Folksonomies: (Un)Controlled Vocabulary?

Alireza Noruzi

University of Tehran, Department of Library and Information Science, Tehran, Iran, <nouruzi@gmail.com>

Alireza Noruzi is a faculty member of the Department of Library and Information Science, University of Tehran, Iran. He received his Ph.D. in Information Science from the University of Paul Cezanne in France. He also received his M.A. degree in Information Science from the University of Tehran in 2001 and B.A. degree in Library and Information Science from Shiraz University in 1998. He is also the Editor-in-chief of *Webology*, an international open access journal. Alireza Noruzi can be contacted at: nouruzi@gmail.com or http://nouruzi.googlepages.com.

Alireza Noruzi. **Folksonomies: (Un)Controlled Vocabulary?** *Knowledge Organization, 33(4)* 199-203. 14 references.

ABSTRACT: Folksonomy, a free-form tagging, is a user-generated classification system of web contents that allows users to tag their favorite web resources with their chosen words or phrases selected from natural language. These tags (also called concepts, categories, facets or entities) can be used to classify web resources and to express users' preferences. Folksonomy-based systems allow users to classify web resources through tagging bookmarks, photos or other web resources and saving them to a public web site like *Del.icio.us*. Thus information about web resources and online articles can be shared in an easy way. The purpose of this study is to provide an overview of the *folksonomy tagging* phenomenon (also called *social tagging* and *social bookmarking*) and explore some of the reasons why we need controlled vocabularies, discussing the problems associated with folksonomy.

1. What is a Folksonomy?

A *folksonomy* is an Internet-based information retrieval methodology consisting of collaboratively generated, open-ended labels that categorize content such as web resources, online photographs, and web links. A folksonomy is most notably contrasted with a *taxonomy*, in that the authors of the labeling system are often the main users (and sometimes originators) of the content to which the labels are applied. The labels are commonly known as tags (also called categories or facets) and the labeling process is called tagging (Folksonomy 2007). Tags help to improve search engine effectiveness because content is categorized using a familiar, accessible, and shared vocabulary.

Tags are words or phrases users attach to a web site or page. Tags are simply labels for web resources, selected to help the user in later retrieval of those web resources. Tags have the additional effect of grouping related web resources together. There is no fixed set of categories or officially approved choices. A user can use words, acronyms, numbers – whatever seems to make sense – without regard for anyone else's needs, interests, or requirements. With tagging, anyone is free to use any appropriate words, without having to agree with anyone else about how something "should" be tagged (Shirky 2005).

The word Folksonomy is a portmanteau of the words *folks* and *taxonomy* coined by Thomas Vander Wal (Smith 2006), which implies that it can be understood as an organization of web contents by folks (users). The classifiers in folksonomy are not dedicated information professionals, and Thomas Vander Wal described this as a "bottom-up social classification" (Vander Wal 2004, 2005a,b), unlike traditional approaches to library classification (e.g., *Dewey Decimal Classification* (*DDC*), and Library of Congress *Classification* (LCC)).

In folksonomy-enabled systems, users of the documents create metadata for their own individual use that are also shared throughout a community

(Mathes 2004) under the same tag, or share different tags assigned to the same piece of content. Web users describe and organize the content (bookmarks, web sites or pages or photos) with their own vocabulary (words) and assign one or more keywords, namely tags, to each single unit of content. Folksonomy is, thus, implemented through the tags assigned and is currently often understood as tagging (Shen and Wu 2005). Tagging terms facilitate users' searching and information interpretation and help users to identify the main ideas around the topic on the Web. Collaborative tagging or folksonomy describes the process by which many users add metadata in the form of keywords or tags to shared content (Golder and Huberman 2005).

Folksonomy-based systems enable users to categorize their bookmarks or links with tags. Folksonomy is understood to be organized by every user while not limited to the authors of the contents and professional editors (Shen and Wu 2005). Imagine a book, with an author and a back-of-the book indexer. The indexer is, here, a folksonomy user. The indexer reads and gives index terms to the book, apart from the words used by the author. Consequently, folksonomy tags are *index terms* from the point of view of the user. *Index term* is the representation of a concept, preferably in the form of a noun or noun phrase derived from natural language. Nouns are chosen because they are the most concrete part of speech. An index term can consist of more than one word. Index terms should be checked for accuracy and acceptability in reference tools, such as dictionaries, encyclopedias, thesauri and classification schemes (ISO 1985).

