

# Folksonomy and Controlled Vocabulary in LibraryThing

## Introduction

“Folksonomy” is a term coined by Thomas Vander Wal from “folk” and “taxonomy” (Smith, 2004) to describe systems in which people (not professional indexers) assign natural language descriptors to resources. Many definitions of varying precision and breadth have been used for “folksonomy” (Vander Wal, 2005), and a variety of similar terms have been proposed (Merholz, 2005), but the term folksonomy appears to have become the most accepted. Because of the echo of “taxonomy”, folksonomy is often incorrectly described as classification, but it is actually more correctly labeled categorization or indexing (Mathes, 2004).

Folksonomies appear in many current web applications, most famously in the “social bookmarking” service **del.icio.us**<sup>1</sup>, in which users bookmark and tag web pages, and the photo sharing site **flickr**<sup>2</sup>, where users can upload and tag photographs and other images.

---

<sup>1</sup> <http://del.icio.us>

<sup>2</sup> <http://flickr.com>

Folksonomies have also begun to appear in more “scholarly” services, such as **Connotea**<sup>3</sup> and **CiteULike**<sup>4</sup>, which are similar to del.icio.us but oriented toward article citations. Folksonomies are also being applied to books, in the University of Pennsylvania’s library catalog (a project called **PennTags**<sup>5</sup>), and in **LibraryThing**<sup>6</sup>, a social site for cataloging personal libraries. These examples are especially interesting, since the resources involved (articles, books) are already indexed using controlled vocabularies, providing a fertile ground for exploring the similarities and differences in the functions of folksonomies and controlled vocabularies. This article examines properties of folksonomies in general and then looks at the example of LibraryThing in particular.

### ***Definitions***

Vander Wal (2005) asserts that the three important components of folksonomy are the following:

1. The **users**;
2. The **resources** being described, with a unique identifier such as a URL or ISBN; and,
3. The descriptors or “**tags**” used to describe the information resource.

The act of assigning descriptors is often referred to as “tagging.”

In addition, the folksonomy has two important aspects:

4. It is performed for **personal** organization and retrieval.
5. It is performed in a **social** environment, and greater numbers of users improve the system.

---

<sup>3</sup> <http://www.connotea.org>

<sup>4</sup> <http://www.citeulike.org>

<sup>5</sup> <http://tags.library.upenn.edu>

<sup>6</sup> <http://www.librarything.com>

The personal aspect makes participation in a folksonomic system worthwhile for users to pursue, since the system directly benefits their own organization and retrieval (Mathes, 2004).

The social aspect makes folksonomies useful not only for retrieval, but also for discovery. With any two of the three components (users, resources, or tags), one can find more of the third—other users with similar interests, resources similar to those already found, or descriptors similar to those already used (Vander Wal, 2005).

### ***Properties of Folksonomy***

Folksonomy uses an uncontrolled vocabulary, so it has much in common with any other system using natural language for descriptors. However, folksonomy also employs a social system absent from other systems of indexing that sets it apart.

### **Folksonomy as Uncontrolled Vocabulary**

As uncontrolled vocabularies, folksonomies suffer from many difficulties such as ambiguity in the meaning of and differences between terms, a proliferation of synonyms, varying levels of specificity, and lack of guidance on syntax and slight variations of spelling and phrasing (Spiteri, 2005; Mathes, 2004). Syntax is also system-dependent: while some systems allow spaces in tags or differentiate between capital and lowercase letters, others do not (Mathes, 2004). Additionally, indexer error may be higher in folksonomies because of incorrect usage; Spiteri (2005) observes the application of “archaeology” as a tag to materials about dinosaurs and prehistoric microbes. And, while controlled vocabularies are used by professional indexers with

some claim to objectivity, the interests of the user of the folksonomy are explicitly subjective.

Uncontrolled vocabularies do have several significant advantages over controlled vocabularies: the barriers to entry in terms of effort, time, and cognitive burden are much lower than in employing a controlled vocabulary (Mathes, 2004). This property allows much more widespread adoption of tagging systems.

### **Folksonomy as Social System**

Large numbers of users in a folksonomy present opportunities not present in traditional indexing, with a limited number of indexers and little overlapping in the materials indexed. The large number of indexers, combined with the fact that indexers and the audience are the same, allows a tight feedback between indexing and use involving repetition and imitation (Udell, 2004; Spiteri, 2005). Through these processes, users negotiate the meanings of terms to reach a consensus, much as markets negotiate prices. Individual meanings need not be given up—this market process does not necessarily result in homogenization—but community meanings emerge from the aggregation of individual effort (Shirky, 2005).

