FOLKSONOMY AS A THING FOR A LIBRARY
An analysis of user generated metadata in LibraryThing

Vincent STERKEN
Document Management Consultant, I.R.I.S. Solutions & Experts

▪ De laatste vijftig jaar is de creatie en beschikbaarheid van informatie gestaag toegenomen. Om hiermee om te kunnen gaan, hebben informatie specialisten naar nieuwe manieren gezocht om de aanwezige kennis toegankelijk te maken. Gedurende de laatste jaren echter hebben gemeenschapsgestuurde categorisatie tools het licht gezien op het internet. Categorisatie met deze zogenaamde "folksonomies" gebeurt door trefwoorden, of "tags", toe te voegen aan opgeslagen informatie. Mijn onderzoek heeft zich toegespitst op de folksonomie gebruikt door de website *LibraryThing*. Deze laat toe om boeken online te catalogeren. Dit artikel analyseert de effectiviteit van de site aangaande inhoudsbeschrijvingen van boeken, alsook het verschil in de manier waarop "leken" en informatiespecialisten tags toekennen. Als conclusie wordt gesteld dat folksonomies kunnen bijdragen aan traditionele classificatie en categorisatie schema's.

▪ Ces cinquante dernières années, la création et la disponibilité de l'information ont connu une évolution constante. Afin de faire face à cela, les professionnels de l'information ont recherché de nouvelles manières de rendre disponible tout ce savoir. Ces dernières années, des outils de catégorisation collaborative, appelées "folksonomies", ont vu le jour sur Internet. Cette catégorisation est élaborée à l'aide de mots-clés, ou "tags", servant à décrire l'information. Mon étude a porté sur la folksonomie utilisée sur le site *LibraryThing*, qui permet le catalogage de livres en ligne. Cet article analyse l'efficacité du site dans la description des livres, tout comme la différence entre l'indexation réalisée par un catalogueur professionnel et un non-professionnel. En conclusion, on verra que les folksonomies peuvent apporter une contribution utile aux schémas traditionnels de classification et de catégorisation.

In light of obtaining a Master's degree in Business Information and Services Management (VUB), I have written a thesis entitled *Classified: Analysis of user generated metadata in the LibraryThing folksonomy*[1], under the guidance of Céline Van Damme. The findings of this work are presented in this article.

The Age of Infoglut

Since the Second World War the amount of information produced has expanded exponentially. Inventions such as typewriters, microfilms, photocopiers, and computers have each in their own way enlarged the available capabilities of data dissemination and storage. At the same time finding the right data and information has become more and more difficult. The bigger a corpus becomes, the more a need arises for an efficient system to gain access to it. In recent years, with the advent of information and communication technology, this problem has been exacerbated. The ever growing power of computers and size of storage media has seen the total size of information production increase into exabytes. The sharing capacity of the internet has acted as a great facilitator in this respect. In 2000 the School of Information management and information systems (SIMS) of the University of Berkeley (USA) estimated that on the (visible) World Wide Web 20 to 50 terabytes was available. During a follow-up study three years later, SIMS noted that the volume had tripled to 167 terabytes, which is almost 17 times the information residing in the repositories of the Library of Congress (Washington)[2]. In the IDC White Paper *The expanding digital universe*, Gantz et al have calculated that in 2006 161 exabytes of digital information was created, captured and replicated. Between 2006 and 2010 this will have increased more than six fold to 988 exabytes of information[3].

If we want to be able to effectively and efficiently use all this information, robust and flexible systems will need to be developed to accommodate search and retrieval. The information science community has several well established tools which it can use, such as classification schemes and thesauri. In an age of ever growing information production traditional methods of classification and categorization sometimes fall short of their goal. Dedicated professionals have been developing new ways of organizing information, while expanding on the knowledge they already have. Until recently, categorization tools were exclusively in the hands of these professionals. Innovative ways of using the internet have changed this. Out of a need for organizing information on the web, a grassroots classification, dubbed "folksonomy", was developed. With the aid of folksonomies the searcher can organize information in a personal, semantically meaningful way through the use of personal keywords, called "tags".

Folksonomies might be able to provide a tool to categorize large amounts of information at a low cost. As we will see later on, it is no perfect solu-

tion, but it is a practical one. Information that would otherwise remain hidden might become accessible thanks to the incremental nature of these systems. Although natural language systems already existed, they have never really been deployed on such a large scale as now. The reason folksonomies have taken a large flight can be found in the fact that they are useful for the user herself, but also allow for the sharing of resources, thus creating a social network and enlarging its capabilities for retrieval. The first folksonomies were used mainly for storing web-based information, i.e. URLs. Quickly, however other uses have seen the light. Tags are being used for academic papers, life goals, movies, books, etc. But before we get ahead of ourselves, let's first explain what a folksonomy is.