User-originated tagging or folksonomy is a common way to organize content for future navigation, filtering or search. In fact, folksonomy is the practice of allowing users to freely attach keywords or tags to content. Folksonomy is most useful when there is nobody in the "librarian" role or there is simply too much content for a single authority to classify; both of these traits are true of the Web, where folksonomy tagging has grown popular. Thus, folksonomy tagging services allow users to publicly tag and share content, so that they can not only categorize information for themselves, but they also can browse the information categorized by others. There are therefore at once both personal and public aspects to folksonomy tagging systems (Golder and Huberman 2005).

2. How to use folksonomy?

In folksonomy-based systems, once users have created accounts, they can then begin bookmarking web resources; each bookmark records the web resource's URL and its title, as well as the time at which the bookmark is created. Thus the address of a web resource one might wish to visit in the future is saved in that users' own web space. To create a collection of individual bookmarks, the user registers with a social bookmarking site, which allows storage of bookmarks, addition of selected tags, and designation of individual bookmarks as public or private. Some sites periodically verify that bookmarks still work; notifying users when a URL no longer functions. Folksonomy-based systems generally group the tagged web resources (links) by day, each link entry indicates the name of the user who tagged it, the title of the web resources (which is also a hyperlink to the linked resource), the URL of the resource and any keyword tags, or comments annotating that entry. Users can tag the bookmark with multiple tags, or keywords, of their choice. Each user has a personal page for display of bookmarks (see for example, *http://del.icio.us/username*). On this page, all the bookmarks the user has ever created are displayed in reverse-chronological order along with a list of all the tags the user has ever given to a bookmark. By selecting a tag, a user can filter the bookmark list so that only bookmarks with that tag are displayed (Golder and Huberman 2005).

The most popular, widely used folksonomy-based systems are:

1. Del.icio.us: www.del.icio.us
2. Flickr: www.flickr.com
3. YouTube: www.youtube.com
4. CiteULike: www.citeulike.org
5. Connotea: www.connotea.org
6. Technorati: www.technorati.com
7. Furl: www.furl.net
8. TagCloud: www.tagcloud.com
9. Yahoo's MyWeb: http://myweb.yahoo.com
10. Simpy: www.simpy.com
11. Unalog: www.unalog.com
12. Shadows: www.shadows.com
13. Spurl: www.spurl.net
14. Scuttle: www.scuttle.org
15. Tagzania: www.tagzania.com
16. Dabble: www.dabble.com
17. LibraryThing: www.librarything.com
18. Wink: www.wink.com

Folksonomy opens the door to a whole new way of gathering and organizing information by tagging and categorizing web resources. The creator of a bookmark assigns tags to each web resource, resulting in a user-directed, *amateur* method of classifying and organizing information. Because folksonomy services indicate who created each bookmark and provide access to that person's other bookmarked web resources, users can easily make social connections with other people interested in just about any topic. In popular systems like *Del.icio.us* users see not only their own bookmarks but all of every other user's bookmarks as well. Users can easily see how many people have used a tag and search for all web resources that have been tagged with that tag. In this way, the community of users over time can develop a unique structure of keywords to define resources.

The "Tags" box lets users optionally add multiple keyword tags describing their favorite resource. These keywords are not from controlled vocabularies, but users can choose to use the same keyword tags repeatedly. If a user bookmarks a web resource about Webometrics on *Del.icio.us*, it might be tagged as "Webometric/s", "Bibliometric/s", "Link analysis", "Hyperlink" and "Web." Another user might come along and search for the tag "Webometric" or "Webometrics," finding the same resource as well as those tagged by everyone else who shared the same tag. Some folksonomy-based systems like *Del.icio.us* suggest some additional tags to consider. *Del.icio.us* infers its knowledge from the tags entered by every other user in the system, creating a folksonomy, a group intelligence derived by association.

3. Advantages and Disadvantages of Folksonomy

Folksonomy-based systems can: 1) store personal bookmarks; 2) analyze users' bookmark histories and extract user groups which have similar interests; and, 3) recommend resources which are commonly preferred. In contrast to the *Favorites* of browsers such as *Internet Explorer*, folksonomy-based systems like *Del.icio.us* allow users to create or remove associations between tags and web resources by adding, replacing or deleting bookmarks or tags. The advantage of saving bookmarks in this way is that once a user's bookmarks are on the Web, they are accessible from any computer, not just the user's own browser. This is helpful if a user uses multiple computers, at home, universities, work, and so on, so this is considered one of the main features of *Del.icio.us*. Through others' personal pages and the "popular" page, users can get a sense of what other people find interesting. By browsing specific people and tags, users can find web resources that are of interest to them and can find people who have common interests.

