Tagging Human Knowledge

Preprint Version, Last Updated November 18, 2009

Paul Heymann, Andreas Paepcke, and Hector Garcia-Molina
Department of Computer Science
Stanford University, Stanford, CA, USA
{heymann, paepcke, hector}@cs.stanford.edu

ABSTRACT

A fundamental premise of tagging systems is that regular users can organize large collections for browsing and other tasks using uncontrolled vocabularies. Until now, that premise has remained relatively unexamined. Using library data, we test the tagging approach to organizing a collection. We find that tagging systems have three major large scale organizational features: consistency, quality, and completeness. In addition to testing these features, we present results suggesting that users produce tags similar to the topics designed by experts, that paid tagging can effectively supplement tags in a tagging system, and that information integration may be possible across tagging systems.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.1.2 [**Models and Principles**]: User/Machine Systems—*Human information processing*

General Terms

Experimentation, Human Factors, Measurement

1. INTRODUCTION

In 1994, two students organized pages on the web into what became the Yahoo! Directory. What they did could be caricatured as the "library approach" to organizing a collection: create a limited taxonomy or set of terms and then have expert catalogers annotate objects in the collection with taxonomy nodes or terms from the pre-set vocabulary.

In 1998, the Open Directory Project (ODP) replaced expert catalogers with volunteers, but kept the predetermined taxonomy. Experts were too expensive, and users of the Internet too numerous to ignore as volunteers.

In 2003, a social bookmarking system named Delicious was started. In Delicious, users annotated objects (in particular, URLs) with tags (i.e., "keywords") of their own choosing. We call this the "tagging approach" to organizing a collection: ask users with no knowledge of how the collection is organized to provide terms to organize the collection. Within a few years, Delicious had an order of magnitude more URLs annotated than either Yahoo! Directory or ODP.

Increasingly, web sites are turning to the "tagging approach" rather than the "library approach" for organizing the content generated by their users. This is both by necessity and by choice. For example, the photo tagging site Flickr has thousands of photos uploaded each second, an untenable amount to have labeled by experts. Popular websites tend to have many users, unknown future objects, and few resources dedicated up-front to data organization—the perfect recipe for the "tagging approach."

However, the "library approach," even as we have caricatured it above, has many advantages. In particular, annotations are generally consistent, of uniformly high quality, and complete (given enough resources). In the tagging approach, who knows whether two annotators will label the same object the same way? Or whether they will use useful annotations? Or whether an object will end up with the annotations needed to describe it? These questions are the subject of this paper: to what extent does the tagging approach match the consistency, quality, and completeness of the library approach? We believe these questions are a good proxy for the general question of whether the tagging approach organizes data well, a question which affects some of the most popular sites on the web.

While there has been work on the dynamics of tagging systems (e.g., [5], [4]), there has been little evaluation of the tagging approach itself. Our previous work [7] compared library controlled vocabulary terms to uncontrolled user tags. The present work considers numerous, broader questions like synonymy, paid tags, information integration, taxonomies ("classifications"), and perceived annotation quality.

This paper looks at social cataloging sites—sites where users tag books. By using books as our objects, we can compare user tags to decades of expert library cataloger metadata. Sometimes, we treat the library metadata as a gold standard, but we do not do so across the board. For example, we test if users prefer annotations produced by catalogers, user taggers, or paid taggers. (We believe we are the first to suggest and evaluate the use of paid taggers.) By using two social cataloging sites (LibraryThing and Goodreads), we can see how consistently users annotate objects across tagging systems. Overall, we give a comprehensive picture of the tradeoffs and techniques involved in using the tagging approach for organizing a collection, though we do focus by necessity on popular tags and topics.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
WSDM'10, February 4–6, 2010, New York City, New York, USA.
Copyright 2010 ACM 978-1-60558-889-6/10/02 ...$10.00.

Our investigation proceeds as follows. In Section 2 we build a vocabulary to discuss tagging and library data. In Section 3, we describe our datasets. In each of Sections 4, 5, and 6, we evaluate the tagging approach in terms of consistency, quality, and completeness. In Section 7 we discuss related work, and we conclude in Section 8.

2. GENERAL PRELIMINARIES

A social tagging system consists of users $u \in U$, annotations $a \in A$, and objects $o \in O$. In this paper we focus on social cataloging sites where the objects are books. More accurately, an object is a *work*, which represents one or more closely related books (e.g., the different editions of a book represent a work).

An object o can be annotated in three ways. First, an object o can be annotated (for free) by a user of the site, in which case we call the annotation a *tag* or (in some contexts) a user tag (written $t_i \in T$). For example, the top 10 most popular tags in our LibraryThing dataset are "non-fiction," "fiction," "history," "read," "unread," "own," "reference," "paperback," "biography," and "novel." Second, in a variety of experiments, we pay non-experts to produce "tags" for a given object. These are functionally the same as tags, but the non-experts may know little about the object they are tagging. As a result, we call these paid non-experts "paid taggers," and the annotations they create "$-tags", or $\$_i \in \$$ to differentiate them from unpaid user tags. Thirdly, works are annotated by librarians. For example, the Dewey Decimal Classification may say a work is in class 811, which as we will see below, is equivalent to saying the book has annotations "Language and Literature", "American and Canadian Literature," and "Poetry." We will call the annotations made by librarians "library terms" (written $l_i \in L$).

In a given system, an annotation a implicitly defines a group, i.e., the group of all objects that have annotation a (we define $O(a)$ to return this set of objects). We call a the *name* of such a group. A group also has a *size* equal to the number of objects it contains (we define $oc(a)$ to return this size). Since an object can have multiple annotations, it can belong to many groups. An object o becomes contained in group a when an annotator annotates o with a. We overload the notation for T, \$, and L such that $T(o_i)$, $\$(o_i)$, and $L(o_i)$ return the bag (multiset) of user tags, paid tags, and library annotations for work o_i, respectively.

2.1 Library Terms

We look at three types of library terms: *classifications*, *subject headings*, and the contents of *MARC 008*.[1]

A *classification* is a set of annotations arranged as a tree, where each annotation may contain one or more other annotations. An object is only allowed to have one *position* in a classification. This means that an object is associated with one most specific annotation in the tree and all of its ancestor annotations in the tree.

A *subject heading* is a library term chosen from a controlled list of annotations. A *controlled list* is a predetermined set of annotations. The annotator may not make up new subject headings. An object may have as many subject headings as desired by the annotator.

Works are annotated with two classifications, the Library of Congress Classification (LCC) and the Dewey Decimal Classification (DDC). A work has a position in both classifications. LCC and DDC *encode* their hierarchy information in a short string annotating a work, for example, GV735 or 811 respectively. The number 811 encodes that the book is about "Language and Literature" because it is in the 800s, "American and Canadian Literature" because it is in the 810s, and "Poetry" most specifically, because it is in the 811s. Likewise, "GV735" is about "Recreation and Leisure" because it is in GV, and "Umpires and Sports officiating" because it is in GV735. One needs a *mapping table* to *decode* the string into its constituent hierarchy information.

