Analysis of user generated metadata in the LibraryThing folksonomy

Vincent Sterken

Master Thesis

Master in Business Information and Service Management

Promotor: Prof. Dr. E. Van Dijck

Academic Year 2007-2008

Table of Contents

Acknowledgements

I would like to thank:

Prof. Dr. E. Van Dijck for supervising my thesis.

Ms. Céline Van Damme for the conceptual and technological help.

Mr. Jan Wera for the technological support.

Ms. Katleen Sledz for proofreading.

Mr. Tim Spalding for providing me the data without which I would not have been able to write this thesis.

Introduction

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 has changed this. Out of a need for organizing information on the web a grassroots classification 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. 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. URL's. Quickly, however other uses have seen the light. Tags are being used for academic papers, life goals, movies, and books, etc.

Up until now, most of the research concerning folksonomies has analyzed social tagging in terms of information retrieval of content that is directly available on the web, i.e. web pages, digital photos, articles. Tags have not been analyzed within the context of bibliographic data in a system that does not allow instant access to the annotated content.

One of the services using a folksonomy is the social cataloging site LibraryThing. LibraryThing lets the user catalog his personal collection of books. Catalogs can also be browsed by other members. Individual books can be discovered by searching or browsing tags. LibraryThing claims to be the largest virtual library in the world. This makes it interesting to see how well the site fares in terms of subject analysis. As a side question, it will be examined if a significant difference can be noted in tagging behavior between professional indexers and non-professionals.

In the first chapter the existing methods for categorization and classification will be discussed. To understand what folksonomies are and in what way they differ from other classification efforts, it is necessary to know what other methods exist. The aspects of metadata will be treated, followed by an overview of the different ways of organizing information.

In the second chapter folksonomies are treated in more detail. An explanation is given of what a folksonomy entails and what its possible effects can be. Further, the advantages and disadvantages inherent in folksonomic systems are regarded. Although the folksonomy has not known a very long existence, some serious research has been devoted to it the last few years. At the end of the chapter a broad synopsis is given.

The analysis of the LibraryThing data follows in the third chapter. The first part consists of an explanation of the website. The different aspects of LibraryThing, together with the underlying philosophy, are discussed. Subsequently, the functionality is treated, with an emphasis on search and retrieval of content. The data itself will first be analyzed on the functions that tags can perform. Secondly a comparison will be made with the descriptors traditionally assigned to bibliographic records, i.e. subject headings.

1. Classification and its contents

1. The Age of Infoglut

Since the Second World War the amount of information produced has expanded exponentially. Inventions such as typewriters, microfilm, 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.[i] 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 estimated that on the (visible) World Wide Web[ii] 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).[1] 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[2] of information.[iii]

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

[i] An exabyte is 1 billion gigabytes, or 10^{18} bytes.

[ii] As opposed to the deep web (or invisible web), "a vast repository of underlying content, such as documents in online databases, that general-purpose web crawlers cannot reach. Both qualitative and quantitative in difference, deep web content is estimated at 500 times that of the surface web, yet has remained mostly untapped due to the limitations of traditional search engines." British Library's strategy 2005-2008 glossary. In: *The British Library, the World's Knowledge*, http://www.bl.uk/aboutus/stratpolprog/redeflib/glossary/ 12 August 2008

[iii] There will only be storage for 600 000 petabytes. This will create problems concerning long term availability and readability. We will not go into the discussion concerning digital preservation since that would lead us too far.

and retrieval. The information science community has several well established tools which it can use, such as classification schemes and thesauri. The traditional systems alone will not be able to keep up with the fast pace of technological evolution. Relatively recently, the ICT world has developed (and is developing) new methods such as ontologies and the semantic web. The youngest member in the family stems from a grassroots movement of collective categorization, i.e. folksonomies.

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 not a perfect solution, but it is a practical one. Information that would otherwise remain hidden, might become accessible thanks to the incremental nature of these systems. In order to understand what a folksonomy might add to our categorization efforts, it is best to see what its place is in the existing tradition. Therefore, in this chapter an overview will be given of the different methods that are currently available to us.

2. Classification and categorization

"The history of classification began with the establishment of the first library at the port of Alexandria in 285 B.C. Ptolemy I (Ptolemaios Soter) was persuaded by Demetrios Phalereus to collect copies of all known books to the library of Alexandria. With a growing set of resources in the library, books and scrolls were kept in piles or pits in order to group like materials together."[3] Although this statement is not exactly true[i], it does bring the point across that we have been trying for quite some time now to find the best way of organizing information resources.

The object of classification and categorization is providing a way of finding and retrieving information. There are several ways of organizing this.

[i] The earliest archives date back to the Mesopotamian era. Even if you don't count the order within these tablets of clay as a proper classification, you mustn't forget the Egyptian Pharaohs' demand for a well organized bureaucracy, and the ensuing need for proper classification methods. Jannsens, G. & Put, E. Geschiedenis, principles en terminologie van de archivistiek. Onuitgegeven syllabus, Vrije Universiteit Brussel, 2005-2006, pp. 19-21

2.1 Metadata and object-based classification

Metadata are at the core of all systems intended for search and retrieval of documents (digital or otherwise).[4] The term is generally described as 'data about data'.[i] The term comes from the field of computer science, where it is commonly used "in the sense of the information necessary to make computer files useful to humans."[5] The Dublin Core Metadata Initiative adds a functional element to the definition by calling it "structured data about data", which "includes data associated with either an information system or an information object for purposes of description, administration, legal requirements, technical functionality, use and usage, and preservation."[6] In this sense the term has outgrown its original emphasis on digital objects and has been broadened to encompass any kind of standardized descriptive information about resources (e.g. library catalogues, archival finding aids, museum documentation).[7]

Although metadata has been in use outside of the digital realm for a long time, *albeit* under a different name, their function as a finding aid is becoming more and more important in this age of *infoglut* to information specialists and 'civilians' alike. Metadata have many purposes, but the most interesting to us (in the light of this dissertation) are the properties which enable them to find documents[ii].

"In the context of digital resources, there exists a wide variety of metadata formats. Viewed on a continuum of increasing complexity, these range from the basic records used by robot-based Internet search services, through relatively simple formats like the Dublin Core Metadata Element Set

---

[i] *Meta* is actually the Latin version of the Greek word *μετά*, meaning 'after', 'beyond' or 'with'. In English it is used as a prefix "in order to indicate a concept which is an abstraction from another concept, used to complete or add to the latter." Meta. In: *Wikipedia,* http://en.wikipedia.org/wiki/Meta; or in other words: "something of a higher or second-order kind." Meta-. In: *Compact Oxford English Dictionary*, http://www.askoxford.com/concise_oed/meta?view=uk, 14 March 2008

[ii] The molecular units that contain data (which make up information) are sometimes referred to as documents, at other times as objects. Both are valid descriptions. The use depends on the background of the user. I will regard both terms as equivalent to each other unless stated otherwise, although a quick definition of the word document might be in order. A document can be described as "recorded information … which can be treated as a unit" regardless of its format. IDA, Model Requirements for the management of electronic records. MoReq Specification. Brussel: CECA-CEE-CEEA, 2001, p. 7

(DCMES)[i] and the more detailed Text Encoding Initiative (TEI)[ii] header and MARC formats[iii], to highly specific formats like the FGDC Content Standard for Digital Geospatial Metadata[iv], the Encoded Archival Description (EAD)[v] and the Data Documentation Initiative (DDI) Codebook[vi]."[8]

The best known standards in the field of archival and library sciences are Dublin Core, MARC21 and EAD.

The *Dublin Core Metadata Element Set* is a list of metadata elements agreed upon during the OCLC/NCSA Metadata workshop in March 1995 in Dublin, Ohio.[9] It was developed into a standard, which was accepted by the International Organization for Standardization (ISO) in 2003.[10] The workshop united international groups of different professional backgrounds. The purpose was the development of a set of metadata elements which could be used by professionals from the archival, library and computer sciences, text encoding, the museum community, and other related fields. DCMES is positioned as an information resource description.[11] "However, importantly it also aims to provide a basis for semantic interoperability between other … formats" and for "resource-embedded description, initially with HTML documents."[12] This cooperation gave birth to the Dublin Core Metadata Initiative (DCMI)[vii], which has as mission "promoting the widespread adoption of interoperable metadata standards and developing specialized metadata vocabularies for describing resources that enable more intelligent information discovery systems."[13] The standard consists of 15 elements[viii] which should be simple to create and maintain, encourage commonly understood semantics and be extensible. After the creation of the original elements, the set has been refined from the Simple Dublin Core with an

---

[i] http://dublincore.org/documents/dces/
[ii] http://www.tei-c.org/index.xml
[iii] http://www.loc.gov/marc/
[iv] http://www.fgdc.gov/metadata/csdgm/
[v] http://www.loc.gov/ead/
[vi] http://www.ddialliance.org/codebook/index.html
[vii] http://dublincore.org/
[viii] Title-Creator-Subject-Description-Publisher-Contributor-Date-Type-Format-Identifier-Source-Language-Relation-Coverage-Rights

additional Qualified Dublin Core. This "includes three additional elements[i] …, as well as a group of element refinements (also called qualifiers) that refine the semantics of the elements in ways that may be useful in resource discovery."[14]

The *MAchine-Readable Cataloging* (MARC) standards were developed in order to make it possible for computers to read and interpret data in a cataloging record and allow the exchange of bibliographic, authority, holdings, classification, and community information data.[15] A cataloging record is a bibliographic record, i.e. the information you can find on a catalog card, which usually consists of a description of the item, a main entry, subject headings and the classification number.[16] A MARC record consists of three elements: the record structure (which implements the structure of ISO 2709[ii]), the content designation[iii], and the data content of the record (based on standards such as Library of Congress Subject Headings[iv]).[17] MARC 21 is the recombination of the US and Canadian versions of the formats, with the aim of adapting the standard to the needs of the 21st century and make it better suited for international exchange.[18] It has formats for five types: authority data, bibliographic data, classification data, community information, holdings data.[19]

The *Encoded Archival Description* (EAD) DTD[v] started out as a project instigated by the University of California (Berkeley) Library in 1993. The aim was the development of a non-proprietary encoding standard for machine readable finding aids created by archives, libraries, museums, and manuscript repositories, which would provide more information than the traditional MARC standards.[20] In order to determine the right techniques

---

[i] Audience-Provenance-RightsHolder

[ii] ISO 2709:1996 - Format for Bibliographic Information Interchange on Magnetic Tape

[iii] "The goal of content designation is to identify and characterize the data elements that comprise a MARC record with sufficient precision to support manipulation of the data for a variety of functions." MARBI, The MARC 21 Formats: background and principles. November 1996, http://www.loc.gov/marc/96principl.html#four, 19 March 2008

[iv] http://authorities.loc.gov/

[v] "The purpose of a DTD (Document Type Definition) is to define the legal building blocks of an XML document. It defines the document structure with a list of legal elements and attributes." W3Schools, DTD Tutorial. http://www.w3schools.com/dtd/default.asp, 19 March 2008

and technology a number of conditions had to be met. At the least it had to be possible to present the large amounts of descriptive information in inventories (including the relations between descriptions), to maintain the hierarchical relations between the different levels in the digital form, to show descriptions that are relevant on several levels, to navigate easily within the hierarchical structure, to be able to index specific elements in order to enhance later retrieval.[21] The best candidate turned out to be the Standard Generalized Markup Language (SGML). Later the eXtensible Markup Language (XML) was incorporated as well. As standard for the descriptive elements ISAD(G) was adopted.[22] The International Standard for Archival Description (General) (ISAD)[23] was developed by an *ad hoc* commission of the International Council on Archives (ICA)[i] in order to standardize (and thus make it internationally exchangeable) archival inventories. ISAD is a model (not a rulebook) which provides guidance for the preparation of archival (multilevel) descriptions.[24]

The examples above illustrate the need for 'good' (as in useful for its intended purpose) metadata in order to be efficient as a finding aid and tool for classification. There are different ways of putting them to use. Each has its own advantages and disadvantages.

2.2 Subject-based classification

Subject-based classification organizes documents based on the content (subject) they're about. Different forms will be discussed below. The relation with metadata is that metadata properties use this type of classification. The difference with the schemes described above, is that here the subjects are being described instead of the objects (documents).[25]

2.2.1 Language

Natural language is the language that people use in everyday life when speaking and writing. When used for information representation and retrieval there are no limits on or, definitions of, vocabulary, syntax, semantics, and relations between terms. Terms can be derived by taking

[i] http://www.ica.org/

them from titles, topic sentences and other important components or parts of a document. Another way is extracting words or phrases directly from people's queries.[26] Recently, massive collective indexing efforts, i.e. folksonomies, can be leveraged for information representation and retrieval.

A *controlled vocabulary* is a finite list of words (terms), which can be used for classification. It can be considered to be a type of metadata functioning as a subset of natural languages.[27] The purpose is to avoid authors designating terms to documents which make it difficult to retrieve them afterwards. Terms can be deemed useless because they are meaningless to everyone except the author, or because they are too broad or narrow. Using controlled vocabularies also eliminates the possibility of misspellings and variations in related and equivalent terms.[28] A distinction can be made between several approaches based on the time of application, meaning whether they are intended for pre- or post-coordination. A post-coordinated language allows users to "coordinate terms at the time of representation and retrieval", while a pre-coordinated language "combines terms before they are used for representation and retrieval."[29] Controlled vocabularies can be divided into classification schemes, thesauri and subject heading lists:

- *Classification scheme*: Classification schemes consist of alphanumeric terms. They are intended mainly for pre-coordination. Such a scheme has as a foundation an artificial framework of knowledge.[30] The most known are the Dewey Decimal Classification (DDC), the Universal Decimal Classification (UDC), and the Library of Congress Classification (LCC). They all have three main components: (1) a set of subjects (classes and further subclasses), (2) notational symbols, (3) an index. The main classes (with their subclasses) are arranged by notations (eg. 100 is philosophy, while 100.1 is Western Philosophy). An alphabetical index is added to enhance the search capabilities.[31] The different systems in use can be divided into three major groups: enumerative, analytico-synthetic, and faceted classification:
    - In an *enumerative classification scheme* all possible classes are enumerated according to certain characteristics, based on a top-

down approach. The basic tenet is that all possible subjects and topics are listed with a predefined class number. The advantage is that the classifier just has to follow the outlined structure. Unfortunately, this makes the scheme very rigid since it is not possible to create new class numbers.

- In an *analytico-synthetic classification scheme* the subject of a document is divided into its constituent elements, for each of which notations will be found using the scheme. These notations will then be joined to prepare the final class number. This way of working increases the flexibility of the system and reduces the scheme's size.
- A *faceted classification scheme* lists the various facets of every subject or main class, together with a set of rules for constructing class numbers through facet analysis. The basic idea of the original scheme, invented by S.R. Ranganathan, was that any component, or facet, of a subject can fit into a number of fundamental categories. In this case: personality, matter, energy, space and time.[32]

- *Thesaurus*: Thesauri are post-coordinated controlled vocabularies, containing a set of terms, the relations between them and often also how they are to be applied. Preferred and non-preferred terms are defined, as well as the semantic relations between them. The relationships between terms are specified by standard notations. The preferred descriptors are defined by Scope Notes (SN), while USE and UF (Used For) indicate where they should be used.[33] The hierarchical relations are made explicit by denoting whether the descriptor is a BF (Broader Term) or a NF (Narrower Term) of another. Associative relations are dubbed RT (Related Terms).[34] Thesauri are used mostly because of their specificity, flexibility and ability to handle complex concepts.[35] There are several standard thesauri, like ISO 2788[i] and

[i] ISO 2788:1986 - Documentation -- Guidelines for the establishment and development of monolingual thesauri, for excerpts see: http://www.collectionscanada.gc.ca/iso/tc46sc9/standard/2788e.htm

5964[i], the Getty Art and Architecture Thesaurus[ii], and the UK Archival Thesaurus (UKAT)[iii].

- *Subject heading list*: While thesauri are intended for post-coordination indexing (and are thus a relatively flexible tool), subject heading lists are designed both for pre- and post-coordination and consist of alphabetical lists of terms, with associated cross-references and notes. The terms are hierarchically organized, where each 'child' (lower term) has a semantic relation with its 'parent'.[36] Commonly used notations are "See" (for preferred terms) and "See also" (indicating the hierarchical and associative relationships between headings).[37] Two widely used models are the Sears' List of Subject Headings and the Library of Congress Subject Headings.

In order to complete the story we must touch on the term *taxonomy.* The origins of the word taxonomy[iv] can be traced back to an eighteenth century Swedish physician called Carolus Linnaeus. He devised a hierarchical system for classifying natural objects.[38] Since then, its meaning has been co-opted by many different research areas. All definitions have one thing in common: they are about the hierarchical organization of knowledge, divided into *taxa* (classes) between which the relations are clearly defined. Within the context of information architecture, a taxonomy can be regarded as a "subject-based classification that arranges the terms in the controlled vocabulary into a hierarchy."[39] Terms are organized in parent-child relationships. This approach allows users to easily browse categories and subcategories in order to find the appropriate document.[40] Simply put, taxonomy is the information scientist's word for hierarchical classification.

2.2.2 Ontology

The word ontology has taken on many definitions. Its origins can be found in the philosophy of the Ancient Greeks, from where it has taken the meaning

[i] ISO 5964:1985 - Documentation -- Guidelines for the establishment and development of multilingual thesauri
[ii] http://www.getty.edu/research/conducting_research/vocabularies/aat/
[iii] http://www.ukat.org.uk/
[iv] "The word comes from the Greek *τάξις*, *taxis*, 'order' + *νόμος*, *nomos*, 'law' or 'science'." Taxonomy. In: *Wikipedia*, http://en.wikipedia.org/wiki/Taxonomy, 11 April 2008

of studying the essence of "Being".[41] Generally speaking, existence itself is the object of inquiry, including a classification of the elements of the existing world based on their essence. In computer and information science it is described as a conceptual schema, with a corresponding vocabulary, used to represent a particular domain of knowledge.[42] It is a data model that represents a set of concepts within a domain and the relationships between those concepts. Although an ontology is similar to a taxonomy, it surpasses the latter in richness of knowledge representation. The difference can be found in the rich relationships between concepts. "Taxonomies in effect are simplified ontologies. Where taxonomies generally classify categories in "broader" or "narrower" terms, ontologies can include more descriptive classifiers such as 'located in' or 'part of'."[43]

2.2.3 Topic Maps

Topic Maps provide, as an accepted ISO standard,[i] a standardized notation for the representation of information about the structure of information resources.[44] The basic elements of a topic map are topics, occurrences, and associations. A topic is a resource that reifies a real-world subject. A subject can be anything, going from abstract concepts to a specific document section. "Exactly what one chooses to regard as topics in any particular application will vary according to the needs of the application, the nature of the information, and the uses to which the topic map will be put: In a *thesaurus*, topics would represent terms, meanings, and domains."[45] To any given topic three characteristics can be added: names, its associations and its occurrences (resources). Names as such are strictly speaking not necessary, but are useful to the people using the topic maps. Associations (relationships) indicate how topics relate to one another, while the occurrences point to the relevant information resources. These characteristics aren't considered to be universal in nature. Their assignment of is situated within a certain scope (context) where they are regarded as valid. Since multiple topics can represent a single subject, the identity of

[i] ISO/IEC 13250:2003 - Information Technology -- SGML Applications -- Topic Maps; ISO/IEC 13250-2:2006 - Information technology -- Topic Maps -- Part 2: Data model; ISO/IEC 13250-3:2007 - Information technology -- Topic Maps -- Part 3: XML syntax

said subject can be defined through resources called subject indicators.[46] The model makes a clear distinction between the domain model, expressed as topics and its associations, and the indexed resources, expressed as occurrences. This way the topic map can function as a high-level overview of a knowledge domain, which can be adapted dynamically and still be easily used for search and retrieval by experts and non-experts alike.[47]

2.3 Folksonomy

A *folksonomy* is a user-generated taxonomy with which web content can be categorized and retrieved through the use of open-ended labels called tags.[48] It has been dubbed grassroots classification, collaborative tagging, ethnoclassification, folk classification, open tagging, social classification, faceted hierarchy, etc.[49] 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 retrival. 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 as. The people are not so much categorizing as providing a means to connect items and to provide their meaning in their own understanding."[50] Or more simply put, "folksonomies are taxonomies created by users who add tags to things."[51] The defining characteristics of a folksonomy are its bottom-up structure, its lack of hierarchical structure, and the social context in which it is used.[52] The most common examples are the social bookmarking site del.icio.us[i] and the photo sharing site Flickr.[ii] 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[iii] and pictures,[iv] such as music,[i] videos,[ii] books,[iii] academic papers,[iv] events,[v]

[i] http://del.icio.us/. For ease of use, the simpler way of writing "Delicious" instead of "del.icio.us" will be used in the rest of the text.
[ii] http://www.flickr.com/
[iii] http://myweb.yahoo.com, http://www.furl.net/
[iv] http://www.panoramio.com/

blogs,[vi] even life goals.[vii] The primary objective is refindability of saved resources by the user himself. Because of the fact that other users can see (and browse through) the resources that have been saved and can search the saved tags, a communal aspect is inherent to folksonomic systems. The difference with social networking sites is that the emphasis here is on organizing data instead of on creating and maintaining relationships. In the next chapter a more detailed explanation will be given.

