The structure and form of folksonomy tags: the road to the public library catalogue

Louise F. Spiteri

School of Information Management. Dalhousie University. Halifax, Nova Scotia. Canada. Louise.Spiteri@dal.ca

Abstract

Folksonomies have the potential to add much value to public library catalogues by enabling clients to: store, maintain, and organize items of interest in the catalogue using their own tags. The purpose of this paper is to examine how the tags that constitute folksonomies are structured. Tags were acquired over a thirty-day period from the daily tag logs of three folksonomy sites, Del.icio.us, Furl, and Technorati. The tags were evaluated against section 6 (choice and form of terms) of the National Information Standards Organization (NISO) guidelines for the construction of controlled vocabularies. This evaluation revealed that the folksonomy tags correspond closely to the NISO guidelines that pertain to the types of concepts expressed by the tags, the predominance of single tags, the predominance of nouns, and the use of recognized spelling. Potential problem areas in the structure of the tags pertain to the inconsistent use of the singular and plural form of count nouns, and the incidence of ambiguous tags in the form of homographs and unqualified abbreviations or acronyms. Should library catalogues decide to incorporate folksonomies, they could provide clear guidelines to address these noted weaknesses, as well as links to external dictionaries and references sources such as Wikipedia to help clients disambiguate homographs and to determine if the full or abbreviated forms of tags would be preferable.

Keywords: Collaborative tagging, Controlled vocabularies, Folksonomies, Guidelines.

Resumen

Las folksnomías tienen el potencial de proporcionar valor añadido a los catálogos de las bibliotecas públicas permitiendo a los clientes almacenar, mantener y organizar ítems de interés en el catálogo utilizando sus propias etiquetas. El propósito de esta comunicación es examinar de qué modo las etiquetas que constituyen folksonomías están estructuradas. Las etiquetas han sido recogidas durante un período de treinta días a partir de tres sitios de folksonomías: Del.icio.us, Furl y Technorati. Las etiquetas fueron evaluadas siguiendo la sección 6 (elección y forma de los términos) de las directrices para la construcción de vocabularios controlados de la National Information Standards Organization (NISO). La

evaluación reveló que las etiquetas usadas en las folksonomías se adaptan a las directrices de la NISO que abogan por el predominio de términos simples, de sustantivos y por el uso de una grafía reconocible. Los aspectos potencialmente problemáticos en la estructura de las etiquetas son el uso inconsistente del singular y plural de los nombres contables, la incidencia de etiquetas ambiguas en el caso de los conceptos homógrafos, y problemas de identificación de abreviaturas o acrónimos. En el caso de que los catálogos de las bibliotecas decidan incorporar folksonomías, han de proporcionar directrices claras para evitar las debilidades reseñadas, al igual que enlaces a diccionarios y obras de referencia como Wikipedia que permitan a los usuarios desambiguar homógrafos y facilitar la elección entre las formas completas o abreviadas de las etiquetas.

Palabras clave: Directrices, Etiquetado colaborativo, Folksonomías, Vocabularios controlados.

1 Introduction

Digital document repositories such as library catalogues normally index the subject of their contents via keywords or subject headings. Traditionally, such indexing is performed either by an authority, such as a librarian or a professional indexer, or else is derived from the authors of the documents; in contrast, collaborative tagging, or folksonomies, allows anyone to freely attach keywords or tags to content. Demspey (2003) and Ketchell (2000) recommend that clients be allowed to annotate resources of interest and to share these annotations with other clients with similar interests. Folksonomies can thus make significant contributions to public library catalogues by enabling clients to organize personal information spaces, namely to create and organize their own personal information space in the catalogue. Clients find items of interest (items in the library catalogue, citations from external databases, external web pages, etc.) and store, maintain, and organize them in the catalogue using their own tags.

In order to understand more fully these applications, it is important to examine how folksonomies are structured and used, and the extent to which they reflect user needs not found in existing lists of subject headings. The purpose of this study is to evaluate the structure and form of folksonomies against section 6 of the NISO guidelines for the construction of controlled vocabularies (NISO, 2005), which looks specifically at the choice and form of terms.