Folksonomy systems have a number of mechanisms to facilitate this shared meaning. These include suggesting popular tags used by others for the same resource, exposing the statistical relationships between tags that are often used together, and allowing users to collaborate to combine tags they think are equivalent. These replace some of the functions of the thesaurus in a controlled vocabulary.

However, Shirky (2005) points out that tags that seem equivalent, which might be collapsed to a single preferred term in a controlled vocabulary, may exhibit significant differences in the consensus meanings achieved through folksonomies, e.g. *movies v. cinema* or *queer v. homosexual*. These subtle differences are achievable through a social system, where wide subjective usage results in an emergent consensus that does not occur easily by deliberation alone. In addition, Shirky has pointed out that tags can be probabilistic; a resource can be considered “partially” something, unlike in a single-indexer controlled vocabulary, where the resource either fits the category or doesn’t.

## **The Example of LibraryThing**

LibraryThing is a website created by Tim Spalding in which users can catalog their personal collections of books, leveraging bibliographic data from Amazon.com and from libraries. Users can categorize books using tags and discover other users with similar books and interests. To date, LibraryThing has over 4 million books cataloged (*LibraryThing Zeitgeist*, 2006).

Since LibraryThing operates on books and makes use of library data, many of its books have subject headings assigned to them. The majority of these are likely to be Library of Congress Subject Headings (LCSH), but because the data is from many libraries, there are likely to be other English-language subject heading systems in use (such as the Sears subject headings), and there are subject headings from non-U.S. libraries including some not in English. Since the interface of LibraryThing is English,

however, the vast majority of users are English-speaking and the most popular tags are English.

LibraryThing's tagging system is an example of a folksonomy, meeting the definition and having the properties discussed previously. Subject headings comprise a thesaurus of controlled terms, assigned by cataloging librarians to describe books using a relatively objective protocol.

The next section analyzes the similarities and differences in tags and subject headings in LibraryThing.

### ***Analysis***

As of July 26, 2006, LibraryThing has 213,862 unique tags and 174,072 unique subject headings, but these numbers include minor variations of spelling and punctuation (*LibraryThing Zeitgeist*, 2006; Spalding, personal communication). Subject headings are available only for books for which the data is derived from library catalogs (excluding the Amazon.com data, which does not include subject headings). As a result, the coverage of subject headings in LibraryThing is narrower than that for tags (Spalding, personal communication).

The terms in subject headings include topical subjects, geographical locations, time periods, forms, and other categorizations; similar concepts appear in tags. Tags also include information that applies to a particular user's copy, such as "signed", or location tags such as "living room shelf", which would be handled by bibliographic data other than the subject headings in a library record. However, since such tags are personal and context-dependent, they have small numbers of users compared to topical

tags, and so are easily excluded from popularity-based information. Other context-dependent tags that are more universal include “read” and “unread”, indicating whether or not a user has yet read the book; these tend to be more common—“read” is the eighth most popular tag on LibraryThing (*LibraryThing Zeitgeist*, 2006).

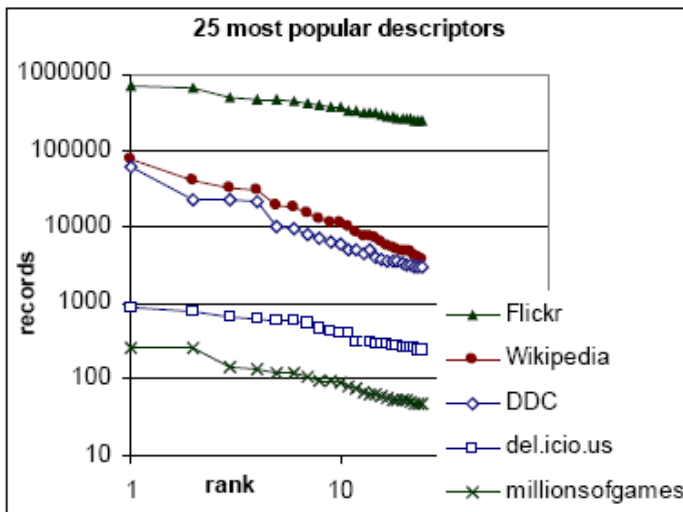
Subject headings are pre-coordinated, meaning that conjoint concepts are coordinated into a single heading at the time of indexing; e.g. “England > Fiction” (Jacob, 2004). In contrast, tags tend to be post-coordinated, embodying a single idea; e.g., the item with subject heading “England > Fiction” might have the tags “England” and “fiction” both applied to it (and can be retrieved in a search using a Boolean AND). The pre- and post-coordinate systems have various merits. In general, post-coordinate systems are simpler because the coordination of terms does not have to be anticipated. This works especially well for de-coupled term combinations, such as a topical subject and a geographic locale or time period. However, it has been observed that in combinations of topical subjects, pre-coordinated terms can be more expressive: compare the subject headings “History > Philosophy” and “Philosophy > History” (Blachly, 2006).