Works are also annotated with zero or more Library of Congress Subject Headings (LCSH).[2] LCSH annotations are structured as one LCSH main topic and zero or more LCSH subtopics selected from a vocabulary of phrases. For example, a book about the philosophy of religion might have the heading "Religion" (Main Topic) and "Philosophy" (Subtopic). In practice, books rarely have more than three LCSH headings for space, cost, and historical reasons. Commonly only the most specific LCSH headings are annotated to a book, even if more general headings apply.

We flatten LCC, DDC, and LCSH for this paper. For example in DDC, 811 is treated as three groups $\{800, 810, 811\}$. LCSH is somewhat more complex. For example, we treat "Religion" more specifically "Philosophy" as three groups {Main:Religion:Sub:Philosophy, Religion, Philosophy}. This is, in some sense, not fair to LCC, DDC, or LCSH because the structure in the annotations provides additional information. However, we also ignore significant strengths of tagging in this work, for example, its ability to have thousands of unique annotations for a single work, or its ability to show gradation of meaning (e.g., a work 500 people tag "fantasy" may be more classically "fantasy" than a work that only 10 people have tagged). In any case, the reader should note that our group model does not fully model the difference between structured and unstructured terms.

A *MARC record* is a standard library record that contains library terms for a particular book. It includes a fixed length string which we call *MARC 008* that states whether the book is a biography, whether the book is fiction, and other details. We define L_{LCC}, L_{DDC}, L_{LCSH}, L_{LM}, and $L_{MARC008}$ to be the set of library terms in LCC, DDC, LCSH, LCSH main topics, and MARC008, respectively.

3. DATASETS

We use a dump of Library of Congress MARC records from the Internet Archive as the source of our library terms. We chose to use only those $2,218,687$ records which had DDC and LCC library terms as well as an ISBN (a unique identifier for a book). We also use a list of approximately $6,000$ groups in LCC from the Internet Archive, and a list of approximately $2,000$ groups in DDC from a library school board in Canada as mapping tables for LCC and DDC.

We started crawling LibraryThing in early April 2008, and began crawling Goodreads in mid June 2008. In both cases,

[1]This section gives a brief overview of library terms and library science for this paper. However, it is necessarily United States-centric, and should not be considered the only way to organize data in a library! For more information, see a general reference such as one by Mann ([10], [9]).

[2]Strictly speaking, we sometimes use any subject heading in MARC 650, but almost all of these are LCSH in our dataset.

our dataset ends in mid-October 2008. We crawled a sample of works from each site based on a random selection of ISBNs from our Library of Congress dataset. LibraryThing focuses on cataloging books (and has attracted a number of librarians in addition to regular users), whereas Goodreads focuses on social networking (which means it has sparser tagging data). We gathered synonym sets (see Section 4.1) from LibraryThing on October 19th and 20th.

We use two versions of the LibraryThing dataset, one with all of the works which were found from our crawl, and one with only those works with at least 100 unique tags. The former dataset, which we call the "full" dataset, has 309,071 works. The latter dataset, which we call the "min100" dataset, has 23,396 works. We use only one version of our Goodreads dataset, a version where every work must have at least 25 tags and there are 7,233 unique ISBNs.

4. EXPERIMENTS: CONSISTENCY

In this and the next two sections, we conduct experiments to determine if tagging systems are consistent, high quality, and complete. Each experiment has a description of a feature of the library approach to be emulated, a summary of the results, zero or more preliminaries sections, and details about background, methodology, and outcome.

The experiments in this section look at *consistency*:

Section 4.1 How big a problem is synonymy? That is, how consistent are users of the same tagging system in choosing the same tag for the same topic?

Section 4.2 How consistent is the tag vocabulary chosen, or used, by users across different tagging systems? That is, do users use the same tags across tagging systems?

Section 4.3 How consistently is a particular tag applied across different tagging systems? That is, do users use the same tags to describe the same objects?

Section 4.4 If paid taggers are asked to annotate objects with $-tags, are those $-tags consistent with user tags?

4.1 Synonymy

Summary

Library Feature: There should not be multiple places to look for a particular object. This means that we would prefer tags not to have synonyms. When a tag does have synonyms, we would prefer one of the tags to have many more objects annotated with it than the others.

Result: Most tags have few or no synonyms appearing in the collection. In a given synonym set, one tag is usually much more common.

Conclusion: Synonymy is not a major problem for tags.

Preliminaries: Synonymy

A group of users named *combiners* mark tags as equivalent. We call two tags that are equivalent according to a combiner *synonyms*. A set of synonymous tags is called a *synonym set*. Combiners are regular users of LibraryThing who do not work directly for us. While we assume their work to be correct and complete in our analysis, they do have two notable biases: they are strict in what they consider a synonym (e.g., "humour" as British comedy is not a synonym of "humor" as American comedy) and they may focus more on finding synonyms of popular, mature tags.

We write the synonym set of t_i, including itself, as $S(t_i)$. We calculate the entropy $H(t_i)$ (based on the probability $p(t_j)$ of each tag) of a synonym set $S(t_i)$ as:

$$p(t_j) = \frac{oc(t_j)}{\sum_{t_k \in S(t_j)} oc(t_k)} \quad H(t_i) = -\sum_{t_j \in S(t_i)} p(t_j) \log_2 p(t_j)$$

$H(t_i)$ measures the entropy of the probability distribution that we get when we assume that an annotator will choose a tag at random from a synonym set with probability in proportion to its object count. For example, if there are two equally likely tags in a synonym set, $H(t_i) = 1$. If there are four equally likely tags, $H(t_i) = 2$. The higher the entropy, the more uncertainty that an annotator will have in choosing which tag to annotate from a synonym set, and the more uncertainty a user will have in determining which tag to use to find the right objects. We believe low entropy is generally better than high entropy, though it may be desirable under some circumstances (like query expansion) to have high entropy synonym sets.

Details

Due to the lack of a controlled vocabulary, tags will inevitably have synonymous forms. The best we can hope for is that users ultimately "agree" on a single form, by choosing one form over the others much more often. For example, we hope that if the tag "fiction" annotates 500 works about fiction, that perhaps 1 or 2 books might be tagged "fictionbook" or another uncommon synonym. For this experiment, we use the top 2000 LibraryThing tags and their synonyms.

