weighted percentage, which shows that only 14% more of the total user-generated tags for history monographs are also used as subject headings.

<table>
<thead>
<tr>
<th>Subject Area</th>
<th>Total Items</th>
<th>Items with 2 or more overlapping terms</th>
<th>% of items with 2 or more overlapping terms</th>
<th>Weighted % of subject headings with overlapping tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>History</td>
<td>103</td>
<td>72</td>
<td>69%</td>
<td>55%</td>
</tr>
<tr>
<td>Political Science</td>
<td>92</td>
<td>47</td>
<td>52%</td>
<td>48%</td>
</tr>
<tr>
<td>Total</td>
<td>195</td>
<td>119</td>
<td>61%</td>
<td>52%</td>
</tr>
</tbody>
</table>

Table 2: Summary Table for items with 2 or more overlapping terms

As political science monographs only have 4% fewer (3.43 vs. 3.68) unique, top-level subject headings assigned than history monographs, it seems unlikely that this is the cause of this difference. One other potential explanation is the popularity, in the context of LibraryThing, for history monographs as compared to political science monographs. The median number of users who have added history monographs from this study to their LibraryThing account is 64% greater (146.5 vs. 89) than the number who have added political science monographs. While further study would be needed to fully ascertain the difference, it is important to consider how the number of users tagging items can potentially influence the value of data. Though these are two potential explanations, further study would be necessary to understand the reasons why the political science sample has significantly fewer monographs that have at least two subject headings that match user tags.
To further examine the potential impact of user-generated tags, the mean and median circulations for monographs that have and do not have at least two user-generated subject tags were calculated. The mean and median circulations for monographs in this study were approximately 23.6% and 23.5%, respectively, higher for those monographs that had two or more user-generated terms. Looking specifically at each subject area, the overall trend is still valid except for median circulations for history, which are only slightly higher for those monographs with two or more user-generated tags.

<table>
<thead>
<tr>
<th>Subject Area</th>
<th>Total Circulations</th>
<th>Mean Circulation per item</th>
<th>Median Circulation per item</th>
</tr>
</thead>
<tbody>
<tr>
<td>History</td>
<td>4002</td>
<td>56</td>
<td>46</td>
</tr>
<tr>
<td>Political Science</td>
<td>1561</td>
<td>33</td>
<td>30</td>
</tr>
<tr>
<td>Total</td>
<td>5593</td>
<td>47</td>
<td>42</td>
</tr>
</tbody>
</table>

**Table 3: Circulation rates for items with 2 or more user-generated subject terms**

<table>
<thead>
<tr>
<th>Subject Area</th>
<th>Total Circulations</th>
<th>Mean Circulation per item</th>
<th>Median Circulation per item</th>
</tr>
</thead>
<tbody>
<tr>
<td>History</td>
<td>1591</td>
<td>51</td>
<td>45</td>
</tr>
<tr>
<td>Political Science</td>
<td>1293</td>
<td>29</td>
<td>27</td>
</tr>
<tr>
<td>Total</td>
<td>2884</td>
<td>38</td>
<td>34</td>
</tr>
</tbody>
</table>

**Table 4: Circulation rates for items with less than 2 user-generated subject terms**

To gain an idea of the potential increase in circulation that tags can generate, linear regression was conducted on the data. The resulting equation, $y=2.4119x + 38.804$, as shown in Figure 1, indicates that approximately 2.4 additional circulations, (over a five year period) will occur for each additional user-generated tag that appears as a subject
heading. Given the average circulation rate of 43 items during this same time frame, each additional tag can potentially increase circulation approximately 6%.

![Total Linear Regression](chart)

**Figure 1: Total Linear Regression**

Along with the previous discussion of median and mean circulation rates, the data presented in this study, while not demonstrating a causal relationship between the presence of user-generated tags as subject headings and circulation rates, does demonstrate the need for continued study into the impacts that the presence of additional metadata, like tags, can have on user searching.
Implications

This study has examined the thirty most-attributed tags from LibraryThing and compared them to the subject headings assigned to the most circulated monographs from an academic library during a five-year period. Noticeable differences in the circulation statistics between the two groups (≥2 tags and <2 tags), while not generalizable, suggests that the inclusion of uncontrolled vocabularies in library bibliographic records may have a positive impact on circulation. In the past, libraries restricted the number of assigned terms due to physical limitations of the card catalog, but with today’s electronic catalogs, libraries should investigate the potential benefits that providing additional terms accessible through keyword searching can provide. Faceted catalogs, which diminish the need for precise searching, and increasing number of e-books, where collocation of items on shelves is non-existent, contribute to the need for closer analysis of user search behavior and investigation into how to better support keyword access to resource discovery.

From the study, it seems evident that the potential impact of user-generated tags may differ across disciplines. Future studies, using similar methodology, can further explore this phenomenon and develop not only a more nuanced understanding of how tags impact individual areas but also as to how users within each discipline approach information seeking. Based on the total circulations for the most circulated items, it seems evident that political science monographs are circulated much less frequently than
history monographs. While this may be related more to the discipline than to how the items are cataloged, it still warrants further investigation.

The study as presented is but one way in which to investigate the value of adding user-generated tags. The collected data highlight other opportunities such as the differences between the vocabulary of everyday users and catalogers. For example, within the political science monographs, thirty-six different monographs were tagged political philosophy, which currently is not an available subject heading, although it can be constructed in a manner that would appear in search results by qualifying political science with philosophy. This, however, was only done on three of the thirty-six titles.

Additionally, some phrases, such as State, The or full spellings, such as United States, are unused by LibraryThing users. In the former case, expectations may be that researchers would use the term, but freshmen who are unaccustomed to advanced research would be unlikely to use this term. It is more likely that individuals would use the full name for the U.S., but as the collected data demonstrates, we, as librarians, should still consider supporting discovery for alternate forms, especially those with demonstrated high-frequency of usage.

Future considerations should be given to the addition of uncontrolled terms to the catalog, whether through LibraryThing or other sources. It is important for libraries to maintain a catalog that is both accommodating to the needs of experienced researchers, as well as to those introductory users who have not yet developed the skills needed to fully utilize the existing library catalog. By carefully considering users’ paradigms when approaching library search interfaces, libraries can improve the overall search process.
References