Another advantage is that users' interests can be identified. Users' lists of tags can be considered descriptive of the interests they hold as well as of their method of classifying those interests. Users' tag lists grow over time, as they discover new interests and add new tags to categorize and describe them. It is possible that the newly growing tag represents a new interest or category to the user (Golder and Huberman 2005).

Among the disadvantages of folksonomy are *low precision* and *lack of collocation* that originate from the absence of properties that characterize controlled vocabularies. These need to be dealt with. However, librarians and information professionals have lessons to learn from the interactive and social aspects exemplified by collaborative tagging systems, as well as their success in engaging users with information management. The future coexistence of controlled vocabularies and collaborative tagging is predicted, with each appropriate for use within distinct information contexts: formal and informal (Macgregor and McCulloch 2006).

Four main problems of folksonomy tagging are ***polysemy***, ***synonymy***, ***plurals***, and ***depth*** (***specificity***) ***of tagging***.

***Polysemy*:** Polysemy refers to a word that has two or more similar meanings. "Poly" means 'many', and "semy" means 'meanings'. A polysemous word is one that has *many* ("poly") related *senses* ("semy"). For example, a "window" may refer to a hole in the wall, or to the pane of glass that resides within it (Pustejovsky 1995). In practice, polysemy dilutes query results by returning related but potentially inapplicable items. Superficially, polysemy is similar to homonymy, where a word has multiple, unrelated meanings. However, homonymy is less a problem because homonyms can be largely ruled out in a tag-based search through the addition of a related term with which the unwanted homonym would not appear. There are, of course, cases where homonyms are semantically related but not polysemous (Golder and Huberman 2005).

Synonymy: Synonymy, different words with similar or identical meanings, presents a greater problem for tagging systems because inconsistency

among the terms used in tagging can make it very difficult for one to be sure that all the relevant items have been found. It is difficult for a folksonomy user to be consistent in the terms chosen for tags; for example, items about the Web may be tagged either World Wide Web or WWW. This problem is compounded in a collaborative system, where all folksonomy users either need to widely agree on a convention, or else accept that they must issue multiple or more complex queries to cover many possibilities. Synonymy is a significant problem because it is impossible to know how many items "out there" a user would have liked a search query to have retrieved (Golder and Huberman 2005). A quick search on *Del.icio.us* reveals that the users exhibit much variety in the sets of tags they employ, even a simple concept ("New York City", for example, is tagged as "NewYorkCity", "New_York_City", "New-York-City", "New.York.City", "New-York", "NewYork", "New.York", "NYC", "NY", etc). An ideal folksonomy-based system would support automatic suggestions for reformatting tags to fit with international trends.

A controlled vocabulary, e.g., a thesaurus, controls the use of synonyms (and near-synonyms), homonyms, homographs, heteronyms, and grammatical variations by establishing a single form of the term, reducing the probability that relevant resources will be missed during a search (for multiple definitions of a controlled vocabulary see: Macgregor and McCulloch 2006). For example, "car", "automobile", "motorcar", or "motor vehicle", etc.

Plurals: Plurals and parts of speech and spelling can undermine a tagging system. For example, if tags *Cat* and *Cats* are distinct, then a query for one will not retrieve both, unless the system has the capability to perform such replacements built into it. For instance, consider a hypothetical researcher who wants a document about *Cat* species native to *Persia* (*Persian cats*). A disadvantage of folksonomy-based systems is that a web resource tagged only *Cat* would not be found by the query *Persian Cats*, though it arguably should be. A searcher may still need to search multiple queries. A tag returns only those resources tagged with that tag, while in library catalogues, there may be a cross-reference (see also) from *Cats* to *Persian cats*.

Depth (specificity) of tagging: Specificity means how specific should the user (classifier) be in translating a concept into index term(s)? Web resources can be tagged to varying levels of specificity, from very broad subjects taken only from the title and abstract to the paragraph level. The depth of tags refers to how many tags there are, relative to a web resource in the system. Tonkin (2006) deduces that the choice of tags is necessarily strongly influenced by user behaviour and habit.