Most tags have no synonyms, though a minority have as many as tens of synonyms (Figure 1(a)). The largest synonym set is 70 tags (synonyms of "19th century"). Unlike one might expect, $|S(t_i)|$ is not strongly correlated with $oc(t_i)$ as shown in Figure 1(b). (Kendall's $\tau \approx 0.208$.)

Figure 1(c) is a histogram of the entropies of the top 2000 tags, minus those synonym sets with an entropy of zero. In 85 percent of cases, $H(t_i) = 0$. The highest entropy synonym set, at $H(t_i) = 1.56$ is the synonym set for the tag "1001bymrbfd," or "1001 books you must read before you die." Less than fifteen tags (out of 2000) have an entropy above 0.5. The extremely low entropies of most synonym sets suggests that most tags have a relatively definitive form.

4.2 Cross-System Annotation Use

Summary

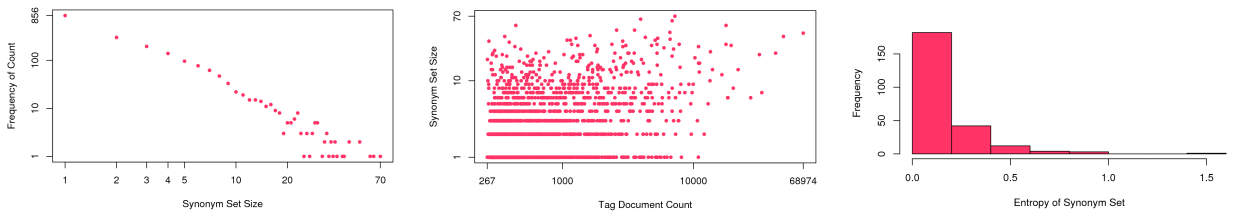
Library Feature: Across tagging systems, we would like to see the systems use the same vocabulary of tags because they are annotating the same type of objects—works.

Result: The top 500 tags of LibraryThing and Goodreads have an intersection of almost 50 percent.

Conclusion: Similar systems have similar tags, though tagging system owners should encourage short tags.

Preliminaries: Information Integration

Federation is when multiple sites share data in a distributed fashion allowing them to combine their collections. *Information integration* is the process of combining, de-duplicating, and resolving inconsistencies in the shared data. Two useful features for information integration are consistent cross-system annotation use and consistent cross-system object annotation. We say two systems have *consistent annotation use* if the same annotations are used overall in both



(a) Frequencies of Synonym Sets (b) Tag Frequency vs Syn. Set Size (c) $H(t_i)$ of Top 2000 Tags, $H(t_i) \neq 0$

Figure 1: Synonym set frequencies, sizes, and entropies.

systems (this section). We say two systems have *consistent object annotation* if the same object in both systems is annotated similarly (Section 4.3). Libraries achieve these two features through "authority control" (the process of creating controlled lists of headings) and professional catalogers.

Details

For both LibraryThing and Goodreads, we look at the top 500 tags by object count. Ideally, a substantial portion of these tags would be the same, suggesting similar tagging practices. Differences in the works and users in the two systems will lead to some differences in tag distribution. Nonetheless, both are mostly made up of general interest books and similar demographics.

The overlap between the two sets is 189 tags, or about 38 percent of each top 500 list.[3] We can also match by determining if a tag in one list is in the synonym set of a tag in the other list. This process leads to higher overlap—231 tags, or about 46 percent. The higher overlap suggests "combiners" are more helpful for integrating two systems than for improving navigation within their own system. An overlap of nearly 50 percent of top tags seems quite high to us, given that tags come from an unlimited vocabulary, and books can come from the entire universe of human knowledge.

Much of the failed overlap can be accounted for by noting Goodreads' prevalence of multi-word tags. Multi-word tags lead to less overlap with other users, and less overlap across systems. We compute the number of words in a tag by splitting on spaces, underscores, and hyphens. On average, tags in the intersection of the two systems have about 1.4 words. However, tags not in the intersection have an average of 1.6 words in LibraryThing, and 2.3 words in Goodreads. This implies that for tagging to be federated across systems users should be encouraged to use fewer words.

While there are 231 tags in the overlap between the systems (with synonyms), it is also important to know if these tags are in approximately the same ranking. Is "fantasy" used substantially more than "humor" in one system? We computed a Kendall's τ rank correlation between the two rankings from LibraryThing and Goodreads of the 231 tags in the overlap of $\tau \approx 0.44$. This means that if we choose any random pair of tags in both rankings, it is a little over twice as likely that the pair of tags is in the same order in both rankings as it is that the pair will be in a different order.

[3]Note that comparing sets at the same 500 tag cutoff may unfairly penalize border tags (e.g., "vampires" might be tag 499 in LT but tag 501 in GR). We use the simpler measurement above, but we also conducted an analysis comparing, e.g., the top 500 in one system to the top 1000 in the other system. Doing so increases the overlap by ≈ 40 tags.

4.3 Cross-System Object Annotation

Summary

Library Feature: We would like annotators to be consistent, in particular, the same work in two different tagging systems should be annotated with the same, or a similar distribution, of tags. In other words, does "Winnie-the-Pooh" have the same set of tags in LibraryThing and Goodreads?
Result: Duplicate objects across systems have low Jaccard similarity in annotated tags, but high cosine similarity.
Conclusion: Annotation practices are similar across systems for the most popular tags of an object, but often less so for less common tags for that object.

Details

We limited our analysis to works in both LibraryThing and Goodreads, where Goodreads has at least 25 tags for each book. This results in 787 works. Ideally, for each work, the tags would be almost the same, implying that given the same source object, users of different systems will tag similarly.

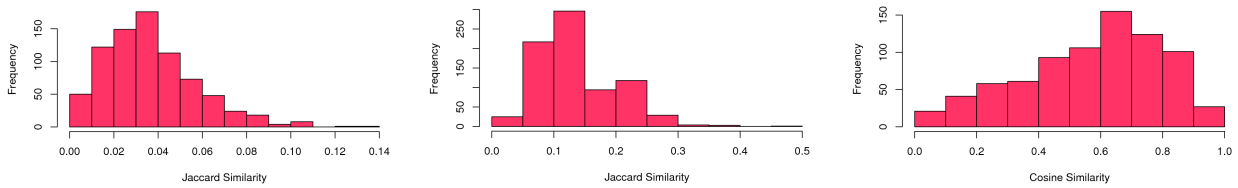
Figure 2 shows distributions of similarities of tag annotations for the same works across the systems. We use Jaccard similarity for set similarity (i.e., each annotation counts as zero or one), and cosine similarity for similarity with bags (i.e., counts). Because the distributions are peaked, Jaccard similarity measures how many annotations are shared, while cosine similarity measures overlap of the main annotations.