2 Definitions of Folksonomies

Folksonomies have been described as “user created metadata ... grassroots community classification of digital assets” (Mathes, 2004). Wikipedia (2006) describes a folksonomy as “an Internet-based information retrieval methodology consisting of collaboratively generated, open-ended labels that categorize content such as Web pages, online photographs, and Web links.” The concept of collaboration is attributed commonly to folksonomies. Thomas Vander Wal, who coined the term *folksonomy*, argues that tagging is done in a social environment (shared and open to others). The act of tagging is done by the person consuming the information (Vander Wal.Net, 2005). It may be more accurate, therefore, to say that folksonomies are created in an environment where, although people may not actively collaborate in their creation and assignation of tags, they may certainly access and use tags assigned by others.

3 Benefits of Folksonomies

Quintarelli (2005) and Fichter (2006) suggest that folksonomies reflect the movement of people away from authoritative, hierarchical taxonomic schemes; the latter reflect an external viewpoint and order that may not necessarily reflect users' ways of thinking. " In a social distributed environment, sharing one's own tags makes for innovative ways to map meaning and let relationships naturally emerge" (Quintarelli, 2005). Vander Wal (2006) adds that "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."

An attractive feature of folksonomies is their inclusiveness; they reflect the vocabulary of the users, regardless of viewpoint, background, bias, and so forth. Folksonomies may thus be perceived to be a democratic system where everyone has the opportunity to contribute and share tags (Kroski, 2006). The development of folksonomies may reflect also the difficulty and expense of applying controlled taxonomies to the Web: Building, maintaining, and enforcing a sound controlled vocabulary is often simply too expensive in terms of development time and of the steep learning curve needed by the user of the system to learn the classification scheme (Fichter, 2006; Kroski, 2006; Quintarelli, 2005; Shirky, 2004). A further limitation of taxonomies is that they may become outdated easily: New concepts or products may emerge that are not yet included in the taxonomy; in comparison, folksonomies accommodate easily such new concepts (Fichter, 2006; Mitchell, 2005). Shirky (2004) points out that the advantage of folksonomies is not that they are better than controlled vocabularies, but that they are better than nothing.

4 Weaknesses of Folksonomies

Folksonomies share the problems inherent to all uncontrolled vocabularies, such as ambiguity, polysemy, synonymy, and basic level variation (Fichter, 2006; Golder and Huberman, 2006; Guy and Tomkin, 2006; Mathes, 2004). The terms in a folksonomy may have inherent ambiguity as different users apply terms to documents in different ways. The polysemous tag *port* could refer to a sweet fortified wine, a porthole, a place for loading and unloading ships, the left-hand side of a ship or aircraft, or a channel endpoint in a communications system. Folksonomies do not include guidelines for use or scope notes. Folksonomies provide for no synonym control; the terms *mac*, *macintosh*, and *apple*, for example, are used to describe Apple Macintosh computers. Similarly, both singular and plural forms of terms appear (e.g., flower and flowers), thus creating a number of redundant headings. The problem with basic level variation is that related terms that describe an item vary along a continuum of specificity ranging from very general to very specific; so, for example, documents tagged *perl* and *javascript* may be too specific for some users, while a document tagged *programming* may be too general for others. Folksonomies provide no formal guidelines for the choice and form of tags, such as the use of compound headings, punctuation, word order, and so forth; for example, should one use the tag *vegan cooking* or *cooking, vegan*? Guy and Tomkin (2006) provide some general suggestions for tag selection best practices, such as the use of plural rather than singular forms, the use of underscore to join terms in a multi-term concept (e.g., open_source), following conventions established by others, and adding synonyms. These suggestions are rather too vague to be of much use, however; for example, under what circumstances should singular forms be used (e.g., non-count nouns), and how should synonyms be linked?

The pitfalls of folksonomies have been well documented; what is missing is an in-depth analysis of the linguistic structure of tags against an established benchmark. While popular opinion suggests that folksonomies suffer from ambiguous and inconsistent structure, the actual extent of these problems is not yet clear; furthermore, analyses conducted so far have not established clear benchmarks of quality pertaining to good tag structure. Although there are no guidelines for the construction of tags, recognized guidelines do exist for the construction of terms that are used in taxonomies. Although these guidelines discuss the elucidation of inter-term relationships (hierarchical, associative, and equivalent), which does not apply to the flat space of folksonomies, they contain sections pertaining to the choice and formation of concept terms, which may, in fact, have relevance for the construction of tags.