Folksonomies

A folksonomy is a user-generated categorization method with which web content can be categorized and retrieved through the use of open-ended labels called tags[4]. It has been dubbed grassroots classification, collaborative tagging, ethnoclassification, folk classification, open tagging, social classification, faceted hierarchy, etc. The neologism folksonomy was first coined by Thomas Van der Wal, who describes it as being "*the result of personal free tagging of information and objects (anything with a URL) for one's own retrieval. The tagging is done in a social environment (shared and open to others). The act of tagging is done by the person consuming the information. The value in this external tagging is derived from people using their own vocabulary and adding explicit meaning, which may come from inferred understanding of the information/object as well. The people are not so much categorizing as providing a means to connect items and to provide their meaning in their own understanding*"[5].

A tag is basically a keyword or reference which you, the user of the system, can add to a resource to describe the resources' aboutness. In order to retrieve the saved information it suffices to either use a search form or click on the term in question in a "tagcloud". A tagcloud represents tags in a stream of words with varying sizes. The size denotes the frequency of the use of a tag in relation to the other tags that are displayed.

The defining characteristics of a folksonomy are its bottom-up structure, its lack of hierarchical control, and the social context in which it is used[6]. The most common examples are the social bookmarking site del.icio.us[7] and the photo sharing site Flickr[8]. The first allows users to tag a URL of a website with relevant keywords, while the latter allows the tagging of photographs. Tags can be applied to a number of resources besides bookmarks and pictures, such as music, videos, books, academic papers, events, blogs, even life goals. The primary objective is refindability of saved resources by the user himself. Because of the fact that other users can see the resources that have been saved and can search the saved tags a communal aspect is inherent to folksonomic systems.

For a more comprehensive explanation (in Dutch) of how a folksonomy works, I recommend reading Ms. Céline Van Damme's article *Van folksonomieën naar ontologieën*, published in this magazine at the beginning of last year[9].

Tagging books

Since its inception folksonomies have been researched for different reasons. Up until now all the research that has been done, has assumed that the system allows a direct access to the source of information that is indicated in the search result. Most folksonomies are web-based, which means that what we're looking for is only a click away. Contrary to what some people believe, not all the paths to enlightenment are to be found on the internet. Sometimes we still need to actually go out and get a book in a library or consult the original document in an archival repository. Nevertheless, folksonomic tools can be of assistance. The question here is: do users tag differently when categorizing paper-based media? In order to examine this question, data was extracted from the social cataloging site *LibraryThing.com* (with the amiable help of Tim Spalding, the site's founder). As a side question, the possible differences in tagging between professional indexers (such as librarians) and non-professionals were regarded.

LibraryThing

"LibraryThing *[LT] is a social network for bibliophiles. You catalog the books you have... or are interested in, and the books you have connect you to other people*"[10]. The site went online on 29 August 2005. It had, at the time of writing my master thesis, over 450 000 registered users (sometimes called *thingamabrarians*), who have saved more than 28 million books with more than 37 million tags. Spalding discerns three levels of use: personal cataloging, social networking, and social cataloging.

Personal cataloging

In an easy to use interface users can create a virtual bookshelf. To add a book you simply use

the provided search box by typing in some words from the title, the author or an ISBN. The data about the books are imported automatically through a connection to libraries (providing MARC and Dublin Core records) and commercial booksellers.

To each book in your library you can add tags. These are designed to be a "*simple way to categorize books according to how you think of them, not how some library official does. Anything can be a tag - just type words or phrases, separated by commas*"[11]. There exist several views of a catalog. One possibility is shown in figure 1. Here you see a library as a list. Another way is showing only the covers in a larger font. The underlying data can be accessed by clicking on a specific cover.
Within a catalog it is possible to search in the different fields, either separately or combined. So, you can find books by typing in keywords, which are then matched with either all fields or specific ones, i.e. titles/authors, tags, reviews, comments, subjects.
Since *LT* uses as a system a folksonomy, tag-

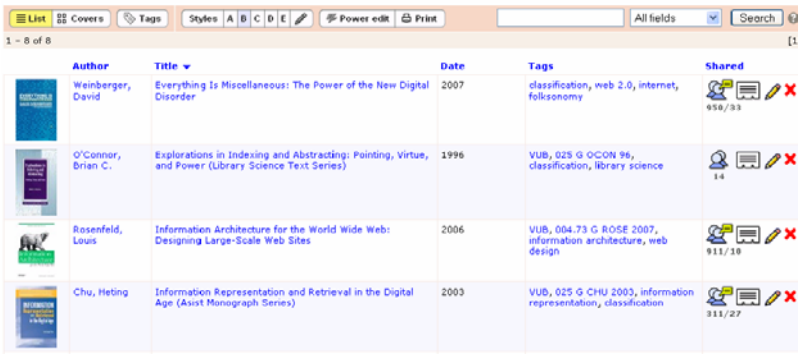

Fig. 1 : Example of a personal library.

clouds are naturally present. Different representations are available. On the highest level, we find a fairly large tagcloud of the top 75 tags, as well as an authorcloud of the top 75 authors[12]. When we go a level down, we notice that each book in *LT* has its own tagcloud. Finally, there is a tag- and authorcloud available for each user with all the tags in the personal catalog (also viewable in the form of a list).