2.4 Semantic Web

The *Semantic Web*, as envisaged by the Tim Berners-Lee,[viii] is an extension of the current Web "in which information is given well-defined meaning, better enabling computers and people to work in cooperation."[53] The two basic components are common formats for integration and combination of data drawn from different resources, and a language for recording how the data relates to real world objects. Thus, it is aimed at providing "a common framework that allows data to be shared and reused across application, enterprise, and community boundaries", based on the Resource Description Framework (RDF).[54] RDF is a standard way of describing relationships between topics, in which metadata is expressed in triples (two terms connected by a third, e.g. university *is a type of* school).[55] At its core it comprises a set of design principles, collaborative working groups, and a set of technologies, including RDF and the Web Ontology Language (OWL[ix]).[56] Although some sites[x] have already made information available in Semantic

---

[i] http://last.fm
[ii] http://youtube.com
[iii] http://www.librarything.com/, http://www.goodreads.com/, http://www.shelfari.com/, PennTags (http://tags.library.upenn.edu/)
[iv] http://www.citeulike.org/, http://www.connotea.org/, http://www.bibsonomy.org/
[v] http://upcoming.org
[vi] http://technorati.com,
[vii] http://www.43things.com/
[viii] Tim Berners-Lee is the inventor of the WWW and HTML, and director of the W3C Consortium. See: http://www.w3.org/People/Berners-Lee/
[ix] "The OWL Web Ontology Language is intended to provide a language that can be used to describe the classes and relations between them that are inherent in Web documents and applications" and is therefore "a language for defining and instantiating Web ontologies." Smith, M.; Welty C.; McGuinness, D. (eds.) OWL Web Ontology Guide. In: *W3C recommendation*, 10 February 2004, http://www.w3.org/TR/owl-guide/, 12 May 2008
[x] E.g. The Neurocommons, http://sciencecommons.org/projects/data/

Web format, it is still hard to find large-scale applications outside of the academic world.[57]

2. Folksonomies

1. Introduction

1.1 YACS?[i]

In the previous chapter a folksonomy was defined as a bottom-up taxonomy, to which metadata is added by users in the form of tags. This is not completely true. Although the word folksonomy is a contraction of folk and taxonomy, it is nothing like a traditional taxonomy. As we have seen, hierarchical classification systems are constructed (and maintained) by professionals who classify objects based on a set of rules. Folksonomies on the other hand work by grace of the adage "anything goes". What is truly remarkable is that both ways of organizing human knowledge work.

Merriam-Websters defines classification as a "systematic arrangement in groups or categories according to established criteria."[ii] There are many different ways to establish the criteria deemed fit to represent a category. The ideal is to create a system which allows users to categorize objects in an objective way. Unfortunately every system is prone to errors, due to the system itself or due to the fact that different people interpret the rules differently. During a practical course on records management for my studies in archival science, Prof. Dr. Frank Scheelings gave the students several documents to classify. The class was split up in groups and each group received the same set of documents but a different method to work with (UDC, DCC, …). Every group experienced a number of problems caused by the ambiguity of the archival records. Because, for instance, a letter can be about different subjects it isn't always easy to give it a place in a pre-defined order. When trying to find the documents again as they were classified by the other groups proved just as difficult. You might argue that this was because of our lack of expertise, but even experienced cataloguers exert variations in the assignment of categories.[1] Whichever way you put it, the

[i] Yet Another Classification System
[ii] Merriam-Webster's Online Dictionary, http://www.merriam-webster.com/dictionary/classification, 14 May 2008

example illustrates my point. The intention of any classification model is to find (again) what is hidden in the accumulated mass of books in a library or in the accumulated bits of the inter- and/or intranet. Therefore, these problems should not be ignored. Especially since the people searching for information are usually not experts in cataloguing.

The question is: who defines the categories and to what end?[i] In his classic article *Ontology is overrated*[2] Clay Shirky states that applying concepts that might work fine in the physical world to the digital realm is a mistake. The classification models we know look the way they do because of historical and cultural reasons combined with the fact that we take physical constraints into account. A letter, a book, a videotape, etc. can only be in one place at a time. The information it contains, however, can be about many different things. Nevertheless, the physical object needs to be put in one particular place. So, it has to be 'declared' to be about one thing in order to be filed in its 'proper' place on the shelf. In the digital world this isn't true anymore. A document can be about many things *and* can be in many places at the same time. David Weinberger takes this reasoning a step further in his book *Everything is miscellaneous. The power of the new digital disorder*, by introducing the concept of the three orders of order. "In the first order of order, we organize things themselves – we put silverware into drawers, books on shelves, photos into albums." Library science has provided us with "a prototypical example of the second order of order: a card catalog … The catalog separates information about the first order objects from the objects themselves. … A code on this second-order object, the catalog card, points to the physical place where the first-order [object] is stored … But now we have bits. Content is digitized into bits, and the information about that content is bits as well. This is the third order of order … [It] removes the limitations we've assumed were inevitable in how we organize information."[3] According to Weinberger, classification efforts can be traced back to the Ancient Greeks' idea that there is a natural order of things that exist; and consequently this order should be applied to our understanding of the

[i] OK, that are actually two questions, but I wasn't expecting the Spanish Inquisition.

world.[i] Thus, many have tried to create a natural order of knowledge. Plato said that reality has natural 'joints'. Hence, a skilled thinker, like a skilled butcher, should be able to know where those 'joints' are.[4] One such attempt at cleaving nature at the joints is the Dewey Decimal Classification. Melvil Dewey adopted as a larger structure the Hegelian reversal of the order proposed by Sir Francis Bacon,[ii] combined with the assumption that the Western Christian culture is the pinnacle of truth. On top of that, he liked decimals to the point that he figured it was the most ideal way to organize libraries. So he created ten top level classes, each with ten divisions, with each having ten sections.[5] The problem here of course is that all knowledge needs to be fitted into sets of ten, while sometimes you need eleven or nine is enough.[iii] Dewey's and other classification systems are also heavily indebted to Aristotle, who said that a category is a definition, a principle, explaining why some things fit into it and others don't. So, you cluster like and split unlike things based on a set of universal principles. Later this idea was transformed into a tree-structure in which everything has only one place.[6] Weinberger says that this Aristotelian logic has stayed with us until this day.

Another reason why a scheme like DCC looks the way it does, is because it had to be thought out on paper; more importantly, by writing each concept on a slip of paper. These slips of paper were than laid out in a certain order. It's obvious that two slips of paper can't be in the same spot at the same time; else you won't be able to see what you are doing. Which brings us back to the original point made by Shirky: the second order of things is ruled by the constraints of the physical world. "The musculature of the Library of Congress categorization scheme looks like it's about concepts. It is organized

---

[i] Shirky has dubbed this *ontological classification*, which is the organization of "a set of entities into groups, based on their essences and possible relations." Shirky, C. Ontology is Overrated: Categories, Links, and Tags. In: *Clay Shirky's Writings About the Internet*, personal weblog, 2005, http://shirky.com/writings/ontology_overrated.html, 22 February 2008

[ii] Bacon's order: history, poesy, and philosophy.

[iii] A second problem is the cultural bias, which is very evident when we take a look at the 200 category, religion: 210 Natural theology - 220 Bible - 230 Christian theology - 240 Christian moral & devotional theology - 250 Christian orders & local church - 260 Christian social theology - 270 Christian church history - 280 Christian sects & denominations - 290 Other religions. As Shirky says: "How much is this not the categorization you want in the 21st century?" Shirky, C. *op. cit.*

into non-overlapping categories that get more detailed at lower and lower levels … But … the supporting structure around which the system is really built, is designed to minimize seek time on shelves. The essence of a book isn't the ideas it contains. The essence of a book is 'book'. Thinking that library catalogs exist to organize concepts confuses the container for the thing contained."[7] The digital world has made it possible to rethink the way we categorize. Only now, since the advent of the internet paradigm, are people seriously contemplating what these possibilities might be in a real, practical sense.

Within this concept folksonomies are considered to represent a particular form of the third order of order. Instead of browsing through a pre-defined set of categories based on the worldview of someone else, folksonomic systems can rearrange themselves in a way which is adapted to the preferences of the user. At the core of a folksonomy lies the usage of *tags*. 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 others tags that are displayed. In figure 1 you see the tagcloud for the most popular tags in the Delicious network as a whole.

Figure 1: Delicious "Popular" tagcloud on 17 May 2008

It is possible to vary in depth. Figure 1 is an example of the most popular tags at the level of the resource. The same is also possible, and generally available, at the level of the user. A personal tagcloud helps the user to navigate easily through the saved resources to retrieve the specific content she is looking for at any given time. Every system that employs a folksonomy has different ways of expressing this in the features made available. In the next chapter the peculiarities of the site upon which research was done, i.e. LibraryThing, will be explained in more detail.

Figure 2: LibraryThing tagcloud for the book *Everything is miscellaneous*.

In effect, the user is adding metadata to an object. Traditionally creation of metadata has been left to professionals, intermediaries. Librarians, archivists, or other people working in the sector of information science assign descriptors to objects based on standardized rules. Although professionally created metadata is considered to be of high quality, it is a very costly and time-consuming process. Further more, since the Second World War we have known an explosion of documents being produced, ushering in the era of infoglut as we know it today. As a result it has become extremely difficult to annotate every single piece of data being produced. The first attempts at solving this problem involved the creator of the document.[i] Within the archives profession research has shown however that most creators don't want to be bothered with the assignment of metadata. Even adding a few descriptors before saving a document is seen as an annoyance and a misappropriation of time. Therefore metadata is added automatically

[i] Based on the principles of the *records continuum*.

as much as possible.[8] For the authors of books and papers this argument is perhaps less valid, given the fact that they have a vested interest in the findability of their work. Whatever may be the case, author generated metadata helps solve a part of the problem of scalability. Adam Mathes states that both approaches, intermediary and author generated metadata, "share a basic problem: the intended and unintended eventual users of the information are disconnected from the process."[9] A third approach is user created metadata, which is of course what folksonomies provide. The users of the resources add information to them in the form of tags, thus creating another way of retrieving said resources and enhancing the scalability of the system. A very important difference with the first two approaches can be found in the social aspect. In most cases, a folksonomy doesn't stand on its own. The fact that it is possible to see the tags and the associated resources of other people leverages the power of the masses.[i] Nevertheless, users do not tag (directly) for the benefit of the community as such, but mainly for themselves. And that is why it works. Altruism is a beautiful ideal, but not one to encourage indexing efforts. Tagging is done to find one's own resources, yet by doing so a happy side effect is created because of the set-up of such systems. Taggers help the rest by helping themselves.

Is a folksonomy Yet Another Classification System? Not really. Classification requires a conscious effort to organize the world of knowledge in a coherent matter by cleaving nature at the right joints. Although in the above text the terms classification and categorization have been used as synonyms, a distinction can be made between both. Categorization reflects the Aristotelian division and clustering of data (and information and knowledge), called lumping and splitting in information science, while classification comprises of "a system of classes, ordered according to a predetermined set of principles and used to organize a set of entities."[10] Categorization tries to organize the world in categories, while classification tries to organize information in classes. Traditional classification is very rigorous and allows an object to be placed within one particular class. Categorization is more

[i] Which is way it has also been dubbed 'mob indexing'.

flexible and draws nonbinding associations between objects (based on the simple recognition of similarities across a set of objects). Elin Jacob has compared both systems based on six systemic properties:[11]

1. Process: the process of classification is done by analyzing the essential characteristics of an object, followed by an assignment to a certain class. The process of categorization is generally unsystematic but also more flexible because it doesn't need to rely on predetermined definitions. Similarity assessments can be made based on immediate context, personal goals, or individual experience.
2. Boundaries: in a classification system classes are clearly delineated by the definition the 'essence' of a particular class, making the classes mutually exclusive and non-overlapping. In a categorization system boundaries are fuzzy, mutable and potentially fluid. Membership to a category does not prohibit membership to another category.
3. Membership: membership to a class is strict, while a membership to a category may vary across time based on the combination of context dependent and independent information used to define membership.
4. Criteria for assignment: in a classification system criteria for assignment are governed by the principles that lie at the basis of the conceptual framework. The criteria for category assignment are potentially variable according to the context within a given category is used.[i]
5. Typicality: here the question is asked how representative a given member is of its class or category. Since class assignments rely on theoretical (and thus abstract) ideal properties, every member should be equally representative of its class. In a category no one object is said to be completely representative because of the fact that context matters.
6. Structure: a classification system is generally a hierarchical structure of well-defined, mutually exclusive, and nonoverlapping classes nested in a series of superordinate-subordinate … relationships. The

[i] E.g. a romantic book: are we talking about Coleridge or Jackie Collins?

structure of a classification system provides a powerful cognitive tool -- an external scaffolding --that minimizes the cognitive load on the individual by embedding information about reality through the organization of classes within the system. … In contrast, the structure of a categorization system consists of variable clusters of entities that may or may not be organized in a hierarchical structure."[12]

Given the analysis of Jacob, we can clearly regard folksonomies as a categorization system. Collaborative tagging can be seen as the creation of categories. These categories are of an implicit nature and have a closer relation to a general worldview with as many facets as there are people tagging. Tags are extremely context driven, allow resources to be in multiple categories at the same time, and lack any kind of hierarchy. This makes it a very flexible tool which, however imprecise at times, can aid the retrieval of relevant information.

David Weinberger pointed out that while tagging meets the criteria of categorization, it nevertheless *feels* different. Categorization feels like putting things in buckets (lumping and splitting), tagging feels like labeling them.[13] One reason why it has become so popular is precisely the feeling that you're not making a life-or-death decision: you're 'just' putting a label on the thing you don't want to lose. This line of reasoning is associated with the idea that there is a lower cognitive cost to tagging. Rashmi Sinha has made a cognitive analysis of the phenomenon. He explains the cognitive process behind that

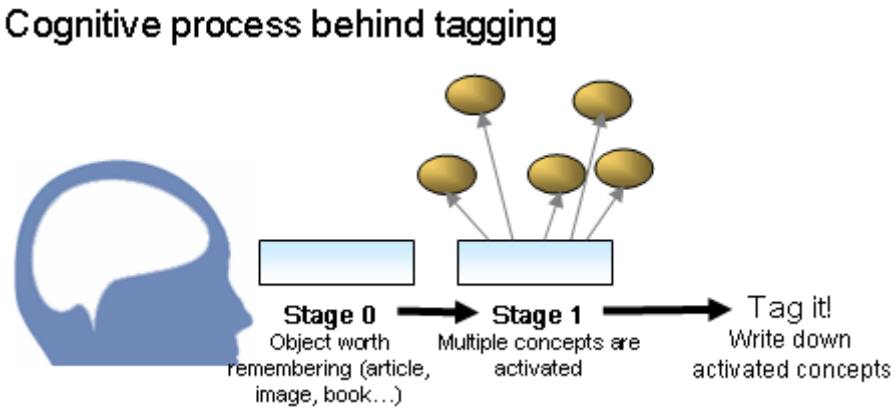


Figure 3: The two stage cognitive process of tagging. (Sinha, 2005)

gets activated when we tag and how this differs from the process of categorization. At the first stage a number of related concepts are activated. Since with tagging there is no filtering involved, the second stage consists of applying one or multiple tags to the resource in question.[14]

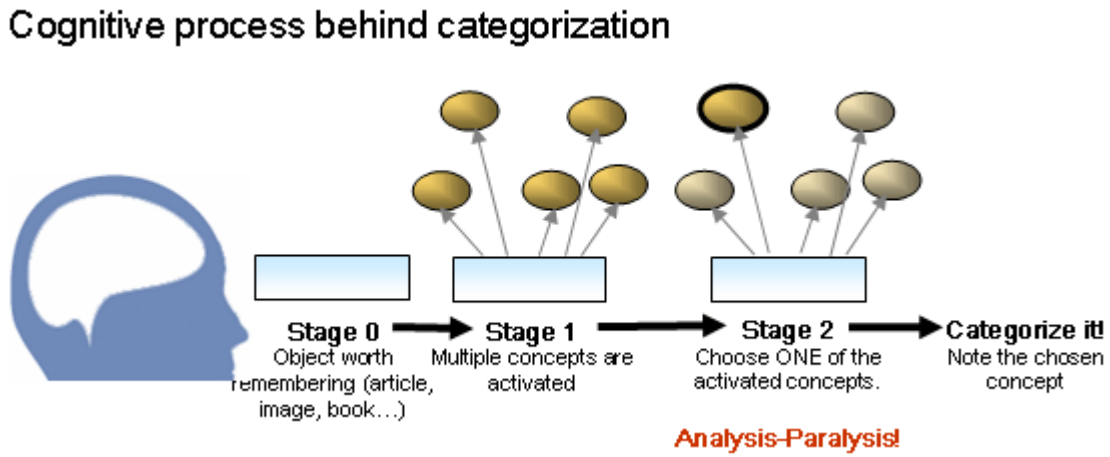


Figure 4: The three stage process of categorizing (Sinha, 2005)

Categorization on the other hand is more formal. After having selected candidate categories a conscious decision needs to be made. Cognitively, the process consists of computing the similarity between the present item and the possible categories. In our everyday life we make these kinds of decisions all the time. Sinha states that we in general it is sufficient to make local decisions. You see a dog, your brain 'knows' it's a mammal. As you encounter more dogs during your life the brain makes subcategories, making you an a bit more of an expert in the subject as you go along. In the digital realm we are less proficient in this kind of categorization, which may lead to what he calls "post-activation analysis paralysis." The object of digital categorization is to optimize future findability. If you don't maintain a balanced scheme it becomes difficult after a while to find your way in the nested tree-structure that you have created. Although it is possible to reorganize digital objects, this is fairly expensive in terms of the time it takes to do so. Think too much about this and you will develop a state of fear that you might make the wrong decision. Tagging on the other hand eliminates the need to make a decision. "The beauty of tagging is that it taps into an existing cognitive process without adding much cognitive cost. At the

cognitive level, people already make local, conceptual observations. Tagging decouples these conceptual observations from concerns about the overall categorical scheme. The challenge for tagging systems is to then do what the brain does - intelligent computation to make sense of these local observations, and an efficient, predictable way to ensure findability."[15]

Summarizing, we can say that folksonomies hold the promise of helping us find our way in the enormous amount of information available to us. Tagging provides us with an additional means of organizing the human body of knowledge. Sometimes the discussion takes on an OR this OR that direction. It should be AND AND. That's the beauty of the digital world. People can choose how they want to view something. So, search and retrieval can be facilitated by folksonomies AND taxonomies AND ontologies AND … well, you get my point.