5 Methodology

Tags were chosen from three popular folksonomy sites: Delicious, Furl, and Technorati[1]. Delicious and Furl function as bookmarking sites, while Technorati enables people to search for, and organize, blogs. These sites were chosen because they provide daily logs of the most popular tags that have been assigned by their members on a given day. The daily tag logs from each of the sites were acquired over a thirty-day period. A list of unique tags for each site was compiled after the thirty-day period; *unique* refers to the single instance of a tag. The analysis of the tag structure in the three lists was conducted by applying the NISO guidelines for thesaurus construction (NISO, 2005), which are the most current set of recognized guidelines for the construction of controlled vocabularies. While folksonomies are not controlled vocabularies, they are lists of terms used to describe content, which means that the NISO guidelines could work well as a benchmark against which to examine how folksonomy tags are structured, as well as the extent to which this structure reflects the widely-accepted norm for controlled vocabularies.

6 Findings

Unless stated otherwise, the number of tags per folksonomy site is 76 for Delicious, 208 for Furl, and 229 for Technorati.

6.1 Homographs

The NISO guidelines recommend that homographs - terms with identical spellings but different meanings - should be avoided as far as possible in the selection of terms (NISO, 2005, p. 32). Homographs constitute 22% of Delicious tags, 12% of Furl tags, and 20% of Technorati tags. Unique entities constitute a significant proportion of the homographs in all three sites, with 71% in Delicious, 43% in Furl, and 55% in Technorati. The most frequently-occurring homographs across the three sites consist predominantly of computer-related products, such as Ajax and CSS.

6.2 Single word vs. multiword terms

The NISO guidelines recommend that terms should represent a single concept expressed by a single term or multiword term, as needed (NISO, 2005, p. 35). Single term tags constitute 93% of Delicious tags, 76% of Furl tags, and 80% of Technorati tags. The preponderance of

[1] http://www.technorati.com

single tags in Delicious may reflect the fact that it does not allow for the use of spaces between the different elements of the same tag, e.g., *open source*.

6.3 Types of concepts

NISO provides a list of seven types of concepts that may be represented by terms; while this list is not exhaustive, it represents the most frequently-occurring types of concept. Table 1 shows the percentage of tags that correspond to each of the seven types of concepts:

Table 1. Concepts Represented by the Tags

	Delicious	**Furl**	**Technorati**
Things	76%	82%	90%
Materials	0%	0%	0.4%
Activities	12%	10%	4%
Events	0%	0%	0%
Properties	8%	6%	4%
Disciplines	4%	3%	1%
Measures	0%	0%	0%

Tags that represent *things* are clearly predominant in the three sites, with activities and properties forming a distant second and third in importance. None of the tags represent events or measures, and only a fraction of the Technorati tags represent materials. None of the tags fell outside the scope of the seven types of concepts.

6.4 Unique Entities

Unique entities may represent the names of people, places, organizations, products, and specific events (NISO 2005, p. 36). Unique entities constitute 22% of Delicious tags, 14% of Furl tags, and 49% of Technorati tags. There is no consistency in the percentage of unique entities: Technorati has nearly twice the percentage of tags than Delicious has, and nearly triple the percentage of tags than Furl has. Computer-related products constitute 100% of the unique entities in Delicious, 63% in Furl, and 38% in Technorati. The remainder of the unique entities in Furl and Technorati represent places, people, and corporate bodies.

6.5 Grammatical forms of terms

Table 2 shows the distribution of the grammatical forms of tags:

Table 2. Grammatical Form of Tags

	Delicious	**Furl**	**Technorati**
Nouns	88%	71%	86%
Verbal Nouns	5%	6%	4%
Noun Phrases - Premodified	1%	15%	4%
Noun Phrases- Postmodified	0%	2%	3%
Adjectives	6%	6%	3%
Adverbs	0%	0%	0%

If all the types of nouns are combined, then 95% of Delicious tags, 94% of Furl tags, and 97% of Technorati tags constitute types of nouns. The grammatical structure of the tags in the three

folksonomy sites thus reflects very closely the NISO recommendations that tags consist of mainly nouns, with the added proviso that adjectives and adverbs be kept to a minimum. None of the folksonomy sites used adverbs as tags, and the number of adjectives was very small, forming an average total of 5% of the tags.

6.6 Nouns (plural and singular forms)

NISO divides nouns into two categories: Count nouns (how many?) and non-count, or mass nouns (how much?). NISO recommends that count nouns appear in the plural form and mass nouns in the singular form (NISO, 2005, p. 40). Count nouns formed 18% of Delicious tags, 35% of Furl tags, and 23% of Technorati tags; the remaining tags are non-count tags. Of the count nouns, 36% of Delicious tags, 62% of Furl tags, and 34% of Technorati tags appeared correctly in the plural form.