Looking at the top 75 tags and subject headings by popularity, some differences become apparent.

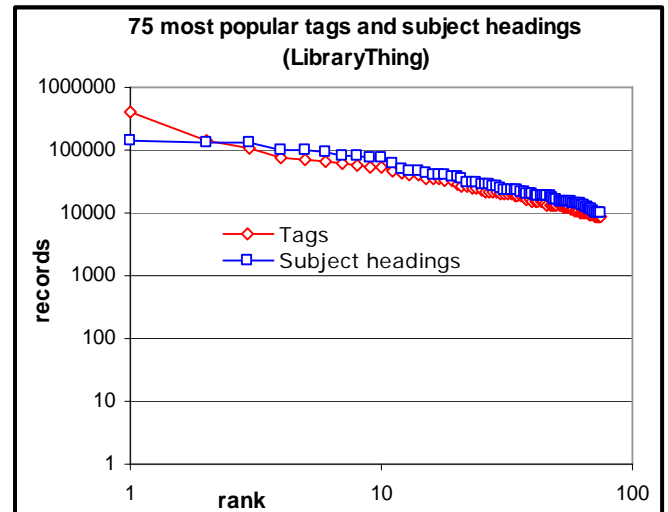
It has been shown that tag distributions for del.icio.us, flickr, and millionsofgames (a video-game tagging site) obey a power law distribution with  $\lambda < 0.6$ , while more structured vocabularies and classifications (the Wikipedia category

system and the Dewey Decimal Classification) are distributed according to a power law with  $\lambda > 0.9$  (Voss, 2006). Figure 1 shows these distributions.

The distributions of tags and subject headings in LibraryThing exhibit somewhat different behavior. Although the tails of the distributions (for tags with rank  $\geq 10$ ) are distributed according to power laws,  $\lambda \approx 0.9$  for both subject headings and tags, indicating that the differences in  $\lambda$  found in Voss may not be the result of differences in vocabulary structure (see Figure 2).



**Figure 1.** Distribution of tags in various systems. Note logarithmic scales. (from Voss, 2006)



**Figure 2.** Tag and subject heading distribution in LibraryThing.

Additionally, the most popular tags and subject headings (with rank  $< 10$ ) are distributed anomalously. The top tags are more popular than the distribution for the tail would suggest, while the top subject headings are less popular (see Figure 2). This may be an artifact of the immense popularity of the number one tag, “fiction”. (It is notable that the subject headings have no term so broad; “Fiction” is coordinated with other terms or appears as a partial concept in terms such as “Science fiction”, “Mystery

fiction”, and so on.) Further quantitative analysis of folksonomies and subject headings systems is needed to understand their statistical structures.

A qualitative evaluation of the top 75 tags and subject headings yields several observations. First, it is clear that both tags and subject headings exist in duplications with minor differences in spelling or punctuation. This is to be expected with tags since there is no vocabulary control among users, but for subject headings it is the result of multiple vocabularies (beside LCSH) or, possibly, because of human error. LibraryThing works around these difficulties in several ways. First, it assumes terms with differences only in capitalization are the same. Second, it supplies links to statistically related tags and subject headings, by calculating the frequency with which they are used in conjunction on items (see Figure 3). Finally, LibraryThing offers a facility for users to combine tags they consider to be equivalent, harnessing human intelligence to gloss over variations of spelling and phrase. For example, the tag “science fiction” is combined with tags such as “scifi”, “sci-fi”, and “science fiction”. To this author’s knowledge, it is the only folksonomy currently using such a feature.

Combining tags essentially makes up for one of the functions of an absent controlled vocabulary. It makes tags more useful, but caution is also needed. Spalding (2005a) has observed that Shirky’s remarks about the differences in “film” versus “cinema” are borne out in LibraryThing, as well as even more subtle differences, such as between “humor” and “humour”.

LibraryThing BETA Jonathanweber [sign out]  
 Catalog your books online.

Your library Add books Your profile Tags Pssst! Search Joy Zeitgeist Groups Talk About Blog

### Tag info: childrens

① Includes: childrens, children's, children's book, children's books, kid lit, childrens book, children's literature, kid book, kid books, childrens lit, kids book, kids books, children lit., childlit, literature.childrens, children's lit, kid-lit, juv, children's lit., child lit, childrens books, childrens fiction, kids, children's literature, childens, childrens literature (what?)