Dilbert, September 9, 2007, http://dilbert.com/fast/2007-09-09/

1.2 Types[i]

According to Thomas Vanderwal there are two types of folksonomies: broad and narrow.[ii] A broad folkonomy, such as LibraryThing, allows for many layers of tagging, which develop patterns of consistency. In a narrow folksonomy, such as Flickr, only a few users supply the tags.[16]

A *broad folksonomy* lets many people tag the same object. Every person, except for the creator, can tag the object in his own way with his own vocabulary.

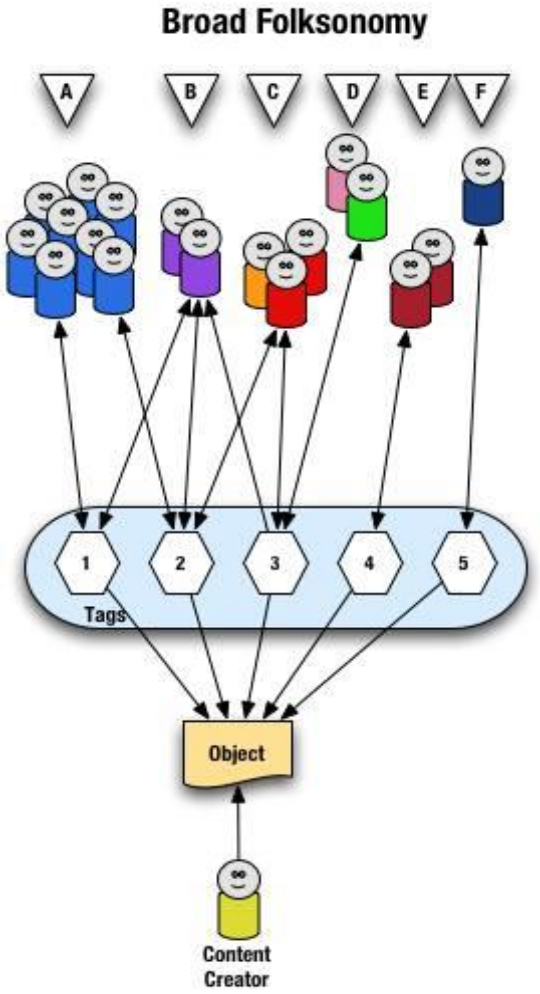


Figure5: Broad folksonomy (Vanderwal)

One person creates the object and makes it available to others. These other people then tag the object with their own terms. In Vanderwal's graphical representation, this is indicated by the arrows pointing to the 'diamonds'. The alphabet letters above the 'people' denote groups using the same vocabulary. The information itself can be found based on the tags, indicated by the arrows pointing back to the 'people'. Many people can and probably will tag an object in a different way. However, a larger amount of users tend to favor a particular (set of) tag(s), which results in what is called a power law curve, a continuously decreasing curve with a

[i] The following explanation is based in its entirety on Vanderwal's post on *Personal InfoCloud* (except where indicated). The images were taken from this post as well. Vanderwal, T. Explaining and showing broad and narrow folksonomies. In: *Personal InfoCloud*, February 21, 2005, http://www.personalinfocloud.com/2005/02/explaining_and_.html, 27 January 2008

[ii] He also added the concept of a personal folksonomy, while discussing Google's e-mail application, Gmail, where it is possible to freely tag one's own e-mail entries. This will not be discussed for this might lead us to far.

long tail at the end.[17] Power law distributions have been observed in economics and linguistics. Vilfredo Pareto,a nineteenth century Italian economist, noted that only a small percentage of people (20%) control most of the wealth (80%). The concept was translated into what is called the 80-20 rule or Pareto's Law.[18] Zipf's Law, named after the linguist George Zipf, tells us that word frequencies fall in a power law pattern. They contain a large number of high frequency words (I, of, the), a moderate amount of common words (book, cup), and a large number of low frequency words (peripatetic, hypognathous).[19] In broad folksonomies a similar distribution can be seen, by providing a means to see trends in how a broad range of people are tagging one object. The trend becomes visible through the spike at the left hand side of the curve. At the right end of the curve, the long tail, we find "a small minority of people who call the object by a term, but those people tagging this object would allow others with a similar vocabulary mindset to find the object, even if they do not use the terms used by the masses over at the left end of the curve."[20]

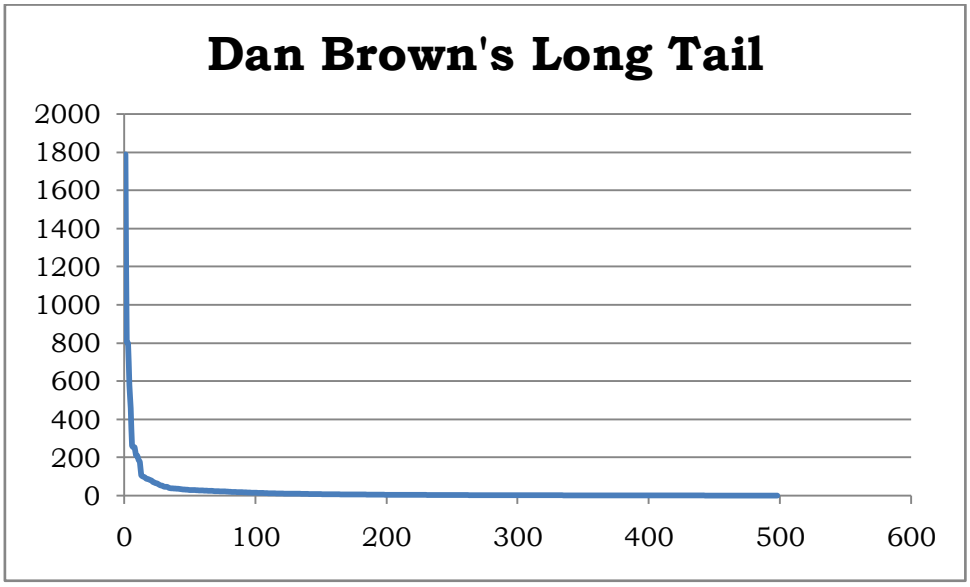


Figure 6: Power Law for Dan Brown's Da Vinci Code in LibraryThing

A *narrow folksonomy* is useful in the context of tagging objects that are not easily searchable or can only be described using text. Tagging is done by one person or a limited number of people. Tags are directly associated with the object.

Vanderwal's representation depicts the person creating the object and applies a descriptor in the form of a tag. The users of the system can also apply tags to describe the object or to help them find it again. In this example group A uses the tag supplied by the creator to find and come back to the object. Groups B uses tag 1, but has applied tag 2 as well. Group C only consumes tags 1, 2 and 3. The same goes for group D for the tags 2 and 3. Group E was not able to find the object because of a mismatch of vocabulary. Group F uses its own tag 3 to find the object, which it has found through other means than the existing tags.

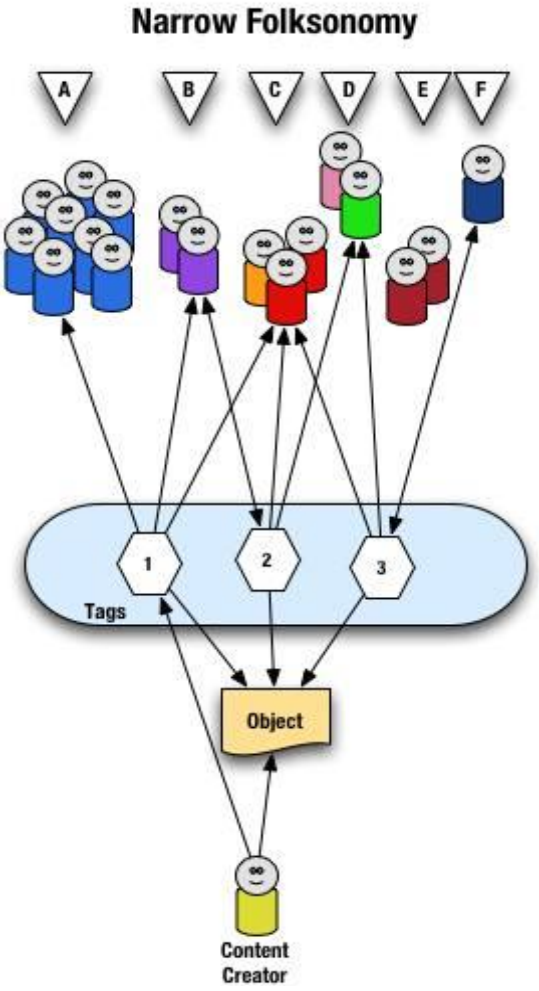


Figure7: Narrow folksonomy (Vanderwal)

Although a narrow folksonomy is not as rich as a broad one, it can still generate a certain value because of the text tags being applied to objects that are not findable using search or text related tools.

1.3 Advantages and disadvantages

The advantages and disadvantages of a folksonomy stem from the fact that tags are based on natural language, which might create problems concerning synonyms, homographs, accuracy and compatibility.[21]

Synonyms are words which refer to the same entity. Science Fiction and sci-fi are equal in meaning. Items tagged with science fiction are not retrieved when searching for sci-fi. The problem is compounded even more with acronyms. SF is a generally accepted acronym for the genre, but it might also stand for San Francisco. Related is the problem of plurals and errors in spelling. If the database deems "cat" and "cats" to be different entities, a

search on either of them will not retrieve the other. The mental model of the tagger defines how she sees the world and subsequently categorizes it. "Reflecting the cognitive aspect of hierarchy and categorization, the 'basic level' problem is that related terms that describe an item vary along a continuum of specificity ranging from very general to very specific; … *cat*, *cheetah* and *animal* are all reasonable ways to describe a particular entity. The problem lies in the fact that different people may consider terms at different levels of specificity."[22] Most people will use the more (but not the most) general description as a basic level than the more specific, e.g. "cat" as opposed to "animal" or "Persian". Ultimately it depends on the need and the level of expertise in a domain of the individual. For some "javascript" might be too specific, while for another "software" is too general.[23]

Homographs are words that are spelled the same but carry different meanings in different contexts. Problems can arise when these words are presented without context, and can thus cause ambiguity. However, the retrieval problem associated with homographs is often more theoretical than actual. Words that are ambiguous on their own usually become clear when used together with other words.[24] A related problem is the issue of polysemy. Polysemous words have many related senses. E.g. "windows" can refer to holes in walls, panes of glass residing in them, or an operating system. Polysemy can dilute query results by returning related but inapplicable items.[25]

A problem which is specific to folksonomies is that of compound tags. Some sites allow multiple word tags, others do not. The result is that users have found ways to circumvent this restriction by compounding tags. Examples include "sciencefiction", "science-fiction", "ScienceFiction" and "science_fiction". Another reason why tags are compounded is to create some sort of a hierarchy in the set of tags an individual uses, e.g. "design:css", "programming/C#". Such tags lower their findability, unless a generally accepted standard form is agreed upon within the community.

The issue of accuracy is a tricky one. Natural languages enhance accuracy since no further interpretation is needed. Controlled vocabularies are

artificial, and thus do not reflect the richness of a natural language. They need to be interpreted. Out of this interpretation inaccuracies can arise.[26] On the other hand, folksonomies lack precision because of the problems cited above. Jason Morrison studied their search effectiveness on web resources as compared to search engines and directories, based on measures of precision, retrieval and recall. Overall folksonomies were the least effective. Search engines have the highest precision and retrieval rate. Directories have the next highest precision rate , but the worst retrieval rate. This in contrast to folksonomies, which had the lowest precision rate, but more than double the retrieval rate of controlled vocabularies.[27] It must be said that this study was done almost two years ago. Since then folksonomies have expanded substantially. They are the most effective when a formidable critical mass is reached. Personally, while doing research into folksonomies sites like Delicious and Cite U Like garnered more relevant results than the Google search engine. The point is, there are several ways of searching for things. Each has its own strengths. A folksonomy is more suited for browsing than for finding, thus enhancing serendipity. By browsing through the content an the interlinked related tag sets it is possible to get the general feel of a subject. For Adam Mathes the difference between browsing a subject area and direct searching to finding relevant documents in a query is like the difference between exploring a problem space looking for the right questions and looking for answers for specific questions.[28]

In terms of cost it is of course much cheaper and less labor intensive to utilize a folksonomy than a full fledged taxonomy. It is not necessary to develop an elaborate system by experts, which is costly not only financially but also in terms of time. Afterwards the architecture needs to be maintained by these experts and users need to be trained in order to get the most out of it. The process to update a rigid controlled vocabulary is lengthy and cumbersome.

2. Previous research

Although folksonomies are the youngest members of the information science family, scholarly discussions ensued fairly quickly. Collaborative tagging became popular around 2004 with the advent of social software applications, ushering Web 2.0 into the world of categorization. Initially, these discussions did not take place in papers and articles, but rather on blogs. One of the first papers is Adam Mathes' *Folksonomies - Cooperative Classification and Communication Through Shared Metadata*, which gives a good description of what folksonomies are about.[29] One of its most ardent proponents from the start is, of course, Thomas Vanderwal. He regularly contributes his thoughts to the debate in lectures and on his blog *Off the Top*.[i] Both have also argued that tagging systems can be used as a tool for personal information management.

The debate has concentrated on the use of folksonomies as general resource discovery and knowledge organization tools. Clay Shirky defends the idea, in his well known speech, *Ontology is overrated*, during the O'Reilly Media Emerging Technology Conference 2005,[ii] that collaborative tagging will eventually supersede traditional classification and categorization schemes.[30] He argues that the "current schemes are incapable of reflecting the transient nature of knowledge and therefore the demands of the modern information user." Since collaborative tagging is inclusive "all users can participate and contribute their own personal vocabularies to generate a collaboratively built 'bottom-up' vocabulary which more accurately reflects users' conceptual model of the world around them."[31] Shirky also notes the economic advantage of utilizing a natural language system in a collaborative fashion (see above). Ian Davis questions these economies. He states that "the total cost of an information retrieval system is the cost of classification plus the cost of discovery."[32] Thus, in formal classification systems a small group of specialists incur a high cost in order to reduce the costs of searching by a large group of people, whereas in a folksonomy it's cheap to classify yet expensive to find. Mathes and Quintarelli have reacted to this kind of

[i] http://www.vanderwal.net/random/index.php
[ii] http://itc.conversationsnetwork.org/shows/detail470.html, 11 July 2008

reasoning by explaining that folksonomies have another purpose, and thus other benefits, than taxonomies.[33] These include the enhancement of serendipity (see above) and the learning curve associated to the search effort itself. Guy and Tonkin have suggested that findability might be greatly improved in folksonomies by improving tag literacy. The measures they propose are aimed at improving tagging systems (e.g. automatic spell checks, synonym suggestions, etc.) and at encouraging users to follow the community's conventions.[34]

Research has been done on different types of sites allowing collaborative tagging in order to define trends and patterns of tag usage. An early research paper is by the hands of Golder and Huberman, in which they have analyzed data from Delicious. Although there is a large variety in the number of tags and what they describe, there is also a certain regularity to be found concerning their nature.[35] The patterns they describe will be discussed in more detail in the next chapter. In general, patterns have been analyzed based on the tuple *User, Resource, Tag*.[36] Tags tend to exhibit a Power Law distribution. When a given resource is added to a folksonomic system a lot of different tags are applied to it. Relatively quickly the number of different keywords drops in frequency.[37] The highest frequency is reserved for a small number of tags. This implies that folksonomies converge towards a consensus concerning the aboutness of a resource.[38] Cattuto, Loreto and Pietronero have stated that the vocabulary used in folksonomies has an emergent nature. Just like in natural languages, they exhibit dynamic aspects, such as the emergence of naming conventions, competition between terms and takeovers by neologisms.[39] Lux, Granitzer and Kern have confirmed these findings. In their sample 80% of co-occurring tags are Power Law distributed. A secondary finding relates to the large amount of tags that are inappropriate for retrieval purposes. These tags can be misspellings and unpopular tags. They can also be personal vocabularies with strong indications for the existence of sub-communities creating semantic islands.[40] Kipp and Campbell have found that collaborative tagging practices work to a certain extent in the same way as conventional indexing, that the tagging

data are to some degree consistent with traditional concepts of aboutness. Next to this, tags also have a distinctly different function. The presence of time and task management tags provides an extra dimension that traditional categorization systems can not accommodate.[41]

There is still some debate about the effectiveness of tagging as a retrieval aid. Chi and Mytkowicz argue that the increasing popularity of a folksonomy decreases its efficiency. More and more resources are added, making the recall of specific tags very great yet imprecise or very low but precise.[42] Others, as we have seen before, claim that the advantages need not be found in exact recall, but rather in the social aspects. Be as it may, a number of authors have proposed improvements and adjustments to the design of folksonomies. Céline Van Damme has made a SWOT analysis of folksonomies and taxonomies in her master thesis *Folksonomies and enterprise folksonomies.*[43] In the section opportunities, Van Damme makes a number of suggestions which might make collaborative tagging systems better. The quality of tags could be improved by setting up a minimum set of rules. Meijas suggest some best practices such as keeping a good balance between idiosyncratic and social tags, the usage of plurals to define categories, including synonyms, and using the conventions of the group. Misspellings could be eliminated by introducing a spelling checker. The credibility of the feedback mechanism could be increased by introducing a ranking mechanism in which particularly useful tags are promoted.[44] This way people might know to which degree certain tags can be trusted. It can also function as a sort of a reputation management system, letting prolific taggers rise to the level of trusted expert. Finally, building a hierarchy in tags could improve retrievability of content.[45] Other suggestions are aimed at the automatic extraction of hierarchies from tags,[46] by automated clustering[47] or the extraction of ontologies.[48]

Although this overview of the research that has been the last few years is hardly complete, it does cover a lot of the ground. More importantly it shows the increasing interest in grassroots classification and the possible implications folksonomies might have for the future of information retrieval.