6.7 Spelling

NISO recommends that spelling should follow the practice of well established dictionaries or glossaries (NISO, 2005, p. 42). The number of tags that do not conform to spelling warrant is very minor, constituting a total of 4% of the Delicious tags, 3% of the Furl tags, and 2% of the Technorati tags. The findings suggest that tags are spelled very consistently and in keeping with recognized warrant across the three folksonomy sites.

6.8 Abbreviations, Initialisms, and Acronyms

NISO recommends that the full form of terms should be used. Abbreviations or acronyms should be used only when they are so well established that the full form of the term is rarely used (NISO, 2005, p. 42). Abbreviations and acronyms constitute 22% of Delicious tags, 16% of Furl tags, and 19% of Technorati tags. The majority of these abbreviations and acronyms pertain to unique entities, such as product names (e.g., *Flash*, *Mac*, and *NFL*). In the case of Delicious and Furl, none of the abbreviated tags is referred to also by its full form. Abbreviations and acronyms play a significant role in the ambiguity of the tags from the three sites; they represent 71% of the abbreviated Delicious tags, 45% of the abbreviated Furl tags, and 73% of the abbreviated Technorati tags. Furl and Technorati are very similar in the proportion of abbreviated tags used, but Delicious is significantly higher. The Delicious tags are focused more heavily upon computer-related products, which may explain why there are so many more abbreviated tags, since many of these products are often referred to by these shorter terms, e.g., CSS, Flash, Apple, etc.

6.9 Neologisms, Slang, and Jargon

The NISO guidelines explain that neologisms, slang, and jargon terms are generally not included in standard dictionaries and should be used only when there is no other widely-accepted alternative (NISO, 2005, p.43). Non-standard tags do not constitute a particularly relevant proportion of the total number of tags per site; they account for 3% of the Delicious tags, 10% of the Furl tags, and 6% of the Technorati tags. The non-standard tags refer almost exclusively to either computer-related concepts, or sex-related concepts, e.g., *Podcast*, *Wiki*, and *Camsex*.

7 Discussion and Recommendations

The tags examined from the three folksonomy sites correspond closely to a number of the NISO guidelines pertaining to the structure of terms, namely in the types of concepts expressed by the tags, the predominance of single tags, the predominance of nouns, the use of

recognized spelling, and the use of primarily alphabetic characters. Potential problem areas in the structure of the tags pertain to the inconsistent use of the singular and plural form of count nouns, the difficulty with creating multiterm tags in Delicious, and the incidence of ambiguous tags in the form of homographs and unqualified abbreviations or acronyms. As has been seen, a significant proportion of tags that represent count nouns appears incorrectly in the singular form. Since many search engines do not deploy default truncation, the use of the singular or plural form could affect retrieval; a search for the tag *computer* in Delicious, for example, retrieved 208,409 hits, while one for *computers* retrieved 91,205 hits.

While all three sites conform to the NISO recommendation that single terms be used whenever possible, some concepts cannot be expressed in this fashion and thus folksonomy sites should accommodate the use of multiterm tags. Furl and Technorati allow for the use of multiterm tags, but make no mention of this feature in their help screens, which means that such tags may be constructed inconsistently, for example, by the insertion of punctuation, where a simple space between the tags will suffice. Delicious should consider allowing for the insertion of spaces between the composite words of a compound tag; without this facility, users may be unaware of how to create compound tags. Alternatively, Delicious should recommend the use of only one punctuation symbol to conflate terms, such as the underscore. Furl and Technorati should explain clearly that compound tags may be formed by the simple convention of placing a space between the terms.

Ambiguous headings constitute the most problematic area in the construction of the tags; these headings take the form of homographs and abbreviations or acronyms. In the case of computer-related product names, it may be safe to assume that in the context of an online environment, it is likely that the meaning of these product names is relatively self evident. The application of the section of the NISO guidelines pertaining to abbreviations and acronyms is particularly difficult, as it is important to balance between using abbreviated forms of concepts that are so well known that the full version is hardly used, versus creating ambiguous tags. The fact that abbreviated forms appear so prominently in the daily logs of the three folksonomy sites suggests that the full forms of these tags are, in fact, very well established. It may be useful for the folksonomy sites to add direct links to an online dictionary and to Wikipedia, and to encourage people to use these sites to determine whether their chosen tags may have more than one application or meaning.

