UDC and folksonomies
[short paper]

Alenka Šauperl
University of Ljubljana (Slovenia)

The full paper will be published in *Knowledge Organization*, 37, 4 (2010).

ABSTRACT: Social tagging systems, known as 'folksonomies', represent an important part of web resource discovery as they enable free and unrestricted browsing through information space. Folksonomies consisting of subject designators (tags) assigned by users, however, have one important drawback: they do not express semantic relationships either hierarchical or associative between tags. As a consequence, the use of tags to browse information resources requires moving from one resource to another, based on coincidence and not on the pre-established meaningful or logical connections that may exist between related resources. We suggest that the semantic structure of the Universal Decimal Classification (UDC) may be used in complementing and supporting tag-based browsing. In this work, two specific questions were investigated: (1) Are terms used as tags in folksonomies included in the UDC? and (2) Which facets of UDC match the characteristics of documents or information objects that are tagged in folksonomies? A collection of the most popular tags from Amazon, LibraryThing, Delicious and 43Things was investigated. The universal nature of UDC was examined through the universality of topics and facets covering diverse human interests which are at the same time interconnected and form a rich and intricate semantic structure. The results suggest that UDC-supported folksonomies could be implemented in resource discovery, in particular in library portals and catalogues.

KEYWORDS: social tagging, folksonomy, UDC, comparison

Introduction

Folksonomies are popular among Web users, allowing them to tag documents (index them), by assigning keywords to resources they find on the Web or that they submit themselves. This enables users to retrieve documents they have already accessed or find new documents other users have tagged. Browsing is aided by tag clouds, i.e. groups of tags, sorted alphabetically and presented by size - their size expresses the frequency of their use.
Inspired by numerous research reports on folksonomies in ILS, we proposed two research questions:

(1) Are terms used as tags in folksonomies included in the UDC? ^

(2) Which facets of the UDC match the characteristics of documents or information objects that are tagged in folksonomies?

Method

We analysed tags in four folksonomies: Del.icio.us, 43 Things, Library Thing and Amazon. Their selection was intentional: we wanted to see whether UDC is in fact universal and can cover subject description of information objects, which are not usually held by libraries.

The site Del.icio.us is a Web service that allows users to save, share and organize their favourite bookmarks, such as addresses of web resources. 200 most popular tags were also selected from 43 Things. This is a site that allows users to note and share personal goals. Library Thing allows users to catalogue books and similar traditional library material. Three works were selected for analysis: *Cold Mountain*, a novel written by Charles Frazier, *The Little Mermaid* by the brothers Grimm, and *The Sound of Music* by Maria von Trapp. In each case, tags associated with the books, sound recordings and movies were analysed. The same sample was also analysed in Amazon, the popular Web bookstore. All samples were collected in June 2009.

Our content analysis consisted of categorizing tags. Some categories were expected and prepared in advance (such as place, time, genre etc.). These expected categories were based on the disciplines expressed by the main UDC numbers and groups of auxiliary numbers. New categories emerged during the analysis (such as^ accessibility, ownership, experience). Slovenian translation of the UDC MRF 2006 was used to identify the appropriate UDC numbers for concepts expressed in tags.

Results and discussion

We expected that terms from 43 Things would be present less frequently in the UDC because the site focuses on people's wishes and plans as opposed to library material. We expected the opposite from Library Thing and Amazon because we selected traditional library material for their parts of the study sample. Our expectations proved to be incorrect. More concepts represented in tags of Del.icio.us and 43 Things were present in UDC than concepts represented in Amazon and Library Thing.

The largest number of tags from Del.icio.us could be found in the UDC class 0. Most tags from this sample were in the area of computer science. This was also observed by Spiteri (2007). This was followed by classes 3, 6 and 7. The highest number of tags from 43 Things belongs to UDC class 6 followed by classes 3 and 7. People seem to be interested in health, the arts and topics of a social and ethical nature.

Most tags from Library Thing fall into ^ class 7. Auxiliary numbers expressing place and time rank 2nd and class 8, expressing literary genre, ranks 3rd. Classes 0 and 9 follow. Class 0 represents carrier such as "DVD". Historical topics are featured in the background of two stories Cold Mountain and The Sound of Music. The results for Amazon are similar. Most tags are in class 7, expressing music and film. Auxiliary numbers expressing time and place follow and classes 8 and 9 rank 3rd.
In total, topic was most frequently expressed by tags in the sample. Names were the second highest ranked attribute of the documents. They were frequently assigned as tags for books, movies and soundtracks and they account for the largest share of concepts excluded from UDC.

Most frequently they were actors' names, authors, literary characters, names of people that were the topic of the work, owners' names and trade names. It is interesting that the names of actors

in the film version of the story were also assigned to books (e.g., Julie Andrews to the book *The Sound of Music*). These names appeared in Library Thing and Amazon. In Del.icio.us, software names appeared (Linux), while in 43 Things only 3 actors and 4 trade names appeared among the most popular tags. The title of the tagged work was repeated among tags for the same work in about 10% of cases in Amazon and Library thing. In this case, it was counted as the name attribute. Titles of other works were also mentioned in Amazon and Library Thing, establishing a connection within the bibliographic universe. It seems that some support for *Functional requirements for bibliographic records* (1999) could also be found in this detail.

Genre ranks third in total but is the second most frequently expressed attribute among Library Thing, Amazon and Del.icio.us tags. It only appears once in 43 Things. The form of the document is represented by terms, such as 'soundtrack,' 'movie' or 'news'. A related attribute, carrier, expresses the technology supporting the media, e.g. 'blue ray', or the media itself, e.g. DVD. Edition is only expressed a few times in Library Thing and Amazon tags. It is also related to the form of a document and can refer to the movie version, as opposed to the book. These two codes were kept separate because in some cases the movie tag would appear among tags assigned to the book. In such a case, this denotes a different edition of the same work, not the form of the information object it describes.

Evaluation is as frequently used an attribute as form in total. Awards are an attribute closely related to evaluation. In contrast to the subjective evaluation of items by users, awards are given by a professional authority and are regarded as an objective evaluation. While subjective evaluation was always discouraged in library catalogues, awards could be used as part of notes (area 7 of ISBD).

Collection or series is a relatively popular group of tags among users of Library Thing and Amazon, but not used in 43 Things or Del.icio.us. This attribute can be closely related to ownership when it is expressed with a tag like 'my DVDs'. This information is part of area 6 of ISBD, but not of subject description in library catalogues.
There are a number of attributes that seem to be important to users and are also part of the UDC. They are listed in table 1.

Table 1: Tags from the four analyzed folksonomies which are 1