Figure 2(a) shows that the Jaccard similarity of the tag sets for a work in the two systems is quite low. For example, about 150 of the 787 works have a Jaccard similarity of the two tag sets between 0.02 and 0.03. One might expect that the issue is that LibraryThing has disproportionately many more tags than Goodreads, and these tags increase the size of the union substantially. To control for this, in Figure 2(b), we take the Jaccard similarity of the top 20 tags for each work. Nonetheless, this does not hugely increase the Jaccard value in most cases. Figure 2(c) shows the distribution of cosine similarity values. (We treat tags as a bag of words and ignore three special system tags.) Strikingly, the cosine similarity for the same work is actually quite high. This suggests that for the same work, the most popular tags are likely to be quite popular in both systems, but that overall relatively few tags for a given work will overlap.

4.4 $-tag Annotation Overlap

Summary

Library Feature: We would like paid taggers to be able to annotate objects in a way that is consistent with users.



(a) Jaccard Similarity of All Tags (b) Jaccard Similarity of Top 20 Tags (c) Cosine Similiarity of All Tags

Figure 2: Distribution of same book similarities.

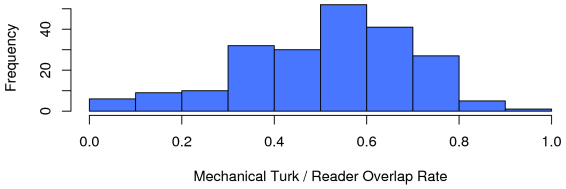


Figure 3: Overlap Rate Distribution.

This reduces dependence on users, and means that unpopular objects can be annotated for a fee.
Result: $-tags produced by paid taggers overlap with user tags on average 52 percent of the time.
Conclusion: Tagging systems can use paid taggers.

Preliminaries: Mechanical Turk

Amazon's *Mechanical Turk* is a marketplace made up of requesters and workers. The *requesters* provide a task and set a price. The *workers* accept or decline the task. A *task* is a unit of work, like determining the type of a tag.

Preliminaries: $-tag Tagging Setup

This section asks whether $-tag terms paid taggers annotate objects with are the same as terms annotated by users as user tags. We randomly selected works from the "min100" dataset with at least three unique $l_i \in L_{LM}$. We then showed paid taggers (in our case, Mechanical Turk workers) a search for the work (by ISBN) on Google Book Search and Google Product Search, two searches which generally provide a synopsis and reviews, but do not generally provide library metadata like subject headings. The paid taggers were asked to add three $-tags which described the given work. Each work was labeled by at least three paid taggers, but different paid taggers could annotate more or fewer books (this is standard on the Mechanical Turk). We provided $2,000$ works to be tagged with 3 $-tags each. Some paid taggers provided an incomplete set of $-tags, leading to a total of $16,577$ $-tags. Paid taggers spent ≈ 90 seconds per work, and we usually spent less than $0.01 per $-tag/work pair. (We analyze $-tags in Sections 4.4, 5.2, and 5.3.)

4.4.1 Details

$-tags matched with tags t_i already annotated to the work at least once on average 52% of the time (standard deviation of 0.21). Thus, paid taggers who had in the vast majority of cases not read the book, overlapped with real book readers more than half the time in what $-tags they applied. A natural followup question is whether some workers are much better at paid tagging than others. We found a range of "overlap rates" among paid taggers (shown in Figure 3), but we are unsure whether higher performance could be predicted in advance.

5. EXPERIMENTS: QUALITY

The experiments in this section look at *quality*:
Section 5.1 Are the bulk of tags of high quality types? For example, are subjective tags like "stupid" common?
Section 5.2 Are $-tags high quality in comparison to library annotations and user tags?
Section 5.3 Can we characterize high quality user tags?

5.1 Objective, Content-based Groups

Summary

Library Feature: Works should be organized objectively based on their content. For example, we would prefer a system with groups of works like "History" and "Biography," to one with groups of works like "sucks" and "my stuff."
Result: Most tags in both of our social cataloging sites were objective and content-based. Not only are most very popular tags ($oc(t_i) > 300$) objective and content-based, but so are less popular and rare tags.
Conclusion: Most tags, rather than merely tags that become very popular, are objective and content-based, even if they are only used a few times by one user.

Preliminaries: Tag Types

We divide tags into six types:
Objective and Content-based *Objective* means not depending on a particular annotator for reference. For example, "bad books" is not an objective tag (because one needs to know who thought it was bad), whereas "world war II books" is an objective tag. *Content-based* means relating to the book contents (e.g., the story, facts, genre). For example, "books at my house" is not a content-based tag, whereas "bears" is.
Opinion The tag implies a personal opinion. For example, "sucks" or "excellent."
Personal The tag relates to personal or community activity or use. For example, "my book", "wishlist", "mike's reading list", or "class reading list".
Physical The tag describes the book physically. For example, "in bedroom" or "paperback".
Acronym The tag is an acronym that might mean multiple things. For example, "sf" or "tbr".
Junk The tag is meaningless or indecipherable. For example, "b" or "jiowefijowef".

Details

If a tagging system is primarily made up of objective, content-based tags, then it is easier for users to find objects. In a

	LT%	GR%
Objective, Content of Book	60.55	57.10
Personal or Related to Owner	6.15	22.30
Acronym	3.75	1.80
Unintelligible or Junk	3.65	1.00
Physical (e.g., "Hardcover")	3.55	1.00
Opinion (e.g., "Excellent")	1.80	2.30
None of the Above	0.20	0.20
No Annotator Majority	20.35	14.30
Total	100	100

Table 1: Tag types for top 2000 LibraryThing and top 1000 GoodReads tags as percentages.

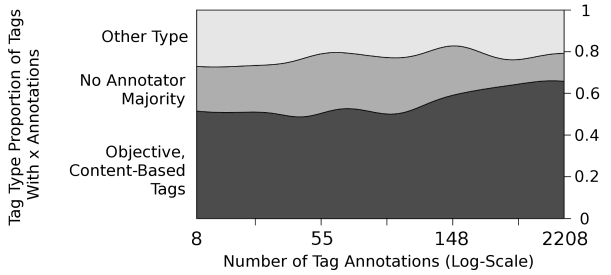


Figure 4: Conditional density plot showing probability of (1) annotators agreeing a tag is objective, content-based, (2) annotators agreeing on another tag type, or (3) no majority of annotators agreeing.

library system, all annotations are objective and content-based in that they do not depend on reference to the annotator, and they refer to the contents of the book.

To produce an unbiased view of the types of tags in our sites, we used Mechanical Turk. We submitted the top $2,000$ LibraryThing tags and top $1,000$ Goodreads tags by annotation count to be evaluated. We also sampled $1,140$ LibraryThing tags, 20 per rounded value of $log(oc(t_i))$, from 2.1 to 7.7. We say a worker provides a *determination* of the answer to a task (for example, the tag "favorite" is an opinion). Overall, 126 workers examined $4,140$ tags, five workers to a tag, leading to a total of $20,700$ determinations. We say the *inter-annotator agreement* is the pair-wise fraction of times two workers provide the same answer. The inter-annotator agreement rate was about 65 percent.