8 Conclusion

The most notable suggested weaknesses of folksonomies are their potential for ambiguity, polysemy, synonymy, basic level variation, and the lack of consistent guidelines for the choice and form of tags. The examination of the tags of the three folksonomy sites in light of the NISO guidelines suggests that ambiguity and polysemy (i.e., homographs) are indeed problems in the structure of the folksonomy tags, although the actual proportion of homographs and ambiguous tags each constitutes less than one-quarter of the tags in each of the three folksonony sites. In other words, although ambiguity and polysemy are certainly problematic areas, most of the tags in each of the three sites are unambiguous in their meaning and thus conform to NISO recommendations. In other areas, the tags conform closely to the NISO guidelines for the choice and form of controlled vocabularies. The tags represent mostly nouns, with very few unqualified adjectives or adverbs. The tags represent the types of concepts recommended by NISO, and conform well to recognized standards of spelling. Most of the tags conform to standard usage; there are few instances of non-standard usage, such as

slang or jargon. In short, the structure of the tags in all three sites is well within the standards established and recognized for the construction of controlled vocabularies.

Should library catalogues decide to incorporate folksonomies, they should consider creating clearly-written recommendations for the choice and form of tags that could include the following areas:

- The difference between count and non-count nouns, as well as an explanation of how the use of the singular and plural forms affects retrieval;
- One standard way in which to construct multiterm tags, e.g., the insertion of a space between the component terms, or the use of an underscore between the terms; and
- A link to a recognized online dictionary and to Wikipedia to enable users to determine the meanings of terms, to disambiguate amongst homographs, and to determine if the full form would be preferable to the abbreviated form. An explanation of the impact of ambiguous tags and homographs upon retrieval would be useful.

With the use of such expanded guidelines and links to useful external reference sources, folksonomies could serve as a very powerful and flexible tool for increasing the user-friendliness and interactivity of public library catalogues and may be useful also for encouraging other activities, such as informal online communities of readers and user-driven readers' advisory services.

References

ANSI/NISO Z39.19-2005. *Guidelines for the construction, format, and management of monolingual controlled vocabularies*. Bethesda, MD: National Information Standards Organization, 2005.

DEMPSEY, L. The recombinant library: portals and people. *Journal of Library Administration*, 2003, vol. 39, n. 4, p. 103-136.

FICHTER, D. Intranet applications for tagging and folksonomies. *Online*, 2006, vol. 30, n. 3, p. 43-45.

GOLDER, S. A.; HUBERMAN, B. A. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 2006, vol. 32, n. 2, p. 198-208.

GUY, M.; TONKIN, E. Tidying up tags? [electronic resource]. *D-Lib Magazine*, 2006, vol. 12, n. 1. <http://www.dlib.org/dlib/january06/guy/01guy.html>. [Consulted: 29 nov. 2006]

KETCHELL, D. S. Too many channels: Making sense out of portals and personalization. *Information Technology and Libraries*, 2000, vol. 19, n. 4, p. 175-179.

KROSKI, E. *The hive mind: folksonomies and user-based tagging* [electronic resource]. <http://infotangle.blogsome.com/2005/12/07/the-hive-mind-folksonomies-and-user-based-tagging/ >. [Consulted: 29 nov. 2006]

MATHES, A. *Folksonomies - cooperative classification and communication through shared metadata*, 2004 [electronic resource]. <http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html>. [Consulted: 29 nov. 2006]

MITCHELL, R. L. Tag teams wrestle with web content. *Computerworld*, 2005, vol. 38, n. 16, p. 31.

QUINTARELLI, E. Folksonomies: power to the people [electronic resource]. In: *Incontro ISKO Italy - UniMIB (2nd: Milano: 2005)* <http://www.iskoi.org/doc/folksonomies.htm>. [Consulted: 29 nov. 2006].

SHIRKY, C. *Folksonomy*, 2004 [electronic resource]. <http://www.corante.com/many/archives/2004/08/25/folksonomy.php>. [Consulted: 29 nov. 2006]

UDELL, J. *Collaborative knowledge gardening*, 2004 [electronic resource]. <http://www.infoworld.com/article/04/08/20/34OPstrategic_1.html>. [Consulted: 29 nov. 2006]

VANDERWAL.NET. *Folksonomy definition and Wikipedia*, 2005 [electronic resource]. <http://www.vanderwal.net/random/entrysel.php?blog=1750 >. [Consulted: 29 nov. 2006]

WIKIPEDIA. *Folksonomy* [electronic resource]. <http://en.wikipedia.org/wiki/Folksonomy >. [Consulted 29 nov. 2006].