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; as discussed above, *Cat*, *Cheetah* and *Animal* are all reasonable ways to describe a particular entity. The problem lies in the fact that different users may consider terms at different levels of specificity to be most useful or appropriate for describing the item in question. The "*basic level*," as opposed to superordinate (more general) and subordinate (more specific) levels, is that which is most directly related to humans' interactions with them. For most people, the *basic level* for *Felines* would be "*Cat*," rather than "*Animal*" or "*Siamese*" or "*Persian*." Experiments demonstrate that, when asked to identify *Dogs* and *Birds*, subjects used "*Dog*" and "*Bird*" more than "*Beagle*" or "*Robin*," and when asked whether an item in a picture is an X, subjects responded more quickly when X was a "*basic level*" (Tanaka and Taylor 1991). These experiments demonstrate general agreement across subjects (Golder and Huberman 2005). For the purposes of tagging systems, however, conflicting *basic levels* can prove disastrous, as documents tagged *Javascript* and *XML* may be too specific for some users, while a document tagged *programming* may be too general for others.

Moreover, 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. Adjectives such as scary, funny, stupid, inspirational tag bookmarks according to the tagger's opinion of the content. Tags beginning with "my," like mystuff and mycomments identify content in terms of its relation to the tagger. Some tags are used by many users, while other tags are used by fewer people (Golder and Huberman 2005).

Folksonomy tagging, then, has the potential to exacerbate the problems associated with the fuzziness of linguistic and cognitive boundaries. As all

folksonomy users' contributions collectively produce a larger classification system, that system consists of idiosyncratically personal categories as well as those that are widely agreed upon (Golder and Huberman 2005).

4. Conclusion

Folksonomy has not only changed the methodology of classification (the distribution and decentralization of labor), but also necessitates a deep change in the way that classifiers organize information. It has removed all concept of hierarchy from the scheme of knowledge organization, facilitating knowledge discovery and web indexing. Although, there is not a perfect system in the world that satisfies every user, we can do better. By controlling vocabularies, search engines could present search results in clusters and attach each cluster to terms having the highest frequency, designating them as the tagging terms of the cluster. They should also be able to recommend tags used by other users: "A lot of users who tagged this 'Open Access' also tagged it 'OA'."

Acknowledgment

The author wishes to thank Mrs. Marjorie Sweetko for her helpful comments in producing the paper.

References

Folksonomy. 2007. In *Wikipedia, the free encyclopedia*. Retrieved November 25, 2006, from http://en.wikipedia.org/wiki/Folksonomy

Golder, Scott A., and Huberman, Bernardo A. 2005. The structure of collaborative tagging systems. Information Dynamics Lab: HP Labs, Palo Alto, CA. Retrieved March 13, 2007, from http://arxiv.org/abs/cs.DL/0508082

ISO (1985). Documentation- Methods for examining documents, determining their subjects, and selecting indexing terms. ISO 5963-1985(E). *ISO Standards Handbook*. Switzerland: International Organization for Standardization. p. 580.

Macgregor, George, and McCulloch, Emma. 2006. Collaborative tagging as a knowledge organisation and resource discovery tool. *Library Review*, 55(5): 291-300.

Mathes, Adam. 2004. Folksonomies – cooperative classification and communication through shared metadata. *Computer Mediated Communication* – LIS590CMC, December 2004. Retrieved March 13, 2007, from http://adammathes.com/academic/computer-mediated-communication/folksonomies.pdf

Pustejovsky, James. 1995. *The generative lexicon*. Cambridge, MA: The MIT Press.

Shen, Kaikai, and Wu, Lide. 2005. Folksonomy as a complex network. September 23, 2005. Retrieved March 13, 2007, from http://arxiv.org/abs/cs.IR/0509072

Shirky, Clay. 2005. Ontology is overrated: Categories, links and tags. Retrieved March 13, 2007, from http://www.shirky.com/writings/ontology_overrated.html

Smith, Gene. 2006. Atomiq: Folksonomy: Social classification. Retrieved March 13, 2007, from http://atomiq.org/archives/2004/08/folksonomy_social_classification.html

Tanaka, James W., and Taylor, Marjorie. 1991. Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, 23(3): 457-482.

Tonkin, Emma. 2006. Folksonomies: The fall and rise of plain-text tagging. *Ariadne*, 47, April 2006. Retrieved March 13, 2007, from http://www.ariadne.ac.uk/issue47/tonkin/intro.html

Vander Wal, Thomas. 2004. You down with folksonomy? August 4, 2004. Retrieved March 13, 2007, from http://www.vanderwal.net/random/entrysel.php?blog=1529

Vander Wal, Thomas. 2005a. Explaining and showing broad and narrow folksonomies. February 21, 2005. Retrieved March 13, 2007, from http://www.vanderwal.net/random/entrysel.php?blog=1635

Vander Wal, Thomas. 2005b. Folksonomy definition and Wikipedia. November 2, 2005. Retrieved March 13, 2007, from http://www.vanderwal.net/random/entrysel.php?blog=1750