3. Tagging rights

"The rise of the Internet is affecting the actual work of organizing information by shifting it from relatively few professional indexers and catalogers to the populace at large.... While not consciously teleological, a self-organizing bibliographical universe nevertheless succeeds in meeting the bibliographic objectives in part, occasionally, and somewhat randomly."[1]

1. Introduction

As we have seen in the previous chapters, there is a clear difference between the way language and its internal relations are handled in controlled vocabularies and folksonomies. The semantic relations in natural language systems do not follow the same rules as hierarchical classifications and categorizations. As a result, a discrepancy has arisen between the language of catalogers and that of users/searchers. The object of any of the aforementioned systems is to enhance the findability of information through the retrieval of documents. In order to do so metadata is added to said documents. Historically the creation of metadata has been the "domain of dedicated professionals working with complex, detailed rule sets and vocabularies."[2] The production of documents (in the broad sense of the word[i]) has expanded exponentially since the Second World War. It has become increasingly difficult for professional intermediaries (catalogers and archivists) to keep up. One approach to counter the tide is the implementation of author generated metadata. Within the archival and the library sciences community there are many proponents of this concept. A well known initiative is Dublin Core (see chapter 1). Recently, with the advent of social bookmarking sites, a new approach has received some

[i] "A document can be described as 'recorded information or object which can be treated as a unit'. Den Teuling adds to this that it is also a 'whole of coherent data, recorded on one or more media', where 'coherent digital data, recorded on more than one medium, form one document if they are to be retrieved in one action.' But also, 'a document of which parts are recorded on different media." Translated from: Sterken, V. Op zoek naar Vaste Waarden. Vooronderzoek Digitale Archivering aan het Vlaams Parlement, 2005-2006. Master Thesis, Free University Brussels (VUB), 2006, pp. 7-8

attention, i.e. user generated metadata. Each method has its advantages and disadvantages in regard to catering to user needs (see chapter 2).

Adding metadata is done, in general, to denote the aboutness of a document. In *Catalogers' common ground and shared knowledge*[3] Alenka Šauperl uses the term "subject identification" in this context. She analyzes the cataloging process in light of the problem of multiple interpretations. Perception of a document is influenced by a person's background and education, but also by the social group or the culture she belongs to. Three levels of interpretation are discerned, i.e. the perspective of the tradition (discipline) and reality (current needs and intentions) of the author, the intermediary, and the user/searcher. "One component is the author's intention (reality) expressed in his or her words (text) and the discourse (language and writing customs) of his or her discipline (tradition). The other component is the reader, the user of the information, who brings his or her own background (discipline—tradition), reasons for selecting the work (reality), and purposes of its use (text). Standing between these two is the cataloger (or indexer), who attempts to link the author's intention and the user's needs within the context of his or her own perceptions."[4] Catalogers are aware of this problem and try to limit the multiple meanings subject headings might have. Nonetheless, Šauperl's research shows that they tend to be more oriented towards their professional community than to authors or users. The sources of inspiration for generating subject headings are: the document, previous experience, the cataloging practice and the local library catalog, other catalogs (notably the Library of Congress), the subject headings list, and reference sources. Of these six sources, only the document is shared with the author and the reference sources are shared with the user.[5] Tennis situates this analysis in the descriptive manifestation of subject cataloging, which is juxtaposed with the prescriptive (textbook) manifestation. The latter is a practice which identifies the needs of users for finding and collocating stock in a library by subject based on precoordinated classification. The prescriptive manifestation is present in a folksonomy in some way when we take its purpose into account. Collaborative tagging systems aim at the sharing and

managing of resources.[6] In this sense, folksonomies bring the interpretation of the user to the foreground. While author generated metadata can be added by the authors themselves and/or automatically, and intermediary generated metadata can be applied by professionals, folksonomies might well fill the gap concerning the view of the user.

In this chapter the differences between intermediate and user generated metadata will be analyzed, using the site LibraryThing[i] (LT). LibraryThing "is an online service to help people catalog their books easily. Because everyone catalogs together, you can also use LibraryThing to find people with similar libraries, get suggestions from people with your tastes and so forth."[7] In most websites that use social tagging systems, the content itself is immediately accessible via the web. When looking for information about a certain topic in a library catalog this is not the case. The searcher will need to select the books he is interested in, in order to retrieve them later. Overall, librarians, who have a professional relationship with the material, have assigned descriptors to facilitate retrieval. LT allows its members to catalog their own books. Metadata is assigned in the form of tags by users, who have a personal relationship with the material they tag. Besides the cataloging of one's own collection, the site can also be searched for books of others regarding certain subjects.

As one might imagine, the service is also very popular with librarians and other information specialists. The questions arises how natural language keywords will relate to controlled vocabularies. Moreover, one would expect that within the group of librarians using LT, the language of the intermediary will be reflected. Given the habit of cataloging for a living, it would be probable that the professional language is maintained in another system because of habit. At the least it is expected that broader and narrower terms of a concept will be more prominently present within this group. On the other hand, no matter what the background, in LT the cataloger is also the user and vice versa. The effect of this ambiguity is

[i] http://www.librarything.com/

difficult to predict, since humans easily switch between roles and viewpoints depending on specific situations.

2. Methodology

First, an in depth explanation will be given about the functionality of LibraryThing. It will become clear that LT is a broad folksonomy with extensive social capabilities. Besides this, the search and retrieval methods of the system will be discussed. These, and other features, are constantly being enhanced, stemming from the founder's implicit Open Source[i] mentality and belief in his product. During the time of writing of this thesis, the new features of the home page and "find friends" was introduced, the idea of an Open Shelves Classification system was launched, and a books Application Programming Interface (API) was released under a Creative Commons[ii] license.[8] The richness of the site makes LT an interesting object of study.

Next, the tags will be examined according to the functions that they perform. The object here is to determine if tags are used for different purposes than from those in other folksonomies, since the tagged content is not immediately accessible. To this end the dataset will be analyzed based on the categories proposed by Golder and Huberman.

Then, a comparison will be made with intermediary generated metadata. To bibliographic records keywords from controlled vocabularies are added in the form of subject headings. After an initial comparison, based on the subject headings which are available in LibraryThing itself, the analysis was narrowed to the authoritative Library of Congress Subject Headings (LCSH). First, it was verified to what extent LCSH terms exist in LT in the form tags.

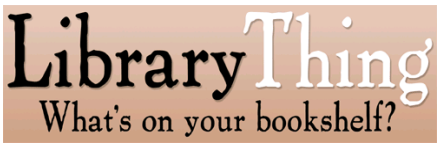
---

[i] Open Source is a development methodology with an emphasis on sharing sources (knowledge and goods). For concise background information, see: Tapscott, D. & Williams, A. Wikinomics. How mass collaboration changes everything. London: Atlantic Books, 2008 (2nd ed.), pp. 65-96

[ii] "Share, Remix, Reuse — Legally. Creative Commons provides free tools that let authors, scientists, artists, and educators easily mark their creative work with the freedoms they want it to carry. You can use CC to change your copyright terms from "All Rights Reserved" to "Some Rights Reserved."" http://creativecommons.org/

This was followed by an analysis of the information value of tags as compared to subject headings.

3. LibraryThing



3.1 What is a LibraryThing?

"LibraryThing is a social network for bibliophiles. You catalog the books you have … or are interested in, and the books you have connects you to other people."[9] The site was (is being) developed by Tim Spalding[i] and went online on 29 August 2005. It has over 450 000 registered users[ii], who have saved more than 28 million books with more than 37 million tags.

3.1.1 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[iii]. The data about the books are imported automatically through a connection to libraries (providing MARC and Dublin Core records) and commercial booksellers. By default the databases of Amazon.com and the American Library of Congress are being searched. This can be expanded by choosing from 690 other sources from around the world, including national, public and university libraries (such as the Charles Darwin University, Yale, Coquitlam Public Library, Zhejiang Provincial Library, Kolding Folkebibliotek, and de Koninklijke Bibliotheek van België), and from commercial sites such as Deastore and Bol-Bruna. Since its origin lies in the United States, English sources are the most represented. However, a fairly large number of other languages are available as well. Some languages are better served than others but, since its addition, this feature has kept on expanding.[iv] If neither of these sources would garner any results, the opportunity remains to add data manually. A final way of adding books is by importing them from other websites such as Goodreads, Shelfari, Amazon wishlists, … or by uploading files such as TEXT, CSV, XML, etc.

---

[i] http://www.librarything.com/profile.php?view=timspalding
[ii] Or *thingamabrarians.*
[iii] International Standard Book Number, a unique, numerical commercial book identifier.
[iv] There are also a number of translated versions of the site available at separate URLs. E.g. www.librarything.nl, www.librarything.de, www.librarything.it

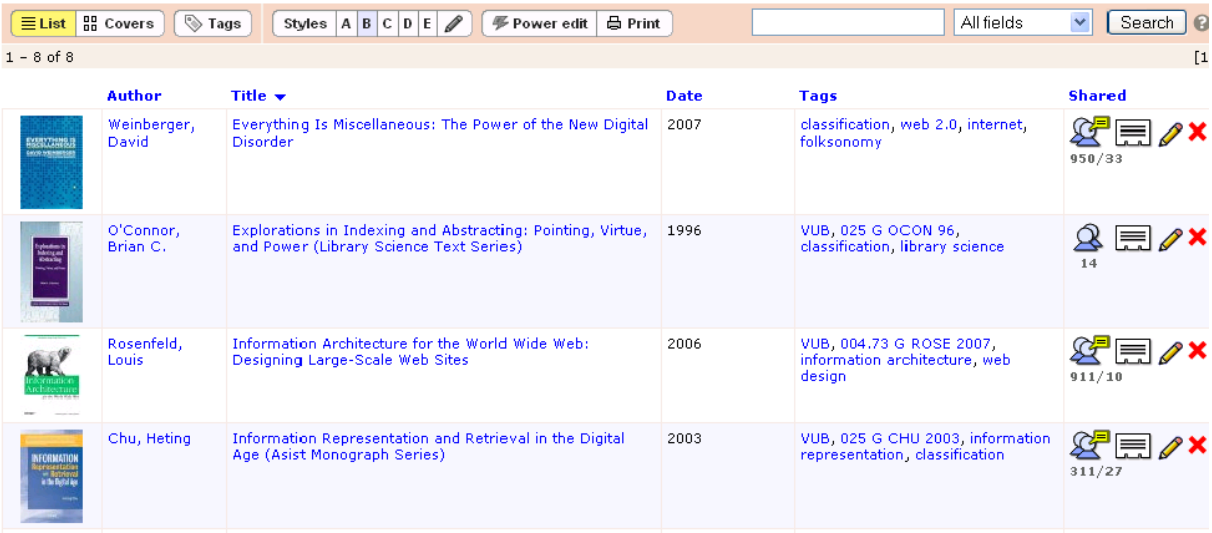


Figure 8: Example of a personal library

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." The only limitations placed on tags are that they can not contain commas or exceed 30 characters.[10] There exist several views of a catalog. One possibility is shown in figure 8. 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. What type of information is shown in a list can be customized. It is possible to define five different styles, denoted as style A to E. In the example above I have chosen to display the book cover, the author, the title of the book, the date of the edition and the tags I have selected. Other options include multiple authors, a rating, comments, reviews, Dewey Decimal Classification numbers, date read, important characters, Google Book search and more.

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, tagclouds are naturally present. Different representations are available. On the highest level, we find a fairly large tagcloud of the top 75 tags,[i] as well as an authorcloud of the top 75 authors.[ii] 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.

3.1.2 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."[11] The site started out as a way of cataloging ones 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 your self," 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."[12] This connection takes place by either becoming 'friends', like on most social networking sites, or placing yourself on a 'private watchlist' or a watchlist of 'interesting libraries'. In either case, you can follow updates of newly cataloged books in the libraries of your connections. Further more there is the possibility of leaving comments on a member's profile page. Most

[i] http://www.librarything.com/tagcloud.php
[ii] http://www.librarything.com/authorcloud.php

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,[i] ranging from Fantasy or Science Fiction Fans to Non-fiction Readers, Graduate Students, Happy Heathens and everything in between.[ii]

On the profile page personal information is shown. Some is added automatically (which tags you have used, to which groups you belong, your LT URLs), other things can be added manually (e.g. about me, favorite authors, about my library, real name, location). A way of connecting with other people is by looking at the pane next to your profile which shows other thingamabrarians who own the same books as you.

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 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.

3.1.2 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.*[13] 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].[iii] 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

[i] http://www.librarything.com/zeitgeist, 27 April 2008
[ii] http://www.librarything.com/groups
[iii] http://www.librarything.com/groups/iseedeadpeoplesbooks

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.

In order to search for a given book there are a number of possibilities. A member can browse through another member's catalog. A second option is to do a keyword search on the title, the author or ISBN. The third solution is to search the tags. It can be argued that because of the presence of a large amount of idiosyncratic tags not all relevant information can be found. When searching for all the books tagged with "fiction", you would like to see all of them without missing out on the ones tagged with "Fiction", "FICTION" or "fition". LT tries to counter this problem by aggregating tags. Users can combine tags, thus creating an enlarged tag space.

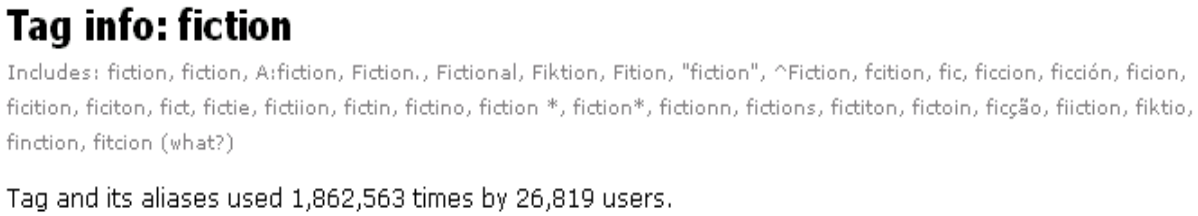
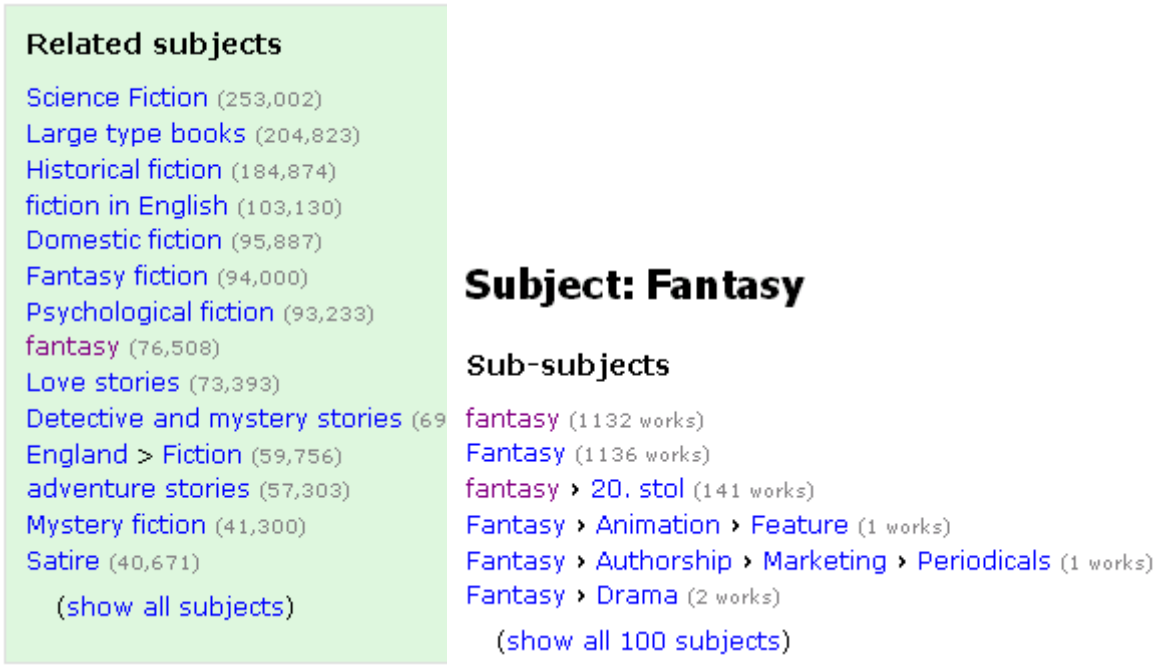


Figure 9 : Info about the tag fiction

Further more, on the right hand side of the screen several other options are presented to help the user to expand or limit her search. A box containing a tag cloud with *related tags* allows broadening or narrowing the search.

In order to dig deeper into the subject matter, and thus to decrease the number of retrieved items the function of *related subjects* can be used. This allows you to explore the related, broader and narrower terms. The significant difference between tags and subjects is that tags are submitted by the users of LT, while subject headings are only available for those books for which data is derived from library catalogs (e.g. Library of Congress, National Library of Australia, Vlaamse Centrale Catalogus, SUDOC, Biblioteca Apostolica Vaticana …).[14] Each subject also lists the related sub-subjects, allowing the searcher to refine her search. Since subject headings are the result of a pre-coordinated effort they have a narrower scope than tags. Imagine searching for a manual on writing a fantasy book. The term "fiction" is a bit too broad to be really useful on its own. So, we narrow the search to the subject heading "fantasy".



In the list of sub-subjects you will find "Fantasy > Handbooks, manuals, etc.", leading you to Allan Kronzek's *The sorcerer's companion: a guide to the*

magical world of Harry Potter. Unfortunately this is not what she is looking for. It is possible to adjust the search using the *related subjects* pane once more. So, let's try "Fantasy fiction, English > History and criticism > Handbooks, manuals, etc". This garners a list of 13 books, none of which are relevant to her query. She could continue browsing through these subjects, but actually finding what she is looking for could take quite some time. By contrast, a Boolean AND search (see next paragraph) brings up a list of 54 books, including Orson Card's *How to write Science Fiction and Fantasy*, Stephen King's *On writing: a memoir of the craft*, David Gerrold's *Worlds of wonder: how to write science fiction & fantasy* and Ursula Le Guin's *Steering the craft: exercises and discussions on story writing for the lone navigator or the mutinous crew*. These three books were found with a single search, whereas if she would have browsed through the subject headings she would have needed to search on various forms of the terms "authorship" and "writing".[i] When searching in a digital environment people tend to use natural vocabulary as keywords. In order to find resources categorized with the term "authorship" a comprehensive list with synonyms, related and preferred terms must be maintained, which is very labor and cost intensive for the developers.

A very interesting feature is the *related tagmashes*. The naming is reminiscent of mashups. A mashup is the creation of something new by combining two or more elements. The roots can be traced back to the Jamaican Dub Mashups (or remixes) which originated in the late 1960's.[15] The term has taken on other meanings since. Digital mashups are digital files containing pre-existing text, graphics, audio, video and/or animation which have been combined to make a new derivative work.[16] The neologism has also found its way into the web 2.0 realm, where it has come to mean a "web application that combines data from more than one source into a single

[i] For Orson Card: "Creative writing"; "Fantasy literature › Authorship"; "Fantasy literature › Technique"; "Science fiction › Authorship"; "Science fiction › Technique"; For Stephen King: "Authors, American › 20th century › Biography"; "Authorship"; "Horror tales › Authorship"; "King, Stephen, 1947- › Authorship"; For David Gerrold: "Science fiction › Authorship"; For Ursula Le Guin: "Authorship › Problems, exercises, etc"; "Creative writing › Problems, exercises, etc"; "Narration (Rhetoric) › Problems, exercises, etc"

integrated tool; an example is the use of cartographic data from Google Maps[i] to add location information to real-estate data, thereby creating a new and distinct web service that was not originally provided by either source."[17]

Tagmashes can be used to combine two and more tags in order to refine your search. As a result of the submitted query, the system will generate a page based on the combination of the search terms. The results of these queries are subsequently saved in the database. This way they can be proposed to users. This feature enhances the scalability of the system. Not only do the tags themselves aid retrieval of resources, the searches do so as well.