Table 1 shows the proportion of top tags by type for LibraryThing and Goodreads. For example, for 60.55% of the top 2000 LibraryThing tags (e.g., $\frac{1211}{2000}$), at least three of five workers agreed that the tag was objective and content-based. The results show that regardless of the site, a majority of tags tend to be objective, content-based tags. In both sites, about 60 percent of the tags examined were objective and content-based. Interestingly, Goodreads has a substantially higher number of "personal" tags than LibraryThing. We suspect that this is because Goodreads calls tags "bookshelves" in their system.

Even if we look at tags ranging from $oc(t_i) = 8$ to $oc(t_i) = 2208$, as shown in Figure 4, the proportion of objective, content-based tags remains very high. That figure shows the probability that a tag will be objective and content-based conditioned on knowing its object count. For example, a tag annotating 55 objects has about a 50 percent chance of being objective and content-based.

H-Scores (by Evaluator)	μ	SD
User Tags	4.46	0.75
LCSH Main Topics	5.18	0.76
$-tags	5.22	0.83

Table 2: Basic statistics for the mean h-score assigned by evaluators to each annotation type. Mean (μ) and standard deviation (SD) are abbreviated.

5.2 Quality Paid Annotations

Summary

Library Feature: We would like to purchase annotations of equal or greater quality to those provided by users.
Result: Judges like $-tags as much as subject headings.
Conclusion: Paid taggers can annotate old objects where users do a poor job of providing coverage and new objects which do not yet have tags. Paid taggers can quickly and inexpensively tag huge numbers of objects.

Preliminaries: $-tag Judging Setup

In this section and the next, we evaluate the relative perceived helpfulness of annotations $t_i \in T$, $\$_i \in \$$ and $l_i \in L_{LM}$. We randomly selected 60 works with at least three tags $t_i \in T$ and three LCSH terms $l_i \in L_{LM}$ from our "min100" dataset.

We created tasks on the Mechanical Turk, each of which consisted of 20 subtasks (a "work set"), one for each of 20 works. Each subtask consisted of a *synopsis* of the work o_i and an *annotation evaluation* section. A synopsis consisted of searches over Google Books and Google Products as in Section 4.4. The annotation evaluation section showed nine annotations in random order, three each from $T(o_i)$, $\$(o_i)$, and $L_{LM}(o_i)$, and asked how helpful the given annotation would be for finding works similar to the given work o_i on a scale of 1 ("not at all helpful") to 7 ("extremely helpful").

We removed three outlier evaluators who either skipped excessive numbers of evaluations, or awarded excessive numbers of the highest score. Remaining missing values were replaced by group means. That is, a missing value for a work/annotation/evaluator triplet was replaced by the mean of helpfulness scores from among all evaluators who had provided scores for that triplet. We abbreviate "helpfulness score" as *h-score* in the following. We say that annotations $t_i \in T$, $\$_i \in \$$, and $l_i \in L_{LM}$ differ in their *annotation type*.

Details

In order to understand the perceived quality of $-tags, we wondered if, given the works that each evaluator saw, they tended to prefer $-tags, user tags, or LCSH on average. To answer this question, we produced a mean of means for each annotation type (i.e., $-tags, user tags, and LCSH main topics) to compare to the other annotation types. We do so by averaging the annotations of a given type within a given evaluator (i.e., to determine what that evaluator thought) and then by averaging the averages produced by each evaluator across all evaluators.

H-Scores	μ	SD	μ 95% CI
$-tags	4.93	1.92	(4.69, 5.17)
Rare User Tags	4.23	2.11	(3.97, 4.50)
Moderate User Tags	5.80	1.47	(5.63, 5.98)
Common User Tags	5.27	1.72	(5.05, 5.48)
LCSH Main Topics	5.13	1.83	(4.91, 5.36)

Table 3: Basic statistics for the mean h-score assigned to a particular annotation type with user tags split by frequency. Mean (μ) and standard deviation (SD) are abbreviated.

Table 2 summarizes the basic statistics by annotation type. For example, the mean evaluator assigned a mean score of 4.46 to user tags, 5.18 to LCSH main topics, and 5.22 to $-tags. At least for our 60 works, $-tags are perceived as being about as helpful as LCSH library annotations, and both are perceived as better than user tags (by about 0.6 h-score). A repeated measures ANOVA showed annotation type differences in general to be significant, and all differences between mean h-scores by annotation type were significant ($p < 0.001$) with the exception of the difference between $-tags and LCSH main topics.

5.3 Finding Quality User Tags

Summary

Library Feature: We would like tag annotations to be viewed as competitive in terms of perceived helpfulness with annotations provided by expert taxonomists.
Result: Moderately common user tags are perceived as more helpful than both LCSH and $-tags.
Conclusion: Tags may be competitive with manually entered metadata created by paid taggers and experts, especially when information like frequency is taken into account.

Details

Section 5.2 would seem to suggest that tags $t_i \in T$ are actually the worst possible annotation type because the average evaluator gave $-tags and LCSH main topics a mean h-score 0.6 higher than user tags. Nonetheless, in practice we found that tags $t_i \in T(o_i)$ often had higher h-scores for the same object o_i than corresponding annotations $\$_i \in \$(o_i)$ and $l_i \in L_{LM}(o_i)$. It turns out that this discrepancy can be explained in large part by the popularity of a user tag.

We define $pop(o_i, t_m)$ to be the percentage of the time that tag t_m is assigned to object o_i. For example, if an object o_i has been annotated ("food", "food", "cuisine", "pizza") then we would say that $pop(o_i, t_{food}) = \frac{2}{4}$. We partitioned the h-scores for T into three sets based on the value $pop(o_i, t_m)$ of the annotation. Those sets were user tag annotations with $pop(o_i, t_m) < 0.11$ ("rare"), those with $0.11 \leq pop(o_i, t_m) < 0.17$ ("moderate"), and those with $0.17 \leq pop(o_i, t_m)$ ("common").[4]

Table 3 shows the basic statistics with these more fine grained categories on a per evaluation basis (i.e., not averaging per annotator). For example, the 95% confidence interval for the mean h-score of moderate popularity user tags is (5.63, 5.98), and the mean h-score of $-tags is 4.93 in our sample. The ANOVA result, Welch-corrected to adjust for unequal variances within the five annotation types, is ($WelchF(4, 629.6) = 26.2$; $p < .001$). All differences among these finer grained categories are significant, with the exception of common user tags versus LCSH, common user tags versus $-tags, and LCSH main topics versus $-tags.