Used **63,869** times by **2,580** users.

**Using the tag "childrens"**

tarpfarmer (1867), shawna (1649), mcghol (1251), Wombat (914), tpewc (902), everdonbooks (872), UberTumbleweed (724), chanale (658), casaloma (653), bluestar50 (650), thefoodpornographer (644), alicebok (617), tripleblessings (549), antimuzak (518), lorrin74 (509), onefear (507), bette1126 (492), Genevieve1 (471), carminowe (453), ErinRebekah (400) — see more

**Most often tagged "childrens"**

Prince Caspian : the return to Narnia by C. S. Lewis (145)  
 The horse and his boy by C. S. Lewis (135)  
 The silver chair by C. S. Lewis (132)  
 The voyage of the Dawn Treader by C. S. Lewis (136)  
 The lion, the witch, and the wardrobe by C. S. Lewis (215)  
 The magician's nephew by C. S. Lewis (141)  
 The last battle by C. S. Lewis (136)  
 Harry Potter and the sorcerer's stone by J.K. Rowling (486)  
 The wind in the willows by Kenneth Grahame (163)  
 The secret garden by Frances Hodgson Burnett (208)  
 Charlotte's web by E. B. White (179)  
 Harry Potter and the Chamber of Secrets by J.K. Rowling (404)  
 Winnie-the-Pooh by A. A. Milne (100)  
 Harry Potter and the prisoner of Azkaban by J.K. Rowling (392)  
 Harry Potter and the goblet of fire by J.K. Rowling (396)  
 Harry Potter and the Order of the Phoenix by J.K. Rowling (403)

② **your tags | LT tag cloud | LT author cloud**

Search tags

② **Related tags** (show numbers)

adventure **animals** art biography **board book**  
**book** British **chapter book** children **Christmas**  
**classic easy reader** England fairy tales  
 family **fantasy fiction** hardcover  
 harry potter historical historical fiction history  
 humor humour **illustrated** library literature  
 magic **mystery** no cover **nonfiction**  
**novel own** Owned paperback **picture**  
**book poetry read** reference school  
 science science fiction **series** short stories  
 unread **young adult**

③ **Related Subjects**

Fantasy (3,820)  
 Children's stories (3,352)  
 Magic > Fiction (2,746)  
 Schools > Fiction (2,512)  
 (2,503)  
 England > Fiction (2,485)  
 Wizards > Fiction (1,945)  
 Wizards > Juvenile fiction (1,840)  
 Fantasy fiction (1,818)  
 Potter, Harry (Fictitious character) > Juvenile

**Figure 3.** LibraryThing tag info page for tag "childrens", showing (1) tag combinations, (2) related tags, (3) related subjects.

In the top 75, tags and subject headings exhibit a reasonable degree of overlap of topics, although the differences between pre- and post-coordinated terms must be taken into account. The top 75 subject headings appear to exhibit a slant because of certain highly popular works, including a heavy dosage of headings such as "Wizards > Fiction", "Magic > Juvenile Fiction", "Potter, Harry (Fictitious character) > Juvenile fiction", "Middle Earth (Imaginary place) > Fiction", and so on. Top tags include information about form that would be characterized by different bibliographic data in a library catalog, such as "series", "literature", or "paperback", so there are no equivalent

subject headings. Subject headings appear to make more fine distinctions in topics than do tags, with a litany of headings depicting characters and their relationships, such as "Fathers and daughters > Fiction", "Orphans > Fiction", and even "Triangles (Interpersonal relations) > Fiction", kinds of concepts which are largely absent from the highly popular tags.

Blachly (2006) has observed that tags in LibraryThing can exhibit the probabilistic behavior described by Shirky by comparing the tags and subject headings for "Dystopia"; where only a limited number of works have the subject heading, many works have the tag, but applied only a few times. Blachly also notes several instances in which tags provide a more natural discovery path for items because of their more direct language; e.g. the tags "queer" and "gay fiction" for Maupin's *Tales of the City*, where the top subject headings are "City and town life > Fiction" and "Humorous stories".

## **Conclusions**

LibraryThing shows that folksonomies and controlled vocabularies can harmoniously coexist, and that correlations between the two can be useful. To this author's knowledge, it is the only example where a folksonomy and a controlled vocabulary are being used in conjunction in this way. Spiteri (2005) has suggested that folksonomies could especially benefit from alignment with focused vocabularies, such as the Getty Thesaurus of Geographic Terms.