Related tagmashes

fiction, literature (105)
classic, fiction (103)
fiction, paperback (103)
favorite, fiction (100)
fiction, literary (100)
fiction, literary, novel (100)
fiction, novel (100)
fiction, tbr (100)
england, fiction (99)
british, fiction (98)
family, fiction (98)
favorite, unread (98)
literature, novel (98)
family, novel (97)
fiction, roman (96)
novel, unread (96)
2007, fiction (94)
fiction, love (94)
film, novel (94)
literature, unread (94)

Here you can see the *related tagmashes* for our top tag "fiction". The associated tagmashes that are shown dig a bit deeper, but do not divert too much from the top level. Imagine that a user would like to read a historical novel set in nineteenth century France. He could type in the search string "history, novel, fiction, 19th Century, France", leading him to Hugo's *Hunchback of Notre Dame*, based on the following result:

Tagmash: 19th century, fiction, france, history, novel

Mashing tags

19th century (Includes: 19th century, 1800's, 1800s, 19 c., 19 century, 19 th century, 19-th century, 19. Jahrhundert, 19. Jh., 19. Jhd, 19. Jhd., 19. sec., 19.Jahrhundert, 19.Jhd., 19c, 19c., 19century, 19eme siecle, 19th c, 19th c., 19th cent, 19th cent., 19th centruy, 19th century ad, 19th century., 19th cenury, 19th cenutry, 19th cty, 19th-c, 19th-c., 19th-century, 19th.C., 19thc, 19thcentury, 19th_century, ad 19th century, ad.19, c19, c19th, CE 19thC, ce19, nineteeenth century, nineteenth centruy, nineteenth century, nineteenth-century, nineteenthcentury, novecento, s.XIX, século xix, siglo 19, siglo xix, xix century, xixe siècle, XIXth century)

fiction (Includes: fiction, fiction, "fiction", A:fiction, fcition, fic, ficção, ficción, ficcion, ficion, ficition, ficiton, fict, fictie, fictiion, fictin, fictino, fiction *, fiction*, Fiction., Fictional, fictionn, fictions, fictiton, fictoin, fiiction, fiktio, Fiktion, finction, fitcion, Fition, ^Fiction)

france (Includes: france, フランス, França, frankreich, g:france, 法國)

history (Includes: history, history, @history, geschichte, geschiedenis, hietory, Hiistory, hisotry, hist, história, histoey, histoire, historia, history., histpry, histroy, histry, hitory, hsitory, hsstory, 歷史, ^History)

novel (Includes: novel, novel, Novel., novela, novels)

Figure 11: Example of a tagmash

[i] http://maps.google.com

The *related tagmashes* pane also generates a more specific result, more closely associated to the newly created search string, including "19th century, french literature" and "historical fiction, paris, romance". There exists a certain overlap in books when using these associated tagmashes, nevertheless your search becomes more specific to what you're looking for.

All these different features aid in exploring the "fiction" tagspace, which is larger than the term itself. As we have seen before, the frequency distribution of all the multiple word tags is significantly higher than the term "fiction" on its own. Within this group exists an implicit hierarchy, in the sense that different levels of relationships can be distinguished. There are related terms like "literary fiction" (and all its variants e.g. "literature & fiction" or Fiction – literature) and "general fiction", and narrower terms like "children's" and "historical fiction". In this example we'd be a bit hard pressed to find a broader term.

All of this is possible because of the fact that all members (or a lot at least) have tagged their own resources, have added other useful metadata and have taken the time to combine tags that are the same in writing and meaning. The result of this, mostly personal, effort is a robust database with fairly accurate data concerning the classification of books containing the largest (virtual) library in the world.

3.2 The dataset and its tags

Data has been collected for the 200 top books[i] during the last week of March. The object of this exercise is to study 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 thesis is Librarians Who LibraryThing,[ii] 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

[i] 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.
[ii] http://www.librarything.com/groups/librarianswholibrar

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",[i] 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. This information was presented in the following form:

- Total tags
    - Librarians: # of tags
    - Non-librarians: # of tags
- Librarian tagging
    - Tag_1: # of tags
    - …
    - Tag_n: # of tags
- Non-librarian tagging
    - Tag_1: # of tags
    - …
    - Tag_n: # of tags

This 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,[ii] the minimum 3985,[iii] with an average of 6187, 86 users per book.

For these 200 books 76 810 unique tags have been applied. Librarians have used 13 503 different tags, non-librarians 70 853. As you can see an overlap exists 7545 tags exists. Thus, roughly half of the tags used by librarians

[i] http://www.librarything.com/z_books.php
[ii] J.K. Rolling's *Harry Potter and the Sorcerer's Stone*
[iii] A.S. Byatt's *Possession : a romance*

were also used by non-librarians. These tags were used more than once for different resources.

The total number of tags in the study was 1 292 111. The maximum comprised 26 843 tags, the minimum 2381, with an average of 6518, 26. The group of librarians has contributed 126 467 tags, or 9,79%, with a maximum of 2898, a minimum of 226 and an average of 652,55. The non-librarians took up 1 147 278 tags, or 88,79%.[i] Their maximum was 23 945, the minimum 2104, with an average of 5 865, 71.

Upon closer inspection the smaller group of librarians resembles the larger group, given the general frequency distribution of applied tags. An overwhelming amount of tags (76,03 %) have been used only once in both groups. The effect of a Power Law,[18] or Long Tail,[ii] is apparent. A limited number of tags constitute a large part of the tag space. This can be illustrated with the distribution of tags in the number one shared book in LT, *Harry Potter and the Sorcerer's Stone*, in the next figures.

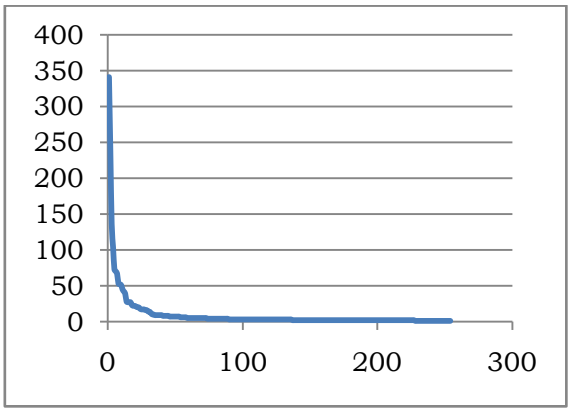


Figure 12: Distribution of tags – Librarians

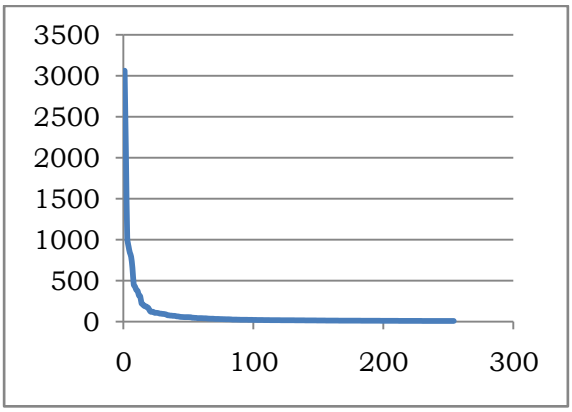


Figure 13: Distribution of tags – Non-librarians

In both cases the highest number of applied tags drops rapidly. 24 861 users have tagged this resource 26 843 times (5776 and 23 945 times by librarians and non-librarians respectively). For this book it took less than 250 keywords to arrive at the tags used less than ten times.[iii] In more general terms, 257 313 tags (76,03%) have only been used once by the users

[i] A small group of users did not tag their resources. This accounts for the remaining 1,42%.
[ii] Also called Zipf's Law and Pareto distribution, depending on the discipline.
[iii] Librarians: 32; Non-librarians: 198

of our dataset. Tags that have been used twice amount to 33 616 (9,93 %). Going up the ladder of increasing tag usage, their total numbers drop quickly. From tags that have been used only six times until the highest frequencies account for less than 1 % of all unique tags. This means that three quarters of the tags are only used by one person. We could argue that they, in effect, contribute little to the community as a whole. The story is a bit more complicated though.

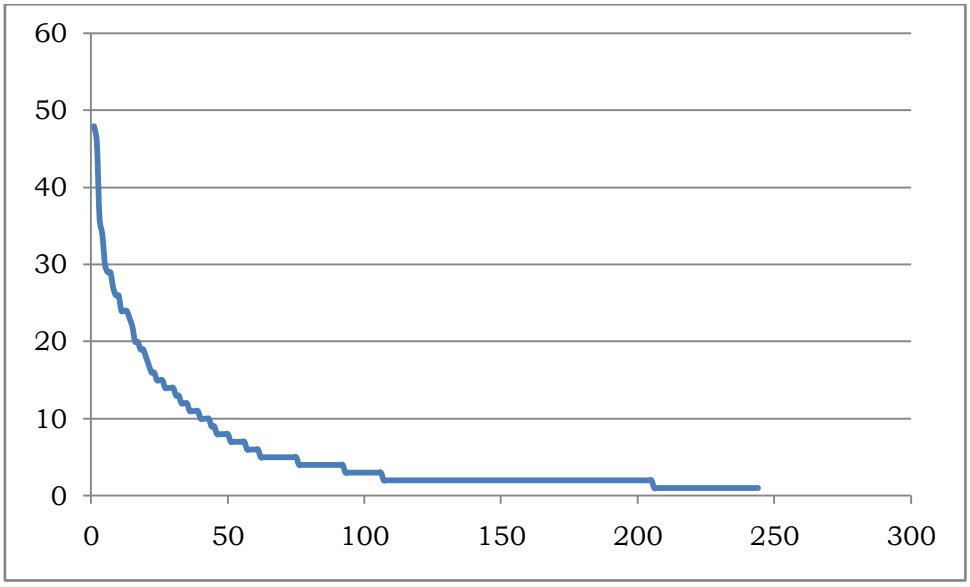


Figure 14: Distribution of a sample of the "fiction" tag space

The most used tag is "fiction" (48 times), followed by "Fiction" (46 times). As you can see, focusing on each tag as a unique keyword does not tell the whole story. By doing a simple query, in which upper and lower case letters are ignored, the frequency increases to 124. By including all the multiple word tags that contain the word "fiction" we get a total of 1802 tags with a frequency distribution of 4158. Within this limited set, the same Power Law is applicable.

As we have seen in figure 11 there is the possibility of combining tags (and authors and books as well). This feature is only available to the paid for accounts. There is even a dedicated group called Combiners![i] The object here here is to combine tags that are identical both in meaning and in use. "In general, tags should be combined only if they ALWAYS overlap in meaning,

[i] http://www.librarything.com/groups/combiners

not if they CAN overlap. For instance, combining "world war 2" with "world war ii" is good; combining "science fiction" with "SF" is not good, since other people use SF to mean "San Francisco" (or something else entirely). Combining acronyms with non-acronyms should be avoided in general."[i] In this way the typical problems related to spelling are partially alleviated. When searching for the tag 'utopia', for instance, books tagged with 'utopias', 'utopian', 'utopie' (Dutch), 'utopian society' and 'utopian_societies' are also returned.[ii] Because this system is based entirely on the input of LT members, it is not perfect. 'YA' is generally accepted as being an abbreviation of young-adult fiction. In LT it is not combined with anything.[iii] Probably this is because of the possibility that it might mean something else (Yorkshire Accent?). 'Young adult' is combined with all the variations on its spelling, but not with 'young adult fiction'.[iv] Here the question can arise if the term relates to the type of fiction or to the protagonists of the story. Finally, 'young adult fiction' does combine works tagged with, amongst others, 'youngadultfiction', 'ya fiction', 'Fiction/YA', 'Fiction: young adult'.[v] Despite this imperfection, the the accuracy of the search results is enhanced dramatically thanks to the work of a group of dedicated volunteers. The abundance of tags that have been used only once in general adds a lot of noise to folksonomies. In LT the noise has been reduced to a certain extent by providing a feature that has the ability to leverage the power of its community. Nevertheless, a significant part of the one time tags that should be combined with others, might escape scrutiny. The power of a folksonomy lays in the fact that the "bad" tags do not have to interfere with the "good" ones. The large number of people tagging on LT ensures that (most) works will not be lost in obscurity. Further more, low frequency tags which are purely idiosyncratic can be very useful for the person using them, without them laying a burden on the rest of the community.

[i] http://www.librarything.com/wiki/index.php/Tag_combining
[ii] http://www.librarything.com/tag/utopia
[iii] http://www.librarything.com/tag/ya
[iv] http://www.librarything.com/tag/young+adult
[v] http://www.librarything.com/tag/Young+Adult+Fiction

3.3 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[19]:

1. Identifying What (or Who) it is About.
2. Identifying What it Is.
3. Identifying Who Owns It.
4. Refining Categories.
5. Identifying Qualities or Characteristics.
6. Self Reference.
7. Task Organizing.

These categories seem to be applicable to LT as well. 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."[20] 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*."[21] 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. …"[22] 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 existing categories. Numbers, especially round numbers (e.g. 25, 100), can perform this function."[23] 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."[24] 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.”[25] 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*.”[26] 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[i] 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.[27] To conduct their experiment, tagging features were added to a movie recommendation site.[ii] 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,*[iii] *self reference, task organization*). The final

[i] http://www.citeulike.org/
[ii] http://www.movielens.org/
[iii] The others assume that users will claim ownership of certain bookmarked URL's. In LT this cannot be the case since the owner of the intellectual content is clearly the author of the book in question.

distribution of tags across these classes was 63% factual, 29% subjective, 3% personal and 5% unknown.[28] 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 classification 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. Even if tags are only used for personal gain, the user will not search very often on terms representing an opinion. Adding opinions in the form of tags can either be done as an emotional reaction to the content, or with the community in mind. If someone is browsing through your personal library, he can see which tags you have added to your books. The first attempt at a search however will be done on the search tab. On the other hand, it is possible to add someone's library to your *private watchlist*. Under the *your profile tab* you can find the subtab *connections*. By clicking on *your private watchlist* a list of recently added books of the libraries on this list is generated. However, there are other ways to express an opinion about a book which are far more effective and with which you can reach a larger audience than adding personal tags. For every book it is possible to make a review. Some of these are quite lengthy and detailed, others are short and emotional. Anyone who wants to know whether he would like a particular book is more likely to read these comments rather than sift through personal libraries looking for tags. On the Thingology blog Tim Spalding has described the Long Tail of Ann Coulter's *Godless: the church of liberalism*. Ann Coulter is an American conservative, right-wing political commentator, known for her controversial points of view. Needless to say that she has an equal amount of avid followers and people who hate everything she does. Apparently this is very much reflected in the tags on Amazon with such terms as 'hateful', 'propaganda', 'the truth', and 'brilliant intellect'. "In LT the same Long Tail is visible, with the significant difference that it has a fairly unremarkable tag cloud, touching on its subject matter and point of view, on Amazon, the tagging has devolved into a shouting match. I don't think the people who tagged the book "asshat," "vomit" or "w h o r e" are using tagging as a

memory aid ("I forget—what books did I think are 'asshat' anyway?"). They're using tagging as a sort of drive-by review. Now, a case can be made that Amazon's tags are signaling something important—this is a "controversial" book indeed! The LibraryThing tag cloud doesn't show that as starkly. On balance, however, I think opinion tags corrupt the value of tagging."[29]

In LT this *drive-by reviewing* is not very prominent. In the end of the Long Tail tags can be found which are only useful by the tagger in question because of their idiosyncratic nature. A certain number of them are still of use to the community thanks to the combining feature. Others are mainly used as memory aids e.g. location tags.

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 does not figure prominently since this kind of information is already supplied by the system. The attribution of characteristics is not predominantly present, probably because there are other ways of expressing certain sentiments.

3.4 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."[30] 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.[31]

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."[32] 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* [i] 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* [ii] has as headings 'Business Intelligence > Fiction', 'Information highway > Fiction' or simply 'Science Fiction'.[iii] Connoisseurs of Science Fiction however know that this is a classic example of the sub-genre

[i] http://www.librarything.com/work/73667
[ii] http://www.librarything.com/work/609
[iii] And curiously enough 'Nervous system > Wounds and injuries > Fiction'.

Cyberpunk.[i] Unsurprisingly in LT it is the book tagged the most with this term.[33] 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.[ii] 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

[i] http://www.librarything.com/tag/cyberpunk; Style of science fiction with a focus on advanced science "coupled with a degree of breakdown or radical change in the social order." Cyberpunk. In: *Wikipedia*, http://en.wikipedia.org/wiki/Cyberpunk, 12 July 2008

[ii] http://www.steve.museum/

formal descriptions, assigned by art historians and other specialists, and the vernacular language used by the general public for searching the database.[34] 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).[35] 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.[36]

The 21% 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 frappes* are perpetuated. It is likely then that they will have 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 can 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 above 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.[i] 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%).[ii] 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. As in most collaborative tagging

[i] http://catalog.loc.gov/
[ii] 2,31% when the lower frequencies are disregarded for the LCSH, 1,64% for the genres.

sites, LT displays tags in a cloud. The relative size of a tag indicates its frequency. Each book has its own tagcloud. By default those which have been used most are shown (although it is possible to click through to see all the tags). 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. A few examples might clarify.

The most shared book in LT is J.K. Rowling's *Harry Potter and the Sorcerer's Stone*. The quality of tags and the LCSH are nearly equivalent in terms of a correct subject analysis.

Figure 15: *Harry Potter & The Sorcerer's Stone*'s tagcloud and subject headings

The book is about the adventures of Harry Potter, a new student at a school for magic in England. This is the first book (American edition) in 6 of the highly popular series of fantasy books. The subject headings situate the contents within the realm of magical schools in England and denote is fiction, more specifically juvenile fiction. The LT tagcloud does more or less the same. Juvenile fiction is being replaced by the tags "children's literature", "children's books", "ya" and "young adult", but the effect is roughly the same. An extra dimension is being added though with the terms "adventure" and "fantasy". The subject heading "schools" and "wizards" are elaborated upon with the tags "boarding school", "witches", "wizardry", and "witchcraft". On the other hand the more complete "Hogwarts School of Witchcraft and Wizardry (Imaginary place)" has been abbreviated to simply "Hogwarts". On the whole, this means that the tagcloud gives a slightly more complete view of what the well known bestseller is about, but only slightly.

For J.R.R. Tolkien's *The Fellowship of the Ring*[i] the tagcloud is clearly more comprehensive.