Using the finer grained categories in Table 3 we can now see that moderately common user tags are perceived as better than all other annotation types. (Furthermore, rare user tags were dragging down the average in the analysis of Section 5.2.) We speculate that rare user tags are too personal and common user tags too general. Despite some caveats (evaluators do not read the work, value of annotations changes over time, works limited by Librarything availability), we are struck by the fact that evaluators perceive moderately common user tags to be more helpful than professional, expert-assigned library annotations.

6. EXPERIMENTS: COMPLETENESS

The experiments in this section look at *completeness*:
Section 6.1 Do user tag annotations cover many of the same topics as professional library annotations?
Section 6.2 Do user tags and library annotations corresponding to the same topic annotate the same objects?

6.1 Coverage

Summary

Library Feature: We believe that after decades of consensus, libraries have roughly the right groups of works. A system which attempts to organize works should end up with groups similar to or a superset of library terms.
Result: Many top tags have equivalent (see below) library terms. Tags contain more than half of the tens level DDC headings. There is a corresponding LCSH heading for more than 65 percent of top objective, content-based tags.
Conclusion: Top tags often correspond to library terms.

Preliminaries: Containment and Equivalence

Our goal is to compare the groups formed by user tags and those formed by library annotations. For instance, is the group defined by tag "History of Europe" equivalent to the group formed by the library term "European History?" We can take two approaches to defining equivalence. First, we could say that group g_1 is equivalent to group g_2 if they both contain the same objects (in a given tagging system). By this definition, the group "Art" could be equivalent to the group "cool" if users had tagged all works annotated with the library term "Art" with the tag "cool." Note that this definition is system specific.

A second approach is to say that group g_1 is equivalent to group g_2 if the names g_1 and g_2 "semantically mean the same." Under this definition, "cool" and "Art" are not equivalent, but "European History" and "History of Europe" are. The latter equivalence holds even if there are some books that have one annotation but not the other. For this definition of equivalence we assume there is a semantic test $m(a, b)$ that tells us if names a and b "semantically mean the same." (We implement m by asking humans to decide.)

In this paper we use the second definition of equivalence (written $g_1 = g_2$). We do this because we want to know

[4]H-scores were sampled for the "common" set for analysis due to large frequency differences between rare user tags and more common tags. Values of $pop(o_i, t_j)$ varied between less than 1 percent and 28 percent in our evaluated works.

(a) Sampled Containment Relationships (*con-pairs*)

Tag	Contained Library Term
spanish	romance → spanish (lc pc 4001.0-4978.0)
pastoral	pastoral theology (lc bv 4000.0-4471.0)
civil war	united states → civil war period, 1861-1865 → civil war, 1861-1865 → armies. troops (lc e 491.0-587.0)
therapy	psychotherapy (lcsh)
chemistry	chemistry → organic chemistry (lc qd 241.0-442.0)

(b) Sampled Equivalence Relationships (*eq-pairs*)

Tag	Equivalent Library Term
mammals	zoology → chordates. vertebrates → mammals (lc ql 700.0-740.8)
fitness	physical fitness (lcsh)
catholic church	catholic church (lcsh)
golf	golf (lcsh)
astronomy	astronomy (lc qb 1.0-992.0)

Table 4: Randomly sampled containment and equivalence relationships for illustration.

to what extent library terms exist which are semantically equivalent to tags (Section 6.1) and to what extent semantically equivalent groups contain similar objects (Section 6.2).

When we compare groups, not only are we interested in equivalence, but also in "containment." We again use semantic definitions: We say a group g_1 *contains* group g_2 (written $g_1 \supseteq g_2$) if a human that annotates an object o with g_2 would agree that o could also be annotated with g_1. Note that even though we have defined equivalence and containment of groups, we can also say that two annotations are equivalent or contain one another if the groups they name are equivalent or contain one another.

Preliminaries: Gold Standard (t_j, l_i) *Relationships*

In this section and the next, we look at the extent to which tags t_j and library terms l_i satisfy similar information needs. We assume a model where users find objects using single annotation queries. If $t_j = l_i$ for a given (t_j, l_i), we say (t_j, l_i) is an *eq-pair*. If $t_j \supseteq l_i$ for a given (t_j, l_i), we say (t_j, l_i) is a *con-pair*. In this section, we look for and describe eq-pairs (where both annotations define the same information need) and con-pairs (where a library term defines a subset of an information need defined by a tag). In Section 6.2, we use these pairs to evaluate the recall of single tag queries—does a query for tag t_j return a high proportion of objects labeled with library terms equivalent or contained by t_j? For both sections, we need a set of gold standard eq- and con-pairs.

Ideally, we would identify all eq- and con-pairs $(t_j, l_i) \in T \times L$. However, this is prohibitively expensive. Instead, we create our gold standard eq- and con-pairs as follows:

Step 1 We limit the set of tags under consideration. Specifically, we only look at tags in T_{738}: the 738 tags from the top 2,000 which were unanimously considered objective and content-based in Section 5.1. (These 738 tags are present in about 35% of tag annotations.)

Step 2 We identify (t_j, l_i) pairs that are likely to be eq- or con-pairs based on how t_j and l_i are used in our dataset. First, we drop all (t_j, l_i) pairs that do not occur together on at least 15 works. Second, we look for (t_j, l_i) pairs with high values of $q(t_j, l_i) = (P(t_j, l_i) - P(t_j)P(l_i)) \times |O(l_i)|$. $q(t_j, l_i)$ is inspired by *leverage* $(P(t_j, l_i) - P(t_j)P(l_i))$ from the association rule mining community [13], though with bias $(|O(l_i)|)$ towards common relationships. We drop all (t_j, l_i) pairs that do not have $q(t_j, l_i)$ in the top ten for a given tag t_j.

Step 3 We (the researchers) manually examine pairs output from Step 2 and judge if they are indeed eq- or con-pairs. At the end of this step, our gold standard eq- and con-pairs have been determined.

Step 4 We evaluate our gold standard using Mechanical Turk workers. We do not change any eq- or con-pair designations based on worker input, but this step gives us an indication of the quality of our gold standard.

The filtering procedures in Steps 1 and 2 allowed us to limit our manual evaluation to 5,090 pairs in Step 3. (Though, the filtering procedures mean we are necessarily providing a lower bound on the eq- and con-pairs present in the data.) In Step 3, we found 2,924 con-pairs and 524 eq-pairs. (Table 4 shows random samples of relationships produced.)