Users can choose whether to keep their library private or public. A private catalog can only be seen by the user himself, while a public one is open for the world to see.

Social networking

LT is not only an online cataloging service. It is "*also an amazing social space, connecting people with similar libraries. It also makes book recommendations based on the collective intelligence of the other libraries*"[13]. The site started out as a way of cataloging one's own library in an easy and cheap manner. The similarities in users' collections became apparent and a social aspect emerged. Like Amazon, automatic recommendations are made about books that you might find interesting. Unlike Amazon, these are based on members' tastes and not on a sales model. "*Generating picks based on an entire collection is far more revealing than focusing on purchases. The stuff that you own is just a very powerful expression of yourself," Mr. Spalding says. "These catalogs represent a lifetime of collecting. Because of this intimacy,* LibraryThing *can also connect likeminded readers -- a sort of MySpace for bookworms. But the object is always to find more books, not to kindle online relationships or cliques. It's not about who you connect with as friends, it's about who you connect with through books," Mr. Spalding explains*[14]. Most of the interaction within the community takes place in the talk pages of the different groups. There exists a possibility to join one of the 3647 groups, ranging from Fantasy or Science Fiction Fans to Non-fiction Readers, Graduate Students, Happy Heathens and everything in between[15].

Up until recently, the first thing you would see after logging in was your personal library. Now every user has a private homepage. In true web 2.0 style, everything on it is customizable of course. The homepage gives an overview of recently added books, recommendations, what connections have been added, the last messages of the talk pages and much more. Another feature that can be seen here is local events. Users can submit events, bookstores and libraries in the local area, which are then pinned on a Google Maps mashup. All of this naturally promotes the social aspects of the site.

Social cataloging

According to Tim Spalding there is a natural ladder of use of *LT*. You start out cataloging your own, personal library. Because of the overlaps in catalogs and aided by the features of the site you develop a social network. All of this together creates what he calls "social cataloging". This can be done implicitly or explicitly. Explicit social cataloging is done for instance by members of the group *I See Dead People['s Books]*[16]. This group enters the private libraries of famous readers as library catalogs. Completed libraries include those of

Thomas Jefferson, Mozart and Tupac Shakur (2Pac). Implicit social cataloging can be considered a side result from using the system. Every bit of information about the books in *LT* that doesn't come from the abovementioned sources is user generated. This includes tags, "common knowledge", and editions. In the common knowledge pane information is added that, in general, does not appear in traditional classification schemes, such as important places and people or characters, and the awards and honors the book has received. In the editions pane all the different editions of the book are combined. This improves the findability. When you're searching for something, you're interested in the information and not necessarily if it's the hard or the soft cover.

The dataset and its tags

Data has been collected for the 200 top books[17] during the last week of March 2008. The object of this exercise is to study the usefulness of tags for the retrieval and description of books. A second element concerns the differences in tagging by information specialists like librarians.
As we have seen before a number of different groups exist. The group which is of special interest to this paper is *Librarians Who* LibraryThing[18], which describes itself as welcoming "*librarians, catalogers, archivists, students… or anyone else who wants to talk about metadata, tagging, FRBR, library 2.0, social software, cataloging, and, of course,* LibraryThing*!*" I believe it relatively safe to assume that, if not everybody, most people who belong to this group are in some way professionally affiliated with classification efforts.

On the Zeitgeist page an overview is given of a number of statistics concerning the users and the available resources. One of the categories is "top books"[19], which cites the 1000 books and authors most shared by the members of *LT*. The site's founder, Tim Spalding, graciously provided a php script which allowed me to extract aggregated data per book, which gives an overview of the different tags used per book and per group. The total number of users that have a given book in their library was added manually, based on the information provided by the top books page. The total population of users for these books is 1 231 385 users. The maximum number of users for a resource was 24 861, the minimum 3985, with an average of 6187,86 users per book.

Functions

In *The structure of collaborative tagging systems* Golder and Huberman have investigated what kinds of distinctions can be made between tags based on their function. Based on their findings they have defined seven categories[20]. Given the limited time and the amount of available tags, it was not possible to make an exhaustive list of possible terms. Therefore analysis was done on a sample of keywords taken from the individual books' tagclouds. These will be discussed in the next paragraphs:

- "***Identifying What (or Who) it is About****. Overwhelmingly, tags identify the topics of bookmarked items. These items include common nouns of many levels of specificity, as well as many proper nouns, in the case of content discussing people or organizations.*" For this category, analysis was done on a query of about 500 keywords (and their variations) of the first 50 books. These included the elements out of the titles and terms like "jesus", "christianity", "big brother", "psychology", "freedom", "growing up", "gender", "solitude", as well as names of characters and the countries or regions where the actions take place. This search resulted in a return of 104 306 applied tags, or 17,56% of the total frequency of 518 945 tags. When differentiating between librarians and non-librarians the percentages vary slightly, 21,36% and 17,53% respectively.
- "***Identifying What it Is.*** *Tags can identify what kind of thing a bookmarked item is, in addition to what it is about. For example, article, blog and book.*" Within *LT* it is pretty obvious that nearly all of the tagged content consists of books. However, the proposed rule is applicable. Against all odds, the tag "book" on its own occurs 398 times (0,12%) in the sample. Books come in different physical carriers, so the query was widened to include soft and hard covers, paper and hard backs, and e-books. A further distinction can be made on the basis of its purported function using the following terms: "fiction", "non-fiction", "textbook", "picture book", "series", "novel", "play", and "poetry". And finally, audio books and film adaptations were taken into account by adding "video", "DVD", and "CD". This adds up to a frequency of 23,85% (with a difference of 2,31% between the two groups).
- "***Identifying Who Owns It****. Some bookmarks are tagged according to who owns or created the bookmarked content…*" The owners of the saved content are of course the authors of the books (or their publishing company). This does not seem to be that relevant. Less than 3% of the tags contain the names of authors. It is not particularly useful to add this information, since the system in itself keeps a record of the author's name, the title, and ISBN numbers.
- "***Refining Categories****. Some tags do not seem to stand alone and, rather than establish categories themselves, refine or qualify exist-*

ing categories. Numbers, especially round numbers (e.g. 25, 100), can perform this function." In this sense tags like "Youth Author", "juvenile fiction", "urban fantasy", "classics", "short stories", and "thriller" are used in the above mentioned way. A query on 31 of these types of tags (and their variations) amounts up to 27,93%. The highest frequencies are noted in the variations on the term "literature" (6,21%), "classic" (excluding literature, 5,06%), "fantasy" (4,32%) and to a lesser extent "science fiction" (2,05%). Refining categories by using numbers only makes up 0,07% of the whole, which implies that it is not deemed all that important. Its relevance is a little bit higher for librarians (0,22%) than for non-librarians (0,06%), but at heart that doesn't change that much.

- "***Identifying Qualities or Characteristics**. Adjectives such as scary, funny, stupid, inspirational tag bookmarks according to the tagger's opinion of the content.*" Based on my own judgment and by scanning the tags in the database, 32 terms were selected to query this category. These included "best", "great", "loved", "hated", "cool", "fun", "overrated", "crap", "hilarious", "no cover", "insight", "signed", and "illustrated". Although it is hard to define the array of possible preferences users might have to express their feelings I believe that a large part is covered by the used terms. The result is rather disappointing. Less than 2% of the dataset is covered by this category. The highest ranked term is "favorite" (0,55%), followed by a steep drop to "edition" (0,17%).
- "***Self Reference**. Tags beginning with my, like mystuff and mycomments identify content in terms of its relation to the tagger.*" Tags beginning with "my" do not seem to be that important when describing books in *LT* (0,15%). The concept of ownership of the physical resource, i.e. the actual book, is expressed by the term "own" (1,47%) or by the owner's name if he is not the holder of the *LT* account. The exact number of names is difficult to ascertain as this would need to be done by comparing the dataset with every possible known name, while excluding character names from the books in question. The analysis for this category was done based on 22 terms, containing the words "wishlist", "room", "box", "shelf", "library", "borrow", "gift", and "acquired". An additional search was done on variations of letters of the alphabet. The total amount of tags are 54 986 (4,65%), which implies that this category is significant. The most important group seems to be tags related to the physical location of the book (1,51%), exemplified by terms like "location", "@home", "at mom's", and "box".
- "***Task Organizing**. When collecting information related to performing a task, that information might be tagged according to that task, in order to group that information together. Examples include toread, jobsearch.*" The total amount of tags related to task organizing takes up 5,96% of the dataset. Terms like like "read", "tbr", "r:*date*" "review", "buy" and "finished" were investigated. Unsurprisingly, the tags related to reading ("read", "unread", "to be read", etc.) take up most of the tags within this category (5,64%).

The percentages mentioned need to be taken with a grain of salt, since it is hard to know to what extent the sample is completely representative. Nevertheless, percentages of 5 and 20 to 30 can be deemed relevant. In summary, the categories *what is*, *what it is about* and *refining categories* account for ±70% of the tags. *Task organizing* only makes up ±7%, but I do believe that this number belies its importance. Tags that are intended for organization of tasks and time management are bound to have a transitory nature. Once a book is read, it makes no sense to keep the related tag "to be read". Those that give information about the year of reading will probably endure longer. In the *common knowledge* pane of the *details* subtab of a book it is possible, by clicking on edit, to tick off "to read". To find books where this is added, the user needs to go to the *common knowledge* page, which can be found through a small font link at the bottom of each page, and search for these words. Unfortunately everybody who has added this subsequently shows up in the search result. As far as I can tell it is not possible to refine your search in order to include only a specific user (at the time of writing). It is doubtful that this function can take the place of the easy method of just searching your own tags. A function like the one in the academic paper bookmarking site CiteUlike.org might be a useful addition of functionality. CiteUlike allows users to add a priority level to the papers being bookmarked, ranging from "I don't really want to read it" to "Top priority!"