Figure 16: *The Fellowship of the Ring*'s tagcloud and subject headings

[i] http://www.librarything.com/work/3203347

This is the first book in the *Lord of the Rings* trilogy, which tells the epic tale of a war between good and evil in an imaginary past of our planet, called Middle Earth. Before the arrival of humans, the land was already populated by other races, such as elves, dwarves, hobbits, wizards and the evil orcs. The tagcloud shows the same elements as the subject headings, i.e. one of the main protagonists of the story, the geographical setting and a genre description. Besides this, it acknowledges the existence of the other races mentioned above, and the fact that it is an epic tale with its own mythology.[i] A refinement of the genre is also provided by the terms "high fantasy" and "epic fantasy", which is specific to works set in alternate realities. The fact that the book is a part of a series is also emphasized by the tags "lord of the rings" and "trilogy". An extra piece of information can be found in the term "Inklings". The Inklings was an informal literary discussion group from Oxford. Between the early 1930's and the early 1960's writers such as Tolkien, C.S. Lewis, Owen Barfield, Charles Williams and others met at a regular basis to listen to each others' unfinished works and to have literary discussions. For readers who are interested in the books of Tolkien and would like to read more of likeminded individuals, this particular tag could provide the starting point for an exciting literary journey. It has another function than that of subject headings however. Regardless, in this case the tagcloud is more complete in terms of subject analysis than the LCSH.

The situation is reversed for Charlotte Bronte's *Jane Eyre.*[ii] The book describes the life of a plain-faced, intelligent English orphan, who becomes a governess later in life and ends marrying her employer. She aspires to more in life than what is traditionally accorded to her sex in Victorian society. It tells the story of a "heroine … whose virtuous integrity, keen intellect, and tireless perseverance broke through class barriers to win equal stature with the man she loved."[iii] The LCSH captures (nearly) all the elements of the story. The same cannot be said for the tagcloud. The top tags capture the

[i] although the term "mythopoeia" would have been more appropriate.
[ii] http://www.librarything.com/work/2204
[iii] Amazon product description, http://www.amazon.com/exec/obidos/ASIN/0553211404/, 17 July 2008

fact that it is a love story situated in England, with a governess as main character.

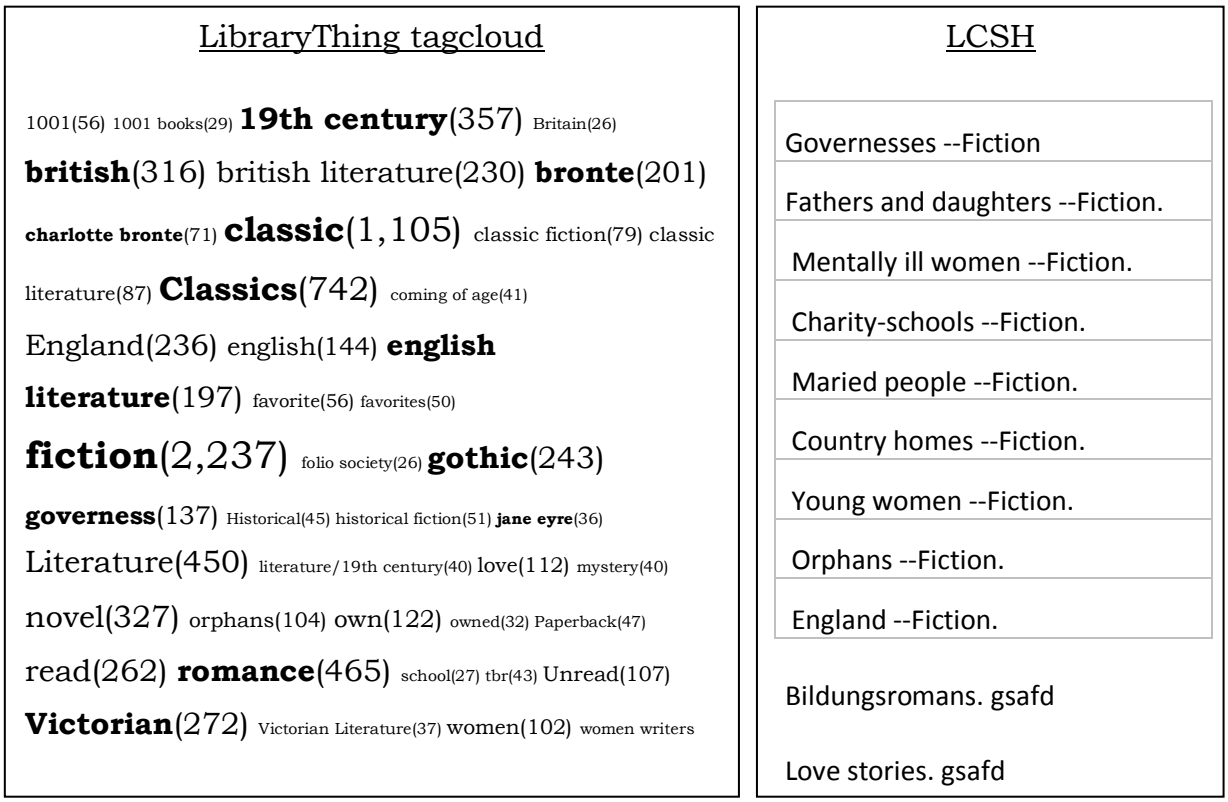


Figure 17: *Jane Eyre*'s tagcloud and subject headings

It does however leave out a lot of other things, including emphasis on the coming-of-age of the heroine, captured by LCSH as "bildungsromans". The cloud is also cluttered by several synonyms, which is a typical problem of folksonomies. Nevertheless, a few concepts are introduced that are useful. These include the era of writing ("Victorian", "Literature/19th century") and certain style elements ("gothic"[i]). Overall, the subject headings seem to give a better representation of the book's contents.

To reiterate: depending on the specific book tags or subject headings are better suited for subject analysis. Whatever the case, when enough people tag a resource, manifestly wrong terms are generally filtered out. Although the LT folksonomy does not always capture all the relevant terms that can be attributed to a resource, they practically always add at least one element

[i] Gothic fiction as a genre combines elements of horror and romance.

(broader, narrower or related term; or a missing concept) that is absent in the subject headings.

An analysis of the tagclouds alone is not sufficient to understand how well LT fares in terms of subject analysis. As mentioned before, the problem associated with plurals, spelling differences, and ambiguity in meaning are in part smoothed out by the efforts of the community. This means that tagclouds actually hide the precision of LT's descriptors. On the other hand, that's not what a tagcloud is used for in most cases. Sinclair and Cardew-Hall have shown that where an information-seeking task requires specific information, users prefer a search interface. Conversely, when the task is more general a tagcloud is preferred.[37] In other words, tagclouds are useful when the searcher just wants to get a general feel of a subject. When she gets more knowledgeable the search boxes will be used. From this we can conclude that a tagcloud will give an overall visual impression of what a resource is about with the added bonus of aiding serendipity, as opposed to providing an accurate subject description.

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,[i] 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"[ii],

[i] E.g. "speculative fiction", "high fantasy", "magical realism"

[ii] "Cyberpunk culture" does exist as subject heading in the LoC catalog. It only yields 6 results (which do not include the classic example *Neuromancer*) as opposed to 9709 hits in LT.

"steampunk"[i] and "paranormal romance"[ii]. 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.

4. What does it all mean?

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

---

[i] A subgenre of fantasy, denoting "works set in an era or world where steam power is still widely used—usually the 19th century, and often set in Victorian era England—but with prominent elements of either science fiction or fantasy, such as fictional technological inventions … or real technological developments like the computer occurring at an earlier date." Steampunk. In: *Wikipedia*, http://en.wikipedia.org/wiki/Steampunk, 11 August 2008

[ii] A subgenre of the romance novel concerning love stories in paranormal settings and/or between different humanoid species. http://en.wikipedia.org/wiki/Paranormal_romance, 11 August 2008

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.”[38] 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.[39]

It has become clear that the different levels of interpretation of a document don’t intermesh very often. Intermediary generated metadata is rooted in the professional environment of indexers and catalogers. User generated metadata takes its cue from the personal experiences and needs of the user in question; and, to a lesser extent, 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 don’t 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.[i] It is essentially a series of widgets designed to enhance library catalogs with LT data and functionality, such as book recommendations, tag browsing and links to other editions and translations.

[i] http://www.librarything.com/forlibraries/

Conclusion

The effect of ICT on society has been profound in many ways. Amongst other things, it has drastically increased the speed of data creation, a process that was set in motion in the 20th century. Large amounts of data and information can be beneficial only if they can be accessed properly. There are many ways information can be organized. Traditionally, classifications were invented and elaborated upon within the realm of libraries. The underlying structure is based on the worldview, and the place of knowledge therein, of the creator. Rigid, all encompassing structures are the result. They need experts to build and maintain them, and are thus cost and labor intensive. When defining the aboutness of a document, the view (level of interpretation) of the intermediary is represented with the intent of being universal and suited for user searches. These practices are firmly rooted in the physical world. Like subjects must be placed together because the physical world demands that the objects (books) can only be in one place. Unfortunately, sometimes it is possible that an object has many subjects. In the digital world however, such distinctions are less important. Objects can be (virtually) in the same and different places at the same time. Folksonomies allow for personalized views of the information under scrutiny. Instead of following the view of one person, the use of tags permits the creation of a view according to the needs of the moment.

Besides the more philosophical sentiment expressed above, there is also a more practical element to be considered. Retrieval of documents is based on the assigned metadata. Up until recently, metadata was added by intermediaries alone. In many cases it has become nearly impossible to keep up with the flow of information being created. One way of alleviating the problem is (automated) author generated metadata. To this, the idea of user generated metadata has been given a platform through the use of folksonomies.

Collaborative tagging has been researched mainly for web based information. Social bookmarking sites are the paradigmatic examples for folksonomies.

They let users store web links, which allow direct access to the resource's content. For my research I have chosen data from the social cataloging site LibraryThing. On LibraryThing users can catalog their books. The resources they add and tag are thus not directly accessible. In effect the site can be seen as a virtual library. People can search the database to find books on a topic. To actually acquire and read a book, the searcher still needs to go out and either buy it or go to a physical library. To answer the question if books would be tagged differently a dataset was analyzed. A second question concerned the differences between intermediary and user generated metadata.

First the functions of tags were analyzed according to the categories proposed by Golder and Huberman. They have determined that tags serve seven functions: identifying what it is about, identifying what it is, identifying who owns it, refining categories, identifying qualities or characteristics, self reference, and task organizing. Broadly speaking, the tags in the dataset correspond to these categories. The emphasis lies first on categorization (or the first four categories), and secondly on managing the resources (the last two categories). Identification of qualities does play a certain part in tagging, but is less important than might be expected.

Then, tags were compared to intermediary assigned descriptors in the form of subject headings for their information value. Subject headings were compared directly to tags. Only a small percentage of the tags corresponded to them, implying that tags would yield a significant amount of supplementary concepts. Tags and subject headings were then compared to each other based on their descriptive strength. Although subject headings generally speaking are quite adequate in describing content, it was found that tags added almost the same number of additional terms which would enhance the findability of the resources.

At the same time it was verified whether the professional background of catalogers would be reflected in the way they tag. The gathered data was divided into two groups. A larger group of 'regular' users and a group of users who belong to Librarians who LibraryThing were considered. The latter

was taken to consist of users with a background as professional catalogers. The hypothesis was that the wording and concepts used in their professional environment would seep into their tags. As it turns out, this hypothesis was wrong. The Librarians who LibraryThing tended slightly more towards the wording used in the subject headings, but not immensely. When they become users, the level of interpretation of the user dominates their interpretation of the intermediary. It would be wise to take into account the lessons they might learn from observing their own and other people's tagging behavior when returning to the workplace.

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.

Bibliography

- Berners-Lee, T.; Hendler, J. & Lassila, O. The Semantic Web. In: *Scientific American*, May, 2001, http://www.sciam.com/article.cfm?id=the-semantic-web, 12 May 2008

- Blachly, A. LibraryThing Press Information. http://www.librarything.com/press/, 27 April 2008

- Blachly, A. Tagging meets Subject Headings. In: *Thing-ology Blog*, May 14 2006, http://www.librarything.com/thingology/2006_05_01_archive.php, 7 July 2008

- Boyd, D. M., & Ellison, N. B. Social network sites: Definition, history, and scholarship. In: *Journal of Computer-Mediated Communication*, Vol 13, Issue 1, article 11, 2007, http://jcmc.indiana.edu/vol13/issue1/boyd.ellison.html, 12 May 2008

- Buranarach, M. A framework for the organization and discovery of information resources in a WWW environment using association, classification and deduction. PhD Thesis, University of Pittsburgh, 2004

- Caplan, P. Metadata fundamentals for all librarians. Chicago: American Library Association, 2003

- Cattuto, C. Semiotic dynamics in online social communities. In: *The European Physical Journal C,* Vol. 46, Nr. 2, 2006, p. 35;

http://www3.isrl.uiuc.edu/~junwang4/langev/localcopy/pdf/cattuto06semioticDynamicsEPJC.pdf, 1 April 2008

- Cattuto, C.; Loreto, V.; Pietronero, L. Semiotic dynamics and collaborative tagging. In: *PNAS*, Vol. 104, Nr. 5, pp. 1461-1464, January 2007, http://dx.doi.org/10.1073/pnas.0610487104, 12 April 2008

- Chi, E. & Mytkowicz, T. Understanding navigability of social tagging systems. 2007, http://www.viktoria.se/altchi/submissions/submission_edchi_0.pdf, 31 January 2008

- Chi, E. & Mytkowicz, T. Understanding the Efficiency of Social Tagging Systems using Information Theory. In *Proceedings of ACM Conference on Hypertext 2008*, Pittsburgh : ACM Press, 2008 (to appear), http://www-users.cs.umn.edu/~echi/papers/2008-hypertext/2008-04-29-hypertext08-tagging-info-theory-fp-046-chi.pdf, 15 July 2008

- Chowdhury, G.G.; Burton, P.F.; McMenemy, D. & Poulter, A. Librarianship: an introduction. London: Facet Publishing, 2008

- Chu, H. Information representation and retrieval in the Digital Age. Medford: Information Today, Inc. for the American Society for Information Science and Technology, 2003

- Coleman, A. Brainstorming Topic Maps. Concept Maps For ADEPT - plans for a NSF-funded workshop, In: *Alexandria Digital Library*, University of California, 13 September 2001, http://alexandria.sdc.ucsb.edu/~acoleman/tmaps.html, 12 May 2008

- Committee on Descriptive Standards, ISAD(G): General Standard International Archival Description. Stockholm: ICA, 2000

- Davis, I. Why tagging is expensive. In: *Panlibus, Talis Corporate Blog*, 7 September 2005, http://blogs.talis.com/panlibus/archives/2005/09/why_tagging_is_.php, 11 July 2008

- Day, M. Metadata in a nutshell. In: *Information Europe* 6(2), Summer 2001, p. 11, Draft, http://www.ukoln.ac.uk/metadata/publications/nutshell/, 14 March 2008

- Delphi Group, Taxonomy & content classification. A Delphi Group White Paper. 2002, http://www.delphigroup.com/research/whitepapers/WP_2002_TAXONOMY.PDF, 20 November 2007

- Dempsey, L. & Heery, R. A review of metadata: a survey of current resource description formats. Bath: UKOLN, University of Bath, 1999, http://www.ukoln.ac.uk/metadata/desire/overview/, 19 March 2008

- Frankowski, D. Data-Mining-AssociationRules-0.10. CPAN, 7 February 2004, http://search.cpan.org/~dfrankow/Data-Mining-AssociationRules-0.10/, 29 July 2008

- Furrie, B. Understanding MARC Bibliographic: Machine-Readable Cataloging. Washington: Library of Congress, Seventh Edition, 2003; see also: http://www.loc.gov/marc/umb/

- Gantz, J. et al, The expanding digital universe. A forecast of worldwide information growth through 2010. IDC White Paper, March 2007, http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf, 5 June 2008

- Garshol, L.M. Metadata? Thesausi? Taxonomies? Topic Maps! Making sense of it all. Ontopia, 26 October 2004, http://www.ontopia.net/topicmaps/materials/tm-vs-thesauri.html, 21 March 2008

- 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

- Guy, M. & Tonkin, M. Folksonomies: Tidying up tags? In: *Dlib Magazin*, Vol. 12, Nr. 1, January 2006, http://www.dlib.org/dlib/january06/guy/01guy.html, 2 April 2008

- Halpin, H.; Robu, V.; Shepherd, H. The complex dynamics of collaborative tagging. In: *Proceedings of the 16 International World Wide Web Conference. Banff, Canada*, 2007, pp. 211-220, http://www2007.org/papers/paper635.pdf, 17 July 2008

- Heymann, P. & Garcia-Molina, H. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Stanford: Stanford University, Technical Report, 2006, http://heymann.stanford.edu/taghierarchy.html, 15 February 2008

- Hillmann, D. Using Dublin Core. DCMI, 11 July 2005, http://dublincore.org/documents/usageguide/, 19 March 2008

- Hotho, A.; Jäschke, R.; Schmitz, C.; Stumme, G. Trend detection in folksonomies. In: *Proceedings of the 1st Conference on Semantics And Digital Media Technology*, 2006, http://www.kde.cs.uni-kassel.de/stumme/papers/2006/hotho2006trend.pdf, 12 July 2008

- IDA, Model Requirements for the management of electronic records. MoReq Specification. Brussel: CECA-CEE-CEEA, 2001; see also: http://www.cornwell.co.uk/edrm/moreq.asp#moreqdownload

- ISO 15836:2003, Information and documentation – The Dublin Core Metadata Element Set

- ISO 2709:1996 - Format for Bibliographic Information Interchange on Magnetic Tape

- ITS.MARC, MARC 21 Format for Bibliographic Data. The Library Corporation, 2001, http://www.itsmarc.com/crs/Bib0001.htm, 19 March 2008

- ITS.MARC, MARC 21 Formats. The Library Corporation, 2001, http://www.itsmarc.com/crs/gen0001.htm, 19 March 2008

- Jacob, E. Classification and categorization: a difference that makes a difference. In: *Library Trends*, Winter 2004, http://findarticles.com/p/articles/mi_m1387/is_3_52/ai_n6080402/print, 17 May 2008

- Jannsens, G. & Put, E. Geschiedenis, principles en terminologie van de archivistiek. Onuitgegeven syllabus, Vrije Universiteit Brussel, 2005-2006

- Jäschke, R.; Marinho, L.; Hotho, A.; Schmidt-Thieme, L.; Stumme, G. Tag recommendations in folksonomies. In: A. Hinneburg, (ed.), *Workshop Proceedings of Lernen - Wissensentdeckung - Adaptivität (LWA 2007)*, 2007, pp. 13-20, http://www.kde.cs.uni-kassel.de/stumme/papers/2007/jaeschke07tagrecommendationsKDML.pdf, 12 July 2008

- Kipp, M. E. & Campbell, G. D. Patterns and inconsistencies in collaborative tagging systems: An examination of tagging practices. In: *Proceedings American Society for Information Science and Technology,* Silver Spring: ASIS&T, 2006, http://dlist.sir.arizona.edu/1704/, 13 February 2008

- Kipp, M. Exploring the context of user, creator and intermediary tagging. In: *Proceedings of the Information Architecture Summit, March 23-27, 2006, Vancouver; Canada*, Silver Spring: ASIS&T, 2006, http://www.iasummit.org/2006/files/109_Presentation_Desc.pdf, 28 February 2008

- Lambe, P. Organising knowledge: taxonomies, knowledge and organisational effectiveness. Oxford: Chandos Publishing Limited, 2007

- Lambiotte, R. & Ausloos, M. Collaborative tagging as a tripartite network. *Lecture Notes in Computer Science*, 3993:1114–1117, Liège: SUPRATECS, Université de Liège, 2005, http://arxiv.org/PS_cache/cs/pdf/0512/0512090v2.pdf, 15 July 2008

- Lau, A. Burning down the shelf: standardized classification, folksonomies and ontological politics. In: Paul Carbone, Christopher Collins, Stacey Meeker (eds.), *InterActions: UCLA Journal of Education and Information Studies*, Vol. 4, Issue 1, University of California, 2008, http://repositories.cdlib.org/gseis/interactions/vol4/iss1/art4/, 13 March 2008

- Leise, F.; Fast, K. & Steckel, M. What is a controlled vocabulary? In: *Boxes and Arrows*, December Issue 2002, http://www.boxesandarrows.com/view/what_is_a_controlled_vocabulary_, 26 March 2008

- Lesk, M. Understanding digital libraries. San Francisco: Elsevier, 2005, 2nd ed.