To evaluate our gold standard in Step 4, we provided Mechanical Turk workers with a random sample of eq- and con-pairs from Step 3 in two scenarios. In a true/false validation scenario, the majority of 20 workers agreed with our $t_j = l_i$ and $t_j \supseteq l_i$ judgments in $\frac{64}{65} = 98\%$ of cases. However, they said that $t_j = l_i$ when $t_j \neq l_i$ or $t_j \supseteq l_i$ when $t_j \not\supseteq l_i$ in $\frac{34}{90} = 38\%$ of cases, making our gold standard somewhat conservative. A χ^2 analysis of the relationship between the four testing conditions (true con-pair, false con-pair, true eq-pair, and false eq-pair) shows a strong correlation between containment/equivalence examples and true/false participant judgments ($\chi^2(3) = 45.3, p < .001$). In a comparison scenario where workers chose which of two pairs they preferred to be an eq- or con-pair, the majority of 30 workers agreed with our judgments in $\frac{138}{150} = 92\%$ of cases.

Details

In this analysis, we ask if tags correspond to library annotations. We ask this question in two directions: how many top tags have equivalent or contained library annotations, and how many of the library annotations are contained or equivalent to top tags? Assuming library annotations represent good topics, the first direction asks if top tags represent good topics, while the second direction asks what portion of those good topics are represented by top tags.

In this section and the next, we use an imaginary "System I" to illustrate coverage and recall. System I has top tags $\{t_1, t_2, t_3, t_4\}$, library terms $\{l_1, l_2, l_3, l_4, l_5, l_6\}$, eq-pairs $\{t_1 = l_1, t_2 = l_2\}$, and con-pairs $\{t_3 \supseteq l_3, t_1 \supseteq l_5\}$. Further, $l_3 \supseteq l_4$ based on hierarchy or other information (perhaps l_3 might be "History" and l_4 might be "European History").

Looking at how well tags represent library terms in System I, we see that 2 of the 4 unique tags appear in eq-pairs, so $\frac{2}{4}$ of the tags have equivalent library terms. Going in the opposite direction, we see that 2 out of 6 library terms have equivalent tags, so what we call eq-coverage below is $\frac{2}{6}$. We also see that 2 of the library terms (l_3, l_5) are directly contained by tags, and in addition another term (l_4) is con-

	X00	XX0	XXX
Con-Coverage	0.3	0.65	0.677
Eq-Coverage	0.1	0.28	0.021

Table 5: Dewey Decimal Classification coverage by tags.

tained by l_3. Thus, a total of 3 library terms are contained by tags. We call this $\frac{3}{6}$ fraction the con-coverage.

We now report these statistics for our real data. Of 738 tags in our data set, 373 appear in eq-pairs. This means at least half ($\frac{373}{738}$) of the tags have equivalent library terms.[5]

To go in the opposite direction, we compute coverage by level in the library term hierarchy, to gain additional insights. In particular, we use DDC terms which have an associated value between 0 and 1000. As discussed in Section 2.1, if the value is of the form X00, then the term is high level (e.g., 800 is Language and Literature); if the value is of the form XX0 it is lower level, and so on (e.g., 810 is American and Canadian Literature). We thus group the library terms into three sets, L_{X00}, L_{XX0} and L_{XXX}. (Set L_{X00} contains all terms with numbers of the form X00). For $L_{rs} \in \{L_{X00}, L_{XX0}, L_{XXX}\}$ being one of these groups, we define two metrics for coverage:

$$concoverage(L_{rs}) = \frac{\sum_{l_i \in L_{rs}} 1\{\exists t_j \in T \quad s.t. \quad t_j \supseteq l_i\}}{|L_{rs}|}$$

$$eqcoverage(L_{rs}) = \frac{\sum_{l_i \in L_{rs}} 1\{\exists t_j \in T \quad s.t. \quad t_j = l_i\}}{|L_{rs}|}$$

Table 5 shows these metrics for our data. For example, the first row, second column says that $\frac{65}{100}$ of XX0 DCC terms are contained by a tag. (More specifically, 65 percent of XXO terms have this property: the term l_i is in L_{XX0} and there is a con-pair (t_j, l_i) in our gold standard, or there is a (t_j, l_k) con-pair where $l_k \in L_{X00}$ and $l_k \supset l_i$.) The second row, third column says that $\frac{21}{1000}$ DDC ones level terms $l_i \in L_{XXX}$ have an eq-pair (t_j, l_i). About one quarter of XXO DDC terms have equivalent T_{738} tags.

6.2 Recall

Summary

Library Feature: A system should not only have the right groups of works, but it should have enough works annotated in order to be useful. For example, a system with exactly the same groups as libraries, but with only one work per group (rather than, say, thousands) would not be very useful.
Result: Recall is low (10 to 40 percent) using the full dataset. Recall is high (60 to 100 percent) when we focus on popular objects (min100).
Conclusion: Tagging systems provide excellent recall for popular objects, but not necessarily for unpopular objects.

Preliminaries: Recall

Returning to our System I example, say that l_1 annotates $\{o_1, o_3\}$, and l_5 annotates $\{o_4, o_5\}$. Because t_1 is equivalent to l_1, and contains l_5, we expect that any work labeled with either l_1 or l_5 could and should be labeled with t_1. We call o_1, o_3, o_4, o_5 the potential objects for tag t_1. Our goal is to see how closely the potential object set actually matches the set of objects tagged with t_1. For instance, suppose that t_1 actually annotates $\{o_1, o_2\}$. Since t_1 annotates one of the four potential works, we say that $recall(t_1) = \frac{1}{4}$.

More formally, if $l_i = t_j$, then we say $l_i \in E(t_j)$. If $t_j \supseteq l_i$, then we say $l_i \in C(t_j)$. Any object annotated with terms from either $E(t_j)$ or $C(t_j)$ should also have a tag t_j. Hence, the *potential object set* for a tag based on its contained or equivalent library terms is:

$$P_{t_j} = \bigcup_{l_i \in (E(t_j) \cup C(t_j))} O(l_i)$$

We define *recall* to be the recall of a single tag query on relevant objects according to our gold standard library data:

$$recall(t_j) = \frac{|O(t_j) \cap P_{t_j}|}{|P_{t_j}|}$$

and that the *Jaccard* similarity between the potential object set and the objects contained by a tag is:

$$J(O(t_j), P_{t_j}) = \frac{|O(t_j) \cap P_{t_j}|}{|O(t_j) \cup P_{t_j}|}$$

Details

In this experiment, we ask whether the tags provided by users have good recall of their contained library terms. An ideal system should have both good coverage (see Section 6.1) and high recall of library terms.