Sen et. al. have examined the factors that influence the way people choose tags and to which degree community members share a vocabulary[21]. To conduct their experiment, tagging features were added to a movie recommendation site. They have adapted the seven categories presented by Golder and Huberman and collapsed them into three broader classes. *Factual tags* identify "facts", such as people, places, or concepts (*what it is, what it is about, refining categories*). *Subjective tags* express user opinions (*characteristics or qualities*). *Personal tags* have as intended audience the tag appliers themselves (*who owns, self reference, task or-*

ganization). The final distribution of tags across these classes was 63% factual, 29% subjective, 3% personal and 5% unknown. The analysis of *LT* is consistent with these findings in the sense that the majority of tags pertain to information about the resource in question, rather than being used for strictly personal comments. The whole point of using any categorization system is to find things again. So it is not illogical that, on a whole, the system doesn't get cluttered with tags that are not particularly useful.

The functions of tags in *LT* can be divided largely into two groups. They are either used for subject analysis or for practical purposes. The first group is represented by categories 1, 2 and 4 of Golder and Huberman, while the latter is represented by categories 6 and 7. Intellectual ownership (category 3) does not figure prominently since this kind of information is already supplied by the system. The attribution of characteristics (category 5) is not predominantly present, probably because there are other ways of expressing certain sentiments.

Information value

The question remains what the information value of tags concerning the aboutness of the resources is. The term "information value" is used here as being "*the information conveyed by the natural language term used in the tag and how this makes the tag useful for retrieval of and distinction between resources or not.*"[22] To understand how well tags fare in terms of subject analysis, a comparison was made with the subject headings assigned to each book. Subject headings in *LibraryThing* are based on the library data *LT* extracts from the different sources mentioned above. A large part will probably come from the Library of Congress Subject Headings, but other systems (mostly English, e.g. Sears, but also other languages) are present as well. Subject headings are available for books for which data has been derived from library catalogs, making their coverage narrower than that of tags. The used terms include topical subjects, geographical locations, time periods, forms and other hierarchical classifications[23].
Subject headings are very useful when browsing a certain subject area. For instance, "*under the tag for "civil war" is a haphazard collection of books. The [*LibraryThing*] subject page for "United States > History > Civil War, 1861-1865", on the other hand, provides a list of subdivisions, giving you the ability to do more educated browsing.*" Moreover, "*the ordered structure of subject headings gives added meaning. "History > Philosophy" is very different from "Philosophy > History' - a distinction that isn't necessarily apparent when searching "history" or "philosophy" separately as tags*"[24]. Terms from subject headings have the advantage of eliminating ambiguity concerning their meaning. They also make the relationships with related and combined concepts. When the searcher is not yet familiar with the subject area, the hierarchy can help provide a certain insight into the matter.

The application of subject headings to books is done by humans. Therefore the system is not infallible. Spalding gives the example of where the classification of the Library of Congress Subject Headings (LCSH) went wrong. Lisa Carey's novel *Love in the asylum* has as a subject heading "Alcoholics > Fiction". The work does not in fact have a lot to say about alcoholics. It does talk about Native Americans, which is nowhere to be seen in the LCSH. The *LT* tag cloud does not mention "alcoholics" or "alcoholism", but does mention "Native Americans". He also shows that certain categories that exist in *LT* and not in LCSH are as real as any *official* category. The rigidity of the existing classification and categorization systems prevents them to include new or emerging classes in a flexible manner. William Gibson's *Neuromancer* has as headings "Business Intelligence > Fiction", "Information highway > Fiction" or simply "Science Fiction". Connoisseurs of Science Fiction however know that this is a classic example of the sub-genre "Cyberpunk". Unsurprisingly in *LT* it is the book tagged the most with this term[25]. This shows that collaborative tagging can add value as a classification system. Cyberpunk is no less a very real category than any other officially sanctioned term. Tags create a certain amount of noise in the system. The sheer amount of users tagging certain content counteracts this problem by creating a consensus concerning the aboutness of a given resource.