- Library of Congress, Development of the Encoded Archival Description DTD. Encoded Archival Description, Version 2002, http://www.loc.gov/ead/eaddev.html, 19 March 2008

- LibraryThing concepts. In: *LibraryThing.com*, http://www.librarything.com/concepts# what, 27 April 2008

- Lund, B.; Hammond, T.; Flack, M.; Hannay, T. Social bookmarking tools (II): A case study – Connotea. In: *Dlib Magazine*, Vol. 11, Nr. 4, April 2005, http://www.dlib.org/dlib/april05/lund/04lund.html, 14 July 2008

- Lux, M.; Granitzer, M. & Kern, R. Aspects of broad folksonomies. In: *Proceedings of the 18th International Conference on Database and Expert Systems Applications (DEXA 2007)*, Washington, DC: IEEE Computer Society, 2007, pp. 283-287, http://www.uni-weimar.de/medien/webis/research/tir/tir-07/proceedings/lux07-aspects-of-broad-folksonomies.pdf, 14 July 2008

- Maass, W.; Kowatsch, T.; Münster, T. Vocabulary patterns in free-for-all collaborative indexing systems. In: Luke Liming Chen, Philippe Cudré-Mauroux, Peter Haase, Andreas Hotho, Ernie Ong (eds.), *Proceedings of the First International Workshop on Emergent Semantics and Ontology Evolution (ESOE-2007), Busan, Korea,* November 2007, http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-292/paper6.pdf, 14 July 2008

- Macgregor, G. & McCulloch, E. Collaborative tagging as a knowledge and organization and resource discovery tool. In: *Library Review*, Vol.

55, N° 5, Preprint, 2006, http://eprints.rclis.org/archive/00005703/, 11 July 2008

- MARBI, The MARC 21 Formats: background and principles. November 1996, http://www.loc.gov/marc/96principl.html, 19 March 2008

- MARC Standards. In: *Wikipedia*, http://en.wikipedia.org/wiki/MARC_standards, 19 March 2008

- Mathes, A. Folksonomies - Cooperative Classification and Communication Through Shared Metadata. Personal weblog, December 2004, http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html, 26 January 2008

- Metadata. In: Mary S. Woodley (ed.), *DCMI Glossary*, 23 April 2004, http://dublincore.org/documents/usageguide/glossary.shtml#M, 14 March 2008

- Metadata. UKOLN, University of Bath, http://www.ukoln.ac.uk/metadata/, 14 March 2008

- Mika, P. Ontologies are us: A unified model of social networks and semantics. In: *International Semantic Web Conference*, ser. Lecture Notes in Computer Science, vol. 3729, International Semantic Web Conference 2005, Springer, November 2005, pp. 522-536. http://citeseer.ist.psu.edu/739485.html, 31 January 2008

- Miller, P. Metadata for the masses. In: *Ariadne*, Issue 5, September 1996, http://www.ariadne.ac.uk/issue5/metadata-masses/, 19 March 2008

- Moore, G. & Ahmed, K. An introduction to Topic Maps. In: *The Microsoft Architecture Journal*, July 2005,

http://msdn.microsoft.com/en-us/library/aa480048.aspx, 12 May 2008

- Morrison, P.J. Tagging and searching: search retrieval effectiveness of folksonomies on the web. In: *Internet News,* November 27, 2007, http://www.websearchguide.ca/netblog/archives/006823.html, 10 July 2008

- Morville, P. Ambient findability. Sebastopol: O'Reilly Media Inc., 2005

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

- Pepper, S. The TAO of Topic Maps. Finding the way in the age of infoglut. April 2002, http://www.ontopia.net/topicmaps/materials/tao.html, 12 May 2008

- Peters, I. & Stock, W.G. Folksonomies in Wissensrepräsentation und Information Retrieval. In: *Information – Wissenschaft und Praxis,* Vol. 59, Nr. 2, 2008, pp. 77-90, http://www.phil-fak.uni-duesseldorf.de/infowiss/admin/public_dateien/files/56/1204547947stock212_h.htm, 12 April 2008

- Pitti, D. Encoded Archival Description. An introduction and overview. In: *D-Lib Magazine*, Volume 5, Number 11, 1999, http://www.dlib.org/dlib/november99/11pitti.html, 19 March 2008

- Quintarelli, E. Folksonomies: power to the people. paper presented at the ISKO Italy-UniMIB meeting : Milan : June 24, 2005, http://www.iskoi.org/doc/folksonomies.htm #broad, 29 April 2008

- Rutkoff, A. Social networking for bookworms. In: *The Wall Street Journal Online*, 27 June 2006, http://online.wsj.com/public/article/SB115109622468789252-i8U6LIHU7ChfgbxG1oZ_iunOIWE_20060727.html, 6 July 2008

- Šauperl, A. Catalogers' common ground and shared knowledge. In: *Journal of the American Society for Information Science and Technology*, Vol. 55, Nr. 1, Wilmington: Wiley Periodicals, 2004, pp. 55-63

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

- Shirky, C. Ontology is Overrated: Categories, Links, and Tags. In: *Clay Shirky's Writings About the Internet*, personal weblog, 2005, http://shirky.com/writings/ontology_overrated.html, 22 February 2008

- Shirky, C. Power Laws, Weblogs, and Inequality. In: *Clay Shirky's writings about the internet*, personal weblog, 2 October 2003, http://www.shirky.com/writings/powerlaw_weblog.html, 18 May 2008

- Siepel A. & de Vries, H. Encoded Archival Description (EAD). Het digitaliseren van toegangen op archieven. Amsterdam: Archiefschool, 2001

- Simpson, E. Clustering tags in enterprise and web folksonomies. HP Labs, Technical Report, July 2007, http://www.hpl.hp.com/techreports/2007/HPL-2007-190.html, 27 April 2008

- Sinclair, J. & Cardew-Hall, M. The folksonomy tag cloud: When is it useful? In: Adrian Dale (ed.), *Journal of Information Science*, Thousand Oaks: Sage Publications, February 2008, pp. 15-29

- Sinha, R. A cognitive analysis of tagging (or how the lower cost of tagging makes it popular). In: *rashmisinha.com*, personal weblog, September 27, 2005, http://www.rashmisinha.com/archives/05_09/tagging-cognitive.html, 27 January 2008

- Smith, T. Cataloging and You: Measuring the Efficacy of a Folksonomy for Subject Analysis. In: Joan Lussky (ed.), *Proceedings of the 18th Workshop of the American Society for Information Science and Technology Special Interest Group in Classification Research*, Milwaukee, Wisconsin, 2007, http://dlist.sir.arizona.edu/2061/, 17 July 2008

- Spalding, T. New Feature: Find Friends. In: *LibraryThing Blog, new features, plans and announcements*, June 17, 2008, http://www.librarything.com/blog/2008/06/new-feature-find-friends.php, 26 June 2008

- Spalding, T. Member home pages. In: *LibraryThing Blog, new features, plans and announcements*, June 21, 2008, http://www.librarything.com/blog/2008/06/member-home-pages.php, 26 June 2008

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

- Spalding, T. Introducing the LibraryThing books API. In: *LibraryThing Blog, new features, plans and announcements*, July 6, 2008, http://www.librarything.com/blog/2008/07/introducing-librarything-books-api.php, 7 July 2008

- Spalding, T. Build the Open Shelves Classification. In: *Thing-ology Blog, meanings, methods and debates*, July 8, 2008, http://www.librarything.com/thingology/2008/07/build-open-shelves-classification.php, 10 July 2008

- Spalding, T. Free Web Services API to Common Knowledge. In: *LibraryThing Blog, new features, plans and announcements*, August 1, 2008, http://www.librarything.com/blog/2008/08/free-web-services-api-to-common.php, 2 August 2008

- Sterken, V. Op zoek naar Vaste Waarden. Vooronderzoek Digitale Archivering aan het Vlaams Parlement, 2005-2006. Master Thesis, Free University Brussels (VUB), 2006

- Sturtz, D. Communal categorization: the folksonomy. Unpublished course paper, Philadelphia: Drexel University, December 16, 2004, http://davidsturtz.com/drexel/622/sturtz-folksonomy.pdf, 10 May 2008

- Svenonius, E. The Intellectual Foundation of Information Organization. Cambridge, MA: MIT Press, 2001

- Tennis, J. T. Social Tagging and the Next Steps for Indexing. In: Furner, J. & Tennis, J. T, (Eds.), *Proceedings of the 17th Workshop of the American Society for Information Science and Technology Special Interest Group in Classification Research* 17, Austin, Texas, 2006, http://dlist.sir.arizona.edu/2091/, 14 July 2008

- Tošić, M. & Milićević, V. The Semantics of Collaborative Tagging System. In: *Proceedings of the 2nd Workshop on Scripting for the Semantic Web*, 2006, http://www.semanticscripting.org/SFSW2006/Paper6.pdf, 14 July 2008

- Trant, J. Exploring the potential for social tagging and folksonomy in art museums: proof of concept. A paper for the *New Review of Hypermedia and Multimedia*, Draft, 13 May 2006, http://www.archimuse.com/papers/steve-nrhm-0605preprint.pdf, 13 March 2008

- Trant, J. More steve … tagger prototype preliminary analysis. In: c*onference.archimuse.*com, 16 October 2006, http://conference.archimuse.com/blog/jtrant/more_steve_tagger_prototype_preliminary_analysis, 21 July 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, http://www.archimuse.com/papers/asist-CR-steve-0611.pdf, 21 July 2008

- Van Damme, C. Folksonomies and entreprise folksonomies. Unpublished master thesis, Vrije Universiteit Brussel, 2006

- Van Damme, C.; Hepp, M. & Siorpaes, K. Folksontology: An integrated approach for turning folksonomies into ontologies. In: *Proceedings of the ESWC 2007 Workshop "Bridging the Gap between Semantic Web and Web 2.0"*, 2007, http://www.kde.cs.uni-kassel.de/ws/eswc2007/proc/FolksOntology.pdf, 31 January 2008

- Van Dijck, P. Emergent i18n effects in folksonomies. In: *Peter Van Dijck's Guide to Ease,* personal weblog, 15 January 2005, http://poorbuthappy.com/ease/archives/2005/01/15/2419/multilingual-folksonomies, 12 May 2008

- Vander Wal, T. Folksonomy definition and Wikipedia. In: *vanderwal.net*, November 2, 2005, http://www.vanderwal.net/random/entrysel.php?blog=1750, 11 May 2008

- Vanderwal, T. Explaining and showing broad and narrow folksonomies. In: *Personal InfoCloud*, February 21, 2005, http://www.personalinfocloud.com/2005/02/explaining_and_.html, 27 January 2008

- Vanderwal, T. Folksonomy explanations. In: *vanderwal.net*, January 18, 2005, http://www.vanderwal.net/random/entrysel.php?blog=1622, 18 May 2008

- Vanderwal, T. Folksonomy provides 70 percent more terms than taxonomy. In: *Personal InfoCloud*, June 12, 2007, http://personalinfocloud.com/2007/06/folksonomy_prov.html, 15 July 2008

- Voorbij, H. Title keywords and subject descriptors: a comparison of subject search entries of books in the Humanities and Social Sciences. In: *Journal of Documentation*, Vol. 54, Nr. 4, September 1998, pp. 466-476

- Weber, J. Folksonomy and controlled vocabulary in LibraryThing. Unpublished Final Project, University of Pittsburgh, 2006, p. 6;

http://dystmesis.net/2006/11/17/tags-and-subject-headings/, 16 March 2008

- Weibel, S.; Godby, J.; Miller, E.; Daniel, R. OCLC/NCSA Metadata Workshop Report. DCMI, 1995, http://dublincore.org/workshops/dc1/report.shtml, 19 March 2008

- Weinberger, D. Everything is miscellaneous. The power of the new digital disorder. New York: Times Books, Henry Holt & Company, LLC, 2007

References

Chapter 1: Classification and its contents

[1] Van Damme, C. Folksonomies and entreprise folksonomies. Unpublished master thesis, Vrije Universiteit Brussel, 2006, p. 8
[2] Gantz, J. 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, 5 June 2008
[3] Buranarach, M. A framework for the organization and discovery of information resources in a WWW environment using association, classification and deduction. PhD Thesis, University of Pittsburgh, 2004, p. 7
[4] Miller, P. Metadata for the masses. In: *Ariadne*, Issue 5, September 1996, http://www.ariadne.ac.uk/issue5/metadata-masses/, 19 March 2008
[5] Caplan, P. Metadata fundamentals for all librarians. Chicago: American Library Association, 2003, p. 1
[6] Metadata. In: Mary S. Woodley (ed.), *DCMI Glossary*, 23 April 2004, http://dublincore.org/documents/usageguide/glossary.shtml#M, 14 March 2008
[7] Day, M. Metadata in a nutshell. In: *Information Europe* 6(2), Summer 2001, p. 11, Draft, http://www.ukoln.ac.uk/metadata/publications/nutshell/, 14 March 2008
[8] Metadata. UKOLN, University of Bath, http://www.ukoln.ac.uk/metadata/, 14 March 2008
[9] Weibel, S.; Godby, J.; Miller, E.; Daniel, R. OCLC/NCSA Metadata Workshop Report. DCMI, 1995, http://dublincore.org/workshops/dc1/report.shtml, 19 March 2008
[10] ISO 15836:2003, Information and documentation – The Dublin Core Metadata Element Set
[11] Sterken, V. Op zoek naar Vaste Waarden. Vooronderzoek Digitale Archivering aan het Vlaams Parlement, 2005-2006. Master Thesis, Free University Brussels (VUB), 2006, p. 33
[12] Dempsey, L. & Heery, R. A review of metadata: a survey of current resource description formats. Bath: UKOLN, University of Bath, 1999, http://www.ukoln.ac.uk/metadata/desire/overview/rev_05.htm, 19 March 2008
[13] About the Initiative. DCMI, http://dublincore.org/about/, 19 March 2008
[14] Hillmann, D. Using Dublin Core. DCMI, 11 July 2005, http://dublincore.org/documents/usageguide/#whatis, 19 March 2008
[15] ITS.MARC, MARC 21 Format for Bibliographic Data. The Library Corporation, 2001, http://www.itsmarc.com/crs/Bib0001.htm, 19 March 2008
[16] Furrie, B. Understanding MARC Bibliographic: Machine-Readable Cataloging. Washington: Library of Congress, Seventh Edition, 2003; see also: http://www.loc.gov/marc/umb/
[17] ITS.MARC, MARC 21 Formats. The Library Corporation, 2001, http://www.itsmarc.com/crs/gen0001.htm, 19 March 2008
[18] MARC Standards. In: *Wikipedia*, http://en.wikipedia.org/wiki/MARC_standards, 19 March 2008
[19] ITS.MARC, *Idem*, http://www.itsmarc.com/crs/gen0011.htm, 19 March 2008
[20] Library of Congress, Development of the Encoded Archival Description DTD. Encoded Archival Description, Version 2002, http://www.loc.gov/ead/eaddev.html, 19 March 2008
[21] Siepel A. & de Vries, H. Encoded Archival Description (EAD). Het digitaliseren van toegangen op archieven. Amsterdam: Archiefschool, 2001, p. 6
[22] Pitti, D. Encoded Archival Description. An introduction and overview. In: *D-Lib Magazine*, Volume 5, Number 11, 1999, http://www.dlib.org/dlib/november99/11pitti.html, 19 March 2008
[23] Committee on Descriptive Standards, ISAD(G): General Standard International Archival Description. Stockholm: ICA, 2000
[24] Committee on Descriptive Standards, *Idem*, p. 7
[25] Garshol, L.M. Metadata? Thesausi? Taxonomies? Topic Maps! Making sense of it all. Ontopia, 26 October 2004 http://www.ontopia.net/topicmaps/materials/tm-vs-thesauri.html#N412, 21 March 2008

[26] Chu, H. Information representation and retrieval in the Digital Age. Medford: Information Today, Inc. for the American Society for Information Science and Technology, 2003, p. 47
[27] Leise, F.; Fast, K. & Steckel, M. What is a controlled vocabulary? In: *Boxes and Arrows*, December Issue 2002, http://www.boxesandarrows.com/view/what_is_a_controlled_vocabulary_, 26 March 2008
[28] Garshol, L.M. *Idem*
[29] Chu, *Idem*, p. 49
[30] Chu, *Idem*, p. 50
[31] Chowdhury, G.G.; Burton, P.F.; McMenemy, D. & Poulter, A. Librarianship: an introduction. London: Facet Publishing, 2008, p. 91
[32] Chowdhury et. al. *Idem*, pp. 92-93; Rowley, J. & Farrow, J. Organizing knowledge. An introduction to managing access to information. Hampshire: Gower Publishing Limited, 2000, 3rd ed., pp. 195-201
[33] Delphi Group, Taxonomy & content classification. A Delphi Group White Paper. 2002, p. 15, http://www.delphigroup.com/research/whitepapers/WP_2002_TAXONOMY.PDF, 20 November 2007
[34] Lesk, M. Understanding digital libraries. San Francisco: Elsevier, 2005, 2nd ed., p. 52
[35] Chu, H. *Idem*, p. 49
[36] Chowdhury, G.G. et al, *Idem*, pp. 115-116
[37] Chu, *Idem*, p. 50
[38] Weinberger, D. Everything is miscellaneous. The power of the new digital disorder. New York: Times Books, Henry Holt & Company, LLC, 2007, p. 72
[39] Garshol, L.M., *Idem*
[40] Van Damme, C. *Op. Cit.*, p. 43
[41] Dethier, H. Het gezicht en het raadsel. Profielen van Plato tot Derrida. Brussel, VUBPress, 1995, 2de druk, pp. 72-73
[42] Lau, A. Burning down the shelf: standardized classification, folksonomies and ontological politics. In: Paul Carbone, Christopher Collins, Stacey Meeker (eds.), *InterActions: UCLA Journal of Education and Information Studies*, Vol. 4, Issue 1, University of California, 2008, http://repositories.cdlib.org/gseis/interactions/vol4/iss1/art4/, 13 March 2008
[43] Delphi Group, *Ibidem*, p. 24
[44] Coleman, A. Brainstorming Topic Maps. Concept Maps For ADEPT - plans for a NSF-funded workshop, In: *Alexandria Digital Library*, University of California, 13 September 2001, http://alexandria.sdc.ucsb.edu/~acoleman/tmaps.html, 12 May 2008
[45] Pepper, S. The TAO of Topic Maps. Finding the way in the age of infoglut. April 2002, http://www.ontopia.net/topicmaps/materials/tao.html, 12 May 2008
[46] Kay, R. QuickStudy: Topic Maps. In: *Computerworld,* IDG, October 4, 2004, p. 2 http://www.computerworld.com/news/2004/story/0,11280,96350,00.html, 12 May 2008
[47] Moore, G. & Ahmed, K. An introduction to Topic Maps. In: *The Microsoft Architecture Journal*, July 2005, http://msdn.microsoft.com/en-us/library/aa480048.aspx, 12 May 2008
[48] Noruzi, A. Folksonomies: why do we need controlled vocabularies? In: Alireza Noruzi (ed.), *Webology*, Vol. 4, Nr. 2, June 2007, http://www.webology.ir/2007/v4n2/editorial12.html, 20 March 2008
[49] Chowdhury et. al. *Ibidem*, p. 100
[50] Vander Wal, T. Folksonomy definition and Wikipedia. In: *vanderwal.net*, personal weblog November 2, 2005, http://www.vanderwal.net/random/entrysel.php?blog=1750, 11 May 2008
[51] Van Dijck, P. Emergent i18n effects in folksonomies. In: *Peter Van Dijck's Guide to Ease,* personal weblog, 15 January 2005, http://poorbuthappy.com/ease/archives/2005/01/15/2419/multilingual-folksonomies, 12 May 2008
[52] 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, 10 May 2008
[53] Berners-Lee, T.; Hendler, J. & Lassila, O. The Semantic Web. In: *Scientific American*, May, 2001, p. 2, http://www.sciam.com/article.cfm?id=the-semantic-web&page=2, 12 May 2008
[54] W3C Semantic Web Homepage, http://www.w3.org/2001/sw/, 12 May 2008
[55] Weinberger, D. *Ibidem*, pp. 192-193