Clearly, tags and subject headings derive from similar principles and work toward similar ends. Equally clearly, they have radically different models of application and use

that cannot be completely reconciled. So, what can systems of subject headings learn from tags? What can folksonomic systems learn from subject headings?

Folksonomic systems generally make excellent use of the wealth of data in exposing related tags. In controlled vocabularies like subject headings, cross-references are generally used to the same end. However, subject heading lists could employ the same statistical techniques as folksonomies to find related subjects (which LibraryThing does), and useful differences from the anticipated cross-references might appear. To date, this type of data has not been leveraged in library catalogs

Spiteri (2005) has also suggested that folksonomies could be used as a basis to develop controlled vocabularies that match the language of users. This might be an especially useful technique in a controlled system, such as a corporate intranet.

Folksonomic systems may benefit from alignment with controlled vocabularies, especially where there is a strong ontological basis for the vocabulary—in geographic names or time periods, for example. However, Shirky (2005) points out that the classification of even something so concrete as geographic locations may be seen very differently by users, and so tagging practices may not line up precisely with a chosen controlled vocabulary.

A system such as LibraryThing, focusing on books, can benefit from observations of the practices of bibliographic description, especially in how cataloging data is shared and customized for local use. It might be beneficial, for example, to segregate personal tags and social tags (or perhaps *container tags* and *content tags* are clearer terms), keeping such tags as “read” and “signed” separate from “science fiction” and “politics”.

Further comparison and contrast of folksonomies and controlled vocabularies will no doubt be instructive as folksonomies continue to increase in popularity.

## References

- Blachly, Abby (2006). "Tagging meets Subject Headings" in *Thing-ology Blog*, May 14, 2006. Retrieved July 31, 2006 from [http://www.librarything.com/thingology/2006\\_05\\_01\\_archive.php](http://www.librarything.com/thingology/2006_05_01_archive.php).
- Jacob, Elin K. (2004). "Classification and categorization: a difference that makes a difference." *Library Trends* 52(3): 515–540.
- LibraryThing Zeitgeist* (2006). Retrieved July 26, 2006 from <http://www.librarything.com/users.php>.
- Mathes, Adam (2004). "Folksonomies—Cooperative Classification and Communication Through Shared Metadata." Retrieved July 31, 2006 from <http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html>.
- Merholz, Peter (2005). "Mob indexing? Folk categorization? Social tagging?" in *Peterme.com*, January 3, 2005. Retrieved July 31, 2006 from <http://www.peterme.com/archives/000444.html>.
- Shirky, Clay (2005). "Ontology is Overrated: Categories, Links, and Tags." Retrieved July 31, 2006 from [http://www.shirky.com/writings/ontology\\_outrated.html](http://www.shirky.com/writings/ontology_outrated.html).
- Smith, Gene (2004). "Folksonomy: social classification" in *Atomiq*, August 3, 2004. Retrieved July 31, 2006 from [http://atomiq.org/archives/2004/08/folksonomy\\_social\\_classification.html](http://atomiq.org/archives/2004/08/folksonomy_social_classification.html).
- Spalding, Tim (2005a). "New: Tag pages and related tags" in *LibraryThing Blog*, November 6, 2005. Retrieved July 31, 2006 from <http://www.librarything.com/blog/2005/11/new-tag-pages-and-related-tags.php>.
- Spalding, Tim (2005b). "Combining tags (heresy!)" in *LibraryThing Blog*, December 18, 2005. Retrieved July 31, 2006 from <http://www.librarything.com/blog/2005/12/combining-tags-heresy.php>
- Spiteri, Louise (2005). "Controlled Vocabularies and Folksonomies." Presentation at Canadian Metadata Forum, Ottawa, ON, Spetermber 27, 2005. Retrieved July 31, 2006 from <http://www.collectionscanada.ca/obj/014005/f2/014005-05209-e-e.pdf>.
- Udell, Jon (2004). "Collaborative Knowledge Gardening" in *InfoWorld*, August 20, 2004. Retrieved July 31, 2006 from [http://www.infoworld.com/article/04/08/20/34OPstrategic\\_1.html](http://www.infoworld.com/article/04/08/20/34OPstrategic_1.html).

Vander Wal, Thomas (2005). "Folksonomy Definition and Wikipedia" in *Off the Top*, November 2, 2005. Retrieved July 21, 2006 from <http://www.vanderwal.net/random/entrysel.php?blog=1750>.

Voss, Jakob (2006). "Collaborative Thesaurus Tagging the Wikipedia Way." Retrieved July 31, 2006 from <http://arxiv.org/abs/cs.IR/0604036>.