There exists a certain overlap between tags and subject headings. When comparing them, the hierarchical relationships between the subject headings get lost in translation (so to speak). Although multiple word tags are allowed in *LT*, an exact comparison would not generate many results if the classes with their subclasses attached would be taken into account. No one in the sample uses the form "Family life > New England > Fiction", nor the more commonly used "Family life -- New England -- Fiction". The available subject headings in *LT* associated with the sampled books were "normalized" in order to make them useful. Associated terms were split up. If we take the example above for instance, the terms "family life", "New England" and "fiction" would be compared with the tags in the dataset. Upper and lower cases were eliminated, as were differences in plurals and singulars. In the sample of *LT* data this accounts for 36% of the tags being equal to

the associated subject headings. Because the term "fiction" is the most used tag the result is somewhat distorted. After disregarding this tag, the percentage drops to 21,24%. In both cases there were no really significant differences between the group of librarians and of the non-librarians. The librarians' tags exhibited a slightly larger overlap than the others (23,37% versus 21%). These tentative results correspond more or less to the findings of the *steve.museum* project. *Steve* was founded in 2005 to address the problems faced by art museums concerning access to their online collections. Their websites knew a growing number of visitors. Yet, these visitors had trouble navigating the digital collections. At the root lay a semantic gap between the formal descriptions, assigned by art historians and other specialists, and the vernacular language used by the general public for searching the database[26]. Jennifer Trant has noted that at least 70% of the tags submitted by regular users of the system (after elimination of misspellings and errant terms) were not in the taxonomy (going up to 90% for the top four most tagged works)[27]. Vanderwal has come to similar conclusions in his discussions with his clients. They have found that 30 to 70% of the terms used in tagging are not represented in their taxonomies[28].

The overlap mentioned above was derived from a direct comparison between the separated subject headings and tags per book. The tags for a given book were retained when these matched the subject headings. Subsequently, the total frequency of the times these tags were applied to the resource was counted and then aggregated. The entire dataset was taken into account. Therefore all the misspellings and idiosyncrasies of individual users were still present. Given the limited time for this research, it was not possible to correct these. However, in an attempt to eliminate a significant part a large number of tags were taken out of the equation. Since *LT* does not have a function that suggests spelling corrections, nor tags used previously by the same or other users, it is doubtful that "des fautes de frappe" are perpetuated. It is likely then that they will have a low frequency count. Although personal tags will be used more often, most of them will not be shared by the larger community. Here, again, a low frequency count can be expected. Following this reasoning, an arbitrary drop-off point was established, i.e. all tags with a frequency lower than 10. When the comparison is made again between subjects and tags, the percentage rises to 47,34. The difference between librarians and non-librarians becomes slightly bigger than before. The conformance to subject headings rises to 55,12% in the group of librarians, while the non-librarians stay closer to the total percentage (47,36%).

The subject headings were taken from the *LT* site itself. The correctness of this automatic extraction is hard to ascertain without having access to the raw data. Therefore the scope of the research was narrowed down. The same comparison was made based only on the Library of Congress Subject Headings (LCSH). For every book the associated Subject Headings were taken manually from the Library of Congress' online catalog. Every book in the online catalog is accompanied by a subject description, containing subject headings, classification numbers and in many cases one or more genres. As expected, an exact comparison between the full LCSH strings, as described above, yields very little result (1,41%). The same goes for the genre descriptions (1,10%). When the strings are split up into separate keywords and combine the result with LCSH and genre descriptions we get a coverage of 8,43%. Here as well, the group of librarians' conformance is higher (10,64%) than that of the non-librarians (8,18%). When we drop the tags with a frequency count below 10, this percentage rises to 13,14%. The difference between the two groups becomes significantly higher however. The librarians then account for 22,81%, while the non-librarians only take up 12,52%.

These findings indicate that the terms used as subject headings only conform to a very limited amount of the terms used in a natural language system such as the folksonomy of *LibraryThing*. The conformance within *Librarians who LibraryThing* is relatively higher. The difference, however, is not as great as one would expect. A possible explanation is that when the professional becomes the user he will act as one, interpreting the resource according to this level.

Subject headings are supposed to be a reliable, standardized way of defining what a book is about, with the intent of optimizing findability and retrieval. In a (relatively) direct comparison, the terms in the *LT* folksonomy only coincide with these in a limited way. Although tagging is used for a variety of functions, it is still aimed at organizing things. When searching on tags, the books are returned that have been tagged the most with that particular term. High frequency tags reflect a consensus within the user community concerning the aboutness of a resource. How well the top tags represent the aboutness of particular documents as compared to LCSH is another question. Simply put: it depends. In some cases they are equivalent, in other they both contribute something in terms of understanding and retrieval possibilities. At times tags are better suited for subject analysis thanks to the personal relationship of the tagger with the resource, yet sometimes they can be wildly incorrect.

To get a better idea of the added value of tags, high frequency tags together with their variations were compared to LCSH per book. Because of time constraints and the sheer size of the dataset, half of the resources (in terms of their frequency) were taken into consideration. For this restricted set 286 LC subject headings and 99 genre descriptions were found. A total of 318 terms were found in the top tags which did not appear in the LCSH. For the sake of clarity synonyms and terms with an almost identical meaning or intention were left out. The "unweighted" number was significantly higher. The terms represent concepts with varying depth. Some can be seen as narrower or related terms of LCSH, others as terms that weren't considered. A small, yet significant, amount comprise neologisms and concepts that exist within a subculture of fans, which have not yet found their way into the official canon of standardized controlled vocabularies, such as "cyberpunk", "steampunk" and "paranormal romance". These supplementary descriptors are relevant to subject analysis in that they enlarge the ways users can search a bibliographic database. To get the most out of a search query, a combination of the traditional tools library science has to offer and a social cataloging system, which has acquired critical mass, seems optimal.