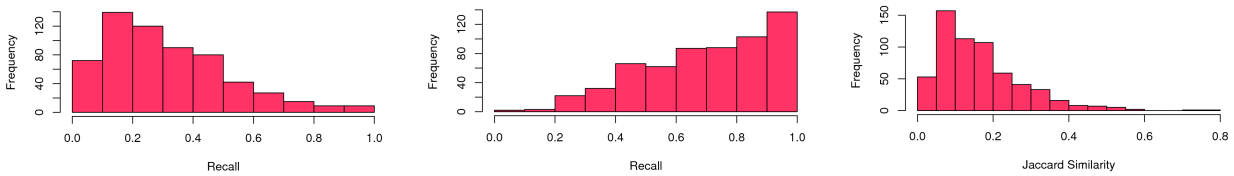
We look at recall for the tags $T_{603} \subset T_{738}$ that have at least one con-pair. Figure 5 shows the distribution of recall of tags $t_j \in T_{603}$ using the full and min100 datasets. Figure 5(a) shows that using the full dataset, most tags have 10 to 40 percent recall. For example, about 140 tags have recall between 10 and 20 percent. Figure 5(b) shows recall using the "min100" dataset. We can see that when we have sufficient interest in an object (i.e., many tags), we are very likely to have the appropriate tags annotated. Recall is often 80 percent and up. Lastly, Figure 5(c) shows the distribution of Jaccard similarity between $O(t_j)$ and P_{t_j}. For most tags, the set of tag annotated works is actually quite different from the set of library term annotated works, with the overlap often being 20 percent of the total works in the union or less. The objects in $O(t_j) - P_{t_j}$ are not necessarily incorrectly annotated with t_j. Since we know that many tags are of high quality (Section 5), a more likely explanation is that the library experts missed some valid annotations.

7. RELATED WORK

Our synonymy experiment in Section 4.1 is similar to previous work on synonymy and entropy in tagging systems. Clements et al. [2] use LibraryThing synonym sets to try to predict synonyms. By contrast, our goal was to determine if synonyms were a problem, rather than to predict them. Chi et al. [1] used entropy to study the evolution of the navigability of tagging systems. They look at entropy as a global tool, whereas we use it as a local tool within synonym sets.

Our experiments relating to information integration in Sections 4.2 and 4.3 (primarily Section 4.2, however), share some similarities to Oldenburg et al. [12] which looked at how to integrate tags across tagging systems, though that work is fairly preliminary (and focused on the Jaccard measure). That work also focuses on different sorts of tagging

[5]Note that this is a lower bound—based on techniques from [7], we suspect more than $\frac{503}{738}$ tags have equivalents.



(a) Recall on Full Dataset (b) Recall on "min100" Dataset (c) Jaccard on Full Dataset

Figure 5: Recall and Jaccard for 603 tags (Full and "min100" datasets).

systems, specifically, social bookmarking and research paper tagging systems, rather than social cataloging systems.

Our tag type experiment in Section 5.1 is related to work like Golder and Huberman [4] and Marlow et al. [11] which looked at common tag types in a tagging systems. However, we believe our work is possibly the first to analyze how tag types change over the long tail of tag usage (i.e., are less popular tags used differently from more popular tags?).

Like Section 5.3, other work has found moderately common terms in a collection to be useful. For instance, Haveliwala et al. [6] propose Nonmonotonic Document Frequency (NMDF), a weighting which weights moderately frequent terms highly. We are not aware of other work that has suggested this particular weighting for tags, however.

The most related work to our experiments in Sections 6.1 and 6.2 is our own work [7], as discussed in Section 1. Some older work, for example, DeZelar-Tiedman [3] and Smith [15] looks at the relationship between tagging and traditional library metadata. However, these works tend to look at a few hundred books at most, and focus on whether tags can enhance libraries. Also related to these experiments, there has been some work on association rules in tagging systems, including work by Schmitz et al. [14] and Heymann et al. [8]. However, that work focused on prediction of tags (or other tagging system quantities). We believe our work is the first to look at relationships between tags and library terms using methods inspired by association rules.

We are unaware of other work either examining $-tags (or even suggesting paying for tags) or attempting to understand how tagging works as a data management or information organization tool (i.e., in the same sense as libraries) in a large-scale, quantitative way.

8. CONCLUSION

We conducted a series of experiments that suggested that tagging systems tend to be at least somewhat consistent, high quality, and complete. These experiments found the tagging approach to be suitable for synonymy, information integration, paid annotation, programmatic filtering for quality, and for situations where an objective and high recall set of annotations covering general topics is needed. In a span of only a few years, LibraryThing has grown to tens of millions of books, and the groups developed by taggers are quite close to the groups developed by professional taxonomists. This is a testament both to the taxonomists, who did a remarkable job of choosing consensus controlled lists and classifications to describe books, and to tags which are unusually adaptable to different types of collections. Strikingly, we found that a particular type of user tag (moderately common user tags) is perceived as even more helpful than expert assigned library annotations. These two sets of experiments are mutually reinforcing. Overall, tags seem to do a remarkably good job of organizing data when viewed either quantitatively in comparison to "gold standard" library metadata or qualitatively as viewed by human evaluators.

9. REFERENCES

[1] E. H. Chi and T. Mytkowicz. Understanding the efficiency of social tagging systems using information theory. In *HT '08*.

[2] M. Clements, A. P. de Vries, and M. J. Reinders. Detecting synonyms in social tagging systems to improve content retrieval. In *SIGIR '08*.

[3] C. DeZelar-Tiedman. Doing the LibraryThing in an Academic Library Catalog. *Metadata for Semantic and Social Applications*, page 211.

[4] S. Golder and B. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198, 2006.

[5] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *WWW '07*.

[6] T. Haveliwala, A. Gionis, D. Klein, and P. Indyk. Evaluating strategies for similarity search on the web. In *WWW '02*.

[7] P. Heymann and H. Garcia-Molina. Contrasting Controlled Vocabulary and Tagging: Experts Choose the Right Names to Label the Wrong Things. In *WSDM '09 Late Breaking Results*.

[8] P. Heymann, D. Ramage, and H. Garcia-Molina. Social Tag Prediction. In *SIGIR '08*.

[9] T. Mann. *Library research models: A guide to classification, cataloging, and computers*. Oxford University Press, USA, 1993.

[10] T. Mann. *The Oxford Guide to Library Research*. Oxford University Press, USA, 2005.

[11] C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *Proc. of HYPERTEXT'06*.

[12] S. Oldenburg, M. Garbe, and C. Cap. Similarity Cross-Analysis of Tag / Co-Tag Spaces in Social Classification Systems. In *CIKM '08: Workshop on Search in Social Media*.

[13] G. Piatetsky-Shapiro. Discovery, analysis, and presentation of strong rules. In *KDD'91*.

[14] C. Schmitz, A. Hotho, R. Jaschke, and G. Stumme. Mining Association Rules in Folksonomies. *IFCS'06*.

[15] T. Smith. Cataloging and You: Measuring the Efficacy of a Folksonomy for Subject Analysis. In *Workshop of the ASIST SIG/CR '07*.