[56] Semantic Web. In: *Wikipedia*, http://en.wikipedia.org/wiki/Semantic_Web, 12 May 2008
[57] Weinberger, D. *Idem*, p. 195

Chapter 2: Folksonomies

[1] Lesk, M. Understanding digital libraries. San Francisco: Elsevier, 2005, 2nd ed., p. 50
[2] Shirky, C. Ontology is Overrated: Categories, Links, and Tags. In: *Clay Shirky's Writings About the Internet*, personal weblog, 2005, http://shirky.com/writings/ontology_overrated.html, 22 February 2008
[3] Weinberger, D. Everything is miscellaneous. The power of the new digital disorder. New York: Times Books, Henry Holt & Company, LLC, 2007, pp. 17-19
[4] Weinberger, D. *Ibidem*, p. 32
[5] Weinberger, D. *Ibidem*, pp. 46-55
[6] Weinberger, D. *Ibidem*, pp. 69-70
[7] Shirky, C. *Idem*
[8] Scheelings, F. Statisch en dynamisch archiefbeheer van administraties en bedrijven. Unpublished lectures, Brussel: VUB, 2005
[9] Mathes, A. Folksonomies - Cooperative Classification and Communication Through Shared Metadata. Personal weblog, December 2004, http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html, 26 January 2008
[10] Jacob, E. Classification and categorization: a difference that makes a difference. In: *Library Trends*, Winter 2004, http://findarticles.com/p/articles/mi_m1387/is_3_52/ai_n6080402/print, 17 May 2008
[11] Jacob, E. *Idem*
[12] Jacob, E. *Idem*
[13] Personal e-mail received on 18 May 2008
[14] Sinha, R. A cognitive analysis of tagging (or how the lower cost of tagging makes it popular). In: *rashmisinha.com*, personal weblog, September 27, 2005, http://www.rashmisinha.com/archives/05_09/tagging-cognitive.html, 27 January 2008
[15] Sinha, R. *Idem*
[16] Vanderwal, T. Folksonomy explanations. In: *vanderwal.net*, personal weblog, January 18, 2005, http://www.vanderwal.net/random/entrysel.php?blog=1622, 18 May 2008
[17] Van Damme, C. *op. cit.*, p. 33
[18] Reid, R. D. & Sanders, N. Operations management. An integrated approach. Hoboken: John Wiley & Sons, Inc., 2007, 3rd ed., p. 153
[19] Shirky, C. Power Laws, Weblogs, and Inequality. In: *Clay Shirky's writings about the internet*, personal weblog, 2 October 2003, http://www.shirky.com/writings/powerlaw_weblog.html, 18 May 2008
[20] Vanderwal, T. Explaining and showing broad and narrow folksonomies. In: *Personal InfoCloud*, February 21, 2005, http://www.personalinfocloud.com/2005/02/explaining_and_.html, 27 January 2008
[21] Chu, H. Information representation and retrieval in the Digital Age. Medford: Information Today, Inc. for the American Society for Information Science and Technology, 2003, pp. 53-55
[22] Golder, S. & Huberman, B. The structure of collaborative tagging systems. Information Dynamics Labs, Hewlett-Packard, 2005, p. 2, http://arxiv.org/ftp/cs/papers/0508/0508082.pdf, February 13, 2008
[23] Golder, S. & Huberman, B. *Idem*, pp. 2-3
[24] Chu, H. *Idem*, p. 53
[25] Golder, S. & Huberman, B. p. 2
[26] Chu, H. *Idem*, p. 54
[27] Morrison, P.J. Tagging and searching: search retrieval effectiveness of folksonomies on the web. In: *Internet News*, November 27, 2007, http://www.websearchguide.ca/netblog/archives/006823.html, 10 July 2008
[28] Mathes, A. *Idem*
[29] Mathes, A. *Idem*

[30] Shirky, C. Ontology is Overrated: Categories, Links, and Tags. In: *Clay Shirky's Writings About the Internet*, personal weblog, 2005, http://shirky.com/writings/ontology_overrated.html, 22 February 2008
[31] Macgregor, G. & McCulloch, E. Collaborative tagging as a knowledge and organization and resource discovery tool. In: *Library Review*, Vol. 55, N° 5, Preprint, 2006, http://eprints.rclis.org/archive/00005703/, 11 July 2008
[32] Davis, I. Why tagging is expensive. In: *Panlibus, Talis Corporate Blog*, 7 September 2005, http://blogs.talis.com/panlibus/archives/2005/09/why_tagging_is_.php, 11 July 2008
[33] Mathes, A. *Idem*; Quintarelli, E. Folksonomies: power to the people. paper presented at the ISKO Italy-UniMIB meeting : Milan : June 24, 2005, http://www.iskoi.org/doc/folksonomies.htm #broad, 29 April 2008
[34] Guy, M. & Tonkin, M. Folksonomies: Tidying up tags? In: *Dlib Magazin*, Vol. 12, Nr. 1, January 2006, http://www.dlib.org/dlib/january06/guy/01guy.html, 2 April 2008
[35] Golder, S. & Huberman, B. *Idem*
[36] E.g. Lambiotte, R. & Ausloos, M. Collaborative tagging as a tripartite network. *Lecture Notes in Computer Science*, 3993:1114–1117, Liège: SUPRATECS, Université de Liège, 2005, http:// arxiv.org/PS_cache/cs/pdf/0512/0512090v2.pdf, 15 July 2008; Hotho, A.; Jäschke, R.; Schmitz, C.; Stumme, G. Trend detection in folksonomies. In: *Proceedings of the 1st Conference on Semantics And Digital Media Technology*, 2006, http://www.kde.cs.uni-kassel.de/stumme/papers/2006/hotho2006trend.pdf, 12 July 2008; Tošić, M. & Milićević, V. The Semantics of Collaborative Tagging System. In: *Proceedings of the 2nd Workshop on Scripting for the Semantic Web*, 2006, http://www.semanticscripting.org/SFSW2006/Paper6.pdf, 14 July 2008
[37] Lund, B.; Hammond, T.; Flack, M.; Hannay, T. Social bookmarking tools (II): A case study – Connotea. In: *Dlib Magazine*, Vol. 11, Nr. 4, April 2005, http://www.dlib.org/dlib/april05/lund/04lund.html, 14 July 2008; Maass, W.; Kowatsch, T.; Münster, T. Vocabulary patterns in free-for-all collaborative indexing systems. In: Luke Liming Chen, Philippe Cudré-Mauroux, Peter Haase, Andreas Hotho, Ernie Ong (eds.), *Proceedings of the First International Workshop on Emergent Semantics and Ontology Evolution (ESOE-2007)*, *Busan, Korea,* November 2007, http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-292/paper6.pdf, 14 July 2008; Peters, I. & Stock, W.G. Folksonomies in Wissens-repräsentation und Information Retrieval. In: *Information – Wissenschaft und Praxis,* Vol. 59, Nr. 2, 2008, p. 80, http://www.phil-fak.uniduesseldorf.de/infowiss/admin/public_dateien/files/56/1204547947stock212_h.htm, 12 April 2008
[38] Cattuto, C. Semiotic dynamics in online social communities. In: *The European Physical Journal C,* Vol. 46, Nr. 2, 2006, p. 35; http://www3.isrl.uiuc.edu/~junwang4/langev/localcopy/pdf/cattuto06semioticDynamicsEPJC.pdf, 1 April 2008
[39] Cattuto, C.; Loreto, V.; Pietronero, L. Semiotic dynamics and collaborative tagging. In: *PNAS*, Vol. 104, Nr. 5, pp. 1461-1464, January 2007, http://dx.doi.org/10.1073/pnas.0610487104, 12 April 2008
[40] Lux, M.; Granitzer, M. & Kern, R. Aspects of broad folksonomies. In: *Proceedings of the 18th International Conference on Database and Expert Systems Applications (DEXA 2007)*, Washington, DC: IEEE Computer Society, 2007, pp. 283-287, http://www.uni-weimar.de/medien/webis/research/tir/tir-07/proceedings/lux07-aspects-of-broad-folksonomies.pdf, 14 July 2008
[41] Kipp, M. E. & Campbell, G. D. Patterns and inconsistencies in collaborative tagging systems: An examination of tagging practices. In: *Proceedings American Society for Information Science and Technology,*, 2006, http://dlist.sir.arizona.edu/1704/, 13 February 2008
[42] Chi, E. & Mytkowicz, T. Understanding navigability of social tagging systems. 2007, http://www.viktoria.se/altchi/submissions/submission_edchi_0.pdf, 31 January 2008; Chi, E. & Mytkowicz, T. Understanding the Efficiency of Social Tagging Systems using Information Theory. In *Proceedings of ACM Conference on Hypertext 2008*, Pittsburgh : ACM Press, 2008 (to appear), http://www-users.cs.umn.edu/~echi/papers/2008-hypertext/2008-04-29-hypertext08-tagging-info-theory-fp-046-chi.pdf, 15 July 2008
[43] Van Damme, C. Folksonomies and entreprise folksonomies. Unpublished master thesis, Vrije Universiteit Brussel, 2006, pp. 47-58

[44] See also: Jäschke, R.; Marinho, L.; Hotho, A.; Schmidt-Thieme, L.; Stumme, G. Tag recommendations in folksonomies. In: A. Hinneburg, (ed.), *Workshop Proceedings of Lernen - Wissensentdeckung - Adaptivität (LWA 2007)*, 2007, pp. 13-20, http://www.kde.cs.uni-kassel.de/stumme/papers/2007/jaeschke07tagrecommendationsKDML.pdf, 12 July 2008
[45] Van Damme, C. *Idem*, pp. 53-55
[46] Heymann, P. & Garcia-Molina, H. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Stanford: Stanford University, Technical Report, 2006, http://heymann.stanford.edu/taghierarchy.html, 15 February 2008
[47] Simpson, E. Clustering tags in enterprise and web folksonomies. HP Labs, Technical Report, July 2007, http://www.hpl.hp.com/techreports/2007/HPL-2007-190.html, 27 April 2008
[48] Mika, P. Ontologies are us: A unified model of social networks and semantics. In: *International Semantic Web Conference*, ser. Lecture Notes in Computer Science, vol. 3729, International Semantic Web Conference 2005, Springer, November 2005, pp. 522-536. http://citeseer.ist.psu.edu/739485.html, 31 January 2008; Van Damme, C.; Hepp, M. & Siorpaes, K. Folksontology: An integrated approach for turning folksonomies into ontologies. In: *Proceedings of the ESWC 2007 Workshop "Bridging the Gap between Semantic Web and Web 2.0"*, 2007, http://www.kde.cs.uni-kassel.de/ws/eswc2007/proc/FolksOntology.pdf, 31 January 2008

Chapter 3: Analysis

[1] Svenonius, E. The Intellectual Foundation of Information Organization. Cambridge, MA: MIT Press, 2001, p. 26; as cited in: Smith, T. Cataloging and You: Measuring the Efficacy of a Folksonomy for Subject Analysis. In: Joan Lussky (ed.), *Proceedings of the 18th Workshop of the American Society for Information Science and Technology Special Interest Group in Classification Research*, Milwaukee, Wisconsin, 2007, p. 2, http://dlist.sir.arizona.edu/2061/, 17 July 2008
[2] Mathes, A. Folksonomies - Cooperative Classification and Communication Through Shared Metadata. Personal weblog, December 2004, http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html, 26 January 2008
[3] Šauperl, A. Catalogers' common ground and shared knowledge. In: *Journal of the American Society for Information Science and Technology*, Vol. 55, Nr. 1, Wilmington: Wiley Periodicals, 2004, pp. 55-63
[4] Šauperl, A. *Idem*, p. 56
[5] Šauperl, A. *Idem*, pp. 61-63
[6] Tennis, J. T. Social Tagging and the Next Steps for Indexing. In: Furner, J. & Tennis, J. T, (Eds.), *Proceedings of the 17th Workshop of the American Society for Information Science and Technology Special Interest Group in Classification Research* 17, Austin, Texas, 2006, p. 4, http://dlist.sir.arizona.edu/2091/, 14 July 2008
[7] Blachly, A. LibraryThing Press Information. http://www.librarything.com/press/, 27 April 2008
[8] Spalding, T. Member home pages. In: *LibraryThing Blog, new features, plans and announcements*, June 21, 2008, http://www.librarything.com/blog/2008/06/member-home-pages.php, 26 June 2008; Spalding, T. New Feature: Find Friends. In: *LibraryThing Blog, new features, plans and announcements*, June 17, 2008, http://www.librarything.com/blog/2008/06/new-feature-find-friends.php, 26 June 2008; Spalding, T. Build the Open Shelves Classification. In: *Thing-ology Blog, meanings, methods and debates*, July 8, 2008, http://www.librarything.com/thingology/2008/07/build-open-shelves-classification.php, 10 July 2008; Spalding, T. Introducing the LibraryThing books API. In: *LibraryThing Blog, new features, plans and announcements*, July 6, 2008, http://www.librarything.com/blog/2008/07/introducing-librarything-books-api.php, 7 July 2008; Spalding, T. Free Web Services API to Common Knowledge. In: *LibraryThing Blog, new features, plans and announcements*, August 1, 2008, http://www.librarything.com/blog/2008/08/free-web-services-api-to-common.php, 2 August 2008
[9] Spalding, T. Presentation during the panel *Creating the future of the Catalog and Cataloging*, ALA Annual Conference, Anaheim, June 29, 2008; See: http://www.librarything.com/thingology/2008/07/future-of-cataloging.php, 5 July 2008

[10] LibraryThing concepts. In: *LibraryThing.com*, http://www.librarything.com/concepts#what, 27 April 2008
[11] Blachly, A. *Idem*
[12] Rutkoff, A. Social networking for bookworms. In: *The Wall Street Journal Online*, 27 June 2006, http://online.wsj.com/public/article/SB115109622468789252-i8U6LIHU7ChfgbxG1oZ_iunOIWE_20060727.html, 6 July 2008
[13] Spalding, T. *Idem*
[14] Weber, J. Folksonomy and controlled vocabulary in LibraryThing. Unpublished Final Project, University of Pittsburgh, 2006, p. 6; http://dystmesis.net/2006/11/17/tags-and-subject-headings/, 16 March 2008
[15] Thill, S. DJ Spooky: how a tiny Caribbean island birthed the mashup. In: *Wired Magazine*, 7 December 2007, http://www.wired.com/entertainment/music/news/2007/07/spooky_QA, 29 April 2008
[16] Hill, M. & Dudley, J. Do the mixed-up movie mash. In: *The Courier Mail,* Brisbane, Australia, 9 March 2006, p.17
[17] Mashup (web application hybrid). In: *Wikipedia*, http://en.wikipedia.org/wiki/Mashup_%28web_application_hybrid%29, 29 April2008
[18] Quintarelli, E. Folksonomies: power to the people. paper presented at the ISKO Italy-UniMIB meeting : Milan : June 24, 2005, http://www.iskoi.org/doc/folksonomies.htm#broad, 29 April 2008
[19] 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
[20] Golder, S. & Huberman, B., *Idem*
[21] Golder, S. & Huberman, B., *Idem*
[22] Golder, S. & Huberman, B., *Idem*
[23] Golder, S. & Huberman, B., *Idem*
[24] Golder, S. & Huberman, B., *Idem*
[25] Golder, S. & Huberman, B., *Idem*
[26] Golder, S. & Huberman, B., *Idem*
[27] Sen et. al. Tagging, communities, vocabulary, evolution. In: *Proceedings of CSCW'06, November 4-8, 2006, Banff, Alberta Canada*, http://www-users.cs.umn.edu/~cosley/research/papers/sen-cscw2006.pdf, 1 April 2008
[28] Sen et. al. *Idem*, pp. 4-5
[29] Spalding, T. The Long Tail of Ann Coulter. In: *Thing-ology Blog, meaning, methods and debate*, May 7, 2008, http://www.librarything.com/thingology/2008/05/long-tail-of-ann-coulter.php, 8 May 2008
[30] Halpin, H.; Robu, V.; Shepherd, H. The complex dynamics of collaborative tagging. In: *Proceedings of the 16 International World Wide Web Conference. Banff, Canada*, 2007, p. 212, http://www2007.org/papers/paper635.pdf, 17 July 2008
[31] 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/, 16 March 2008
[32] Blachly, A. Tagging meets Subject Headings. In: *Thing-ology Blog*, May 14 2006, http://www.librarything.com/thingology/2006_05_01_archive.php, 7 July 2008
[33] Spalding, T. Presentation during the panel *Creating the future of the Catalog and Cataloging*, ALA Annual Conference, Anaheim, June 29, 2008
[34] FAQ, steve: the museum social tagging project, http://steve.museum/index.php?option=com_content&task=blogsection&id=6&Itemid=15, 21 July 2008; See also: Trant, J. Exploring the potential for social tagging and folksonomy in art museums: proof of concept. A paper for the *New Review of Hypermedia and Multimedia*, Draft, 13 May 2006, http://www.archimuse.com/papers/steve-nrhm-0605preprint.pdf, 13 March 2008
[35] Trant, J. More steve … tagger prototype preliminary analysis. In: c*onference.archimuse.*com, 16 October 2006, http://conference.archimuse.com/blog/jtrant/more_steve_tagger_prototype_preliminary_analysis, 21 July 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, 21 July 2008

[36] Vanderwal, T. Folksonomy provides 70 percent more terms than taxonomy. In: *Personal InfoCloud*, June 12, 2007, http://personalinfocloud.com/2007/06/folksonomy_prov.html, 15 July 2008

[37] Sinclair, J. & Cardew-Hall, M. The folksonomy tag cloud: When is it useful? In: Adrian Dale (ed.), *Journal of Information Science*, Thousand Oaks: Sage Publications, February 2008, pp. 15-29

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

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