Conclusions

In *LibraryThing* the different functions tags can have are very similar to those in other folksonomies. The notable difference is that they are mostly aimed at practical aspects instead of more emotive ones. In short, *LT* tags serve as retrieval aids and management tools.

In terms of information value, collaborative tagging provides a rich semantic means for categorization. When compared to traditional bibliographic systems, the *LibraryThing* folksonomy should not be seen as an alternative, but rather as a supplement. Descriptors ascribed by intermediaries are on a whole, fairly accurate in their subject analysis, yet not always complete. In general tags in *LT* are relatively accurate as well, but quite often on a different level. Sometimes they add refining or broader terms, at other times they introduce new or supplementary concepts.
Folksonomies are closer to new developments in new terminology and exhibit a greater and richer variety in terms. At the same time they are also plagued by this variety. A large part of the terms in the system can be considered to be clutter when it comes to subject analysis. Despite this particular drawback, the *LT* folksonomy has its benefits. To fully profit from these, a joining of forces is the best solution. Peter Morville cites in this context the concept of pace layering. He argues that society as a whole is constructed of several layers, each with a unique and suitable rate of change. "*The slow layers provide stability. The fast layers drive innovation. [...] In this discussion of metadata, the potential for a unifying architecture is self-evident. ... standards create a powerful, enduring foundation. [...] the fast-moving, fashionable folksonomies sit on top: flexible, adaptable, and responsive to user feedback. And over time, the lessons learned at the top are passed down [...] This is the future of findability and sociosemantic navigation: a rich tapestry of words and code that builds upon the strange connections between people and content and metadata*"[29]. Lambe translates this as working towards an array of knowledge infrastructure tools. Folksonomies provide the benefit of low design and low costs, while ontologies have the advantage of high precision and low ambiguity. Taxonomies cover the middle ground, attempting to balance design with discovery and precision with serendipity[30].

It has become clear that the different levels of interpretation of a document do not intermesh very often. Intermediary generated metadata (i.e. subject headings, thesaural descriptors,...) are rooted in the professional environment of indexers and catalogers. User generated metadata take their cue from the personal experiences and needs of the user in question; and, to a lesser extent, are coupled with a certain exposure to the community. The results of this research point in the direction of a clear scission between the two groups. The group of professionals in the field of information science does not really differ all that much from the larger community. It would seem that once the librarian becomes the user, she will act as a user and less as a professional cataloger. This is in accordance with the concept of the different layers within society. Every person also has different layers, different identities (e.g. mother/father, indexer, musician, child, etc.). It would be good for the catalogers who make use of a site such as *LT* to remember the potential lessons they have learned from being a user when they return to the workplace. Better yet, social cataloging sites should be used to drive changes, adaptations and updates of the stable layer of taxonomies. The first steps in this direction have already been taken with *LibraryThing* for Libraries[31]. It is essentially a series of widgets designed to enhance library catalogs with *LT* data and functionalities, such as book recommendations, tag browsing and links to other editions and translations.

To the question whether a controlled vocabulary or folksonomy is the best method for subject analysis, can only be answered with yes. As in,

the combination of both will probably yield the best results. The only problem with a folksonomy is that it needs enough users of the system to even out personal preferences. Once critical mass has been acquired a valuable consensus can be reached concerning the aboutness of a document. To this end, for the English speaking world for now, *LibraryThing* would make an excellent starting point.

Vincent Sterken
I.R.I.S. Solutions & Experts
Rue du Bosquet, 10
1348 Louvain-la-Neuve
vincent.sterken@gmail.com

January 2009

Notes

1 Sterken, V. Classified. Analysis of user generated metadata in the LibraryThing folksonomy, Unpublished master thesis, Vrije Universiteit Brussel, 2008.

2 Van Damme, C. *Folksonomies and entreprise folksonomies*. Unpublished master thesis, Vrije Universiteit Brussel, 2006, p. 8.

3 Gantz, J. ; Reinsel D. ; Chute C. et al, The expanding digital universe. A forecast of worldwide information growth through 2010. *IDC White Paper*, March 2007, pp. 1-5, <http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf>, (login June 5, 2008).

4 Noruzi, A. Folksonomies: why do we need controlled vocabularies? In: Noruzi, Alireza (ed.), *Webology*, Vol. 4, Nr. 2, June 2007, <http://www.webology.ir/2007/v4n2/editorial12.html> (login March 20, 2008).

5 Vander Wal, T. Folksonomy definition and Wikipedia. In: *vanderwal.net*, personal weblog November 2, 2005, <http://www.vanderwal.net/random/entrysel.php?blog=1750> (login May 11, 2008).

6 Sturtz, D. *Communal categorization: the folksonomy*. Unpublished course paper, Philadelphia: Drexel University, December 16, 2004, p. 1 <http://davidsturtz.com/drexel/ 622/sturtz-folksonomy.pdf> (login May 10, 2008).

7 <http://delicious.com/> (login February 15, 2009).

8 <http://www.flickr.com> (login February 15, 2009).

9 Van Damme C. Van folksonomieën naar ontologieën. Bladen voor documentatie, March 2008, Vol. 62, n°1, p. 12-17.

10 Spalding, T. Presentation during the panel *Creating the future of the Catalog and Cataloging*, ALA Annual Conference, Anaheim, June 29, 2008; <http://www.librarything.com/thingology/2008/07/future-of-cataloging.php> (login July 5, 2008).

11 LibraryThing concepts. In: *LibraryThing.com*, <http://www.librarything.com/concepts# what> (login April 27 2008).

12 <http://www.librarything.com/tagcloud.php> (login February 15, 2009),
<http://www.librarything.com/authorcloud.php> (login February 15, 2009).

13 Blachly, A. *LibraryThing Press Information*. <http://www.librarything.com/press/> (login April 27, 2008).

14 Rutkoff, A. Social networking for bookworms. *The Wall Street Journal Online*, 27 June 2006, <http://online.wsj.com/public/article/SB115109622468789252-i8U6LIHU7Chfgbx G1oZ_iunOIWE_20060727.html> (login July 6 2008).

15 <http://www.librarything.com/groups> (login February 15, 2009).

16 <http://www.librarything.com/groups/iseedeadpeoplesbooks> (login February 15, 2009).

17 In order to avoid too much overlap I have left out those books that were from the same series. The top 5 books for instance are all from the *Harry Potter* series.

18 <http://www.librarything.com/groups/librarianswholibrar> (login February 15, 2009).

19 <http://www.librarything.com/z_books.php> (login February 15, 2009).

[20] Golder, S.; Huberman, B. *The structure of collaborative tagging systems*. Information Dynamics Labs, Hewlett-Packard, 2005, p. 5, <http://arxiv.org/ftp/cs/papers/0508/ 0508082.pdf> (February 13, 2008); subsequent citations concerning the categories come from the same paper.

[21] Sen, S.; Lam, S.; Al Mamunur, R.; Cosley, D.; Frankowski, D.; Osterhouse, J.; Harper, F.M.; Riedl, J. Tagging, communities, vocabulary, evolution. *Proceedings of CSCW'06, November 4-8, 2006, Banff, Alberta Canada*, <http://www-users.cs.umn.edu/~cosley/ research/papers/sen-cscw2006.pdf> (login April 1, 2008).

[22] Halpin, H.; Robu, V.; Shepherd, H. The complex dynamics of collaborative tagging. *Proceedings of the 16 International World Wide Web Conference. Banff, Canada*, 2007, p. 212, <http://www2007.org/papers/paper635.pdf> (login July 17, 2008).

[23] Weber, J. *Folksonomy and controlled vocabulary in LibraryThing.* Unpublished Final Project, University of Pittsburgh, 2006, p. 5-6; <http://dystmesis.net/2006/11/17/tags-and-subject-headings/> (login March 16 2008).

[24] Blachly, A. *Tagging meets Subject Headings. Thing-ology Blog*, May 14 2006, <http://www.librarything.com/thingology/2006_05_01_archive.php> (login July 7 2008).

[25] Spalding, T. Presentation during the panel *Creating the future of the Catalog and Cataloging*, ALA Annual Conference, Anaheim, June 29, 2008.

[26] FAQ, steve: the museum social tagging project, <http://steve.museum/index.php?option=com_content&task=blogsection&id=6&Itemid=15> (login July 21 2008).

[27] Trant, J. More steve... tagger prototype preliminary analysis. In: conference.archimuse. com, 16 October 2006, <http://conference.archimuse.com/blog/jtrant/ more_steve_tagger_prototype_preliminary_analysis> (login July 21 2008);
Trant, J. Social classification and folksonomy in art museums: early data from the steve.museum tagger prototype. A paper for the *ASIST-CR Social Classification Workshop*, November 4, 2006, Draft, October 10, 2006, pp. 16-21 <http://www.archimuse.com/papers/asist-CR-steve-0611.pdf>, (login July 21, 2008).

[28] Vanderwal, T. Folksonomy provides 70 percent more terms than taxonomy. *Personal InfoCloud*, June 12, 2007, <http://personalinfocloud.com/2007/06/folksonomy_prov.html> (login: 15 July 2008).

[29] Morville, P. *Ambient findability*. Sebastopol: O'Reilly Media Inc., 2005, html edition, Section 6.2

[30] Lambe, P. *Organising knowledge: taxonomies, knowledge and organisational effectiveness*. Oxford: Chandos Publishing Limited, 2007, 253-255.

[31] <http://www.librarything.com/forlibraries/> (login February 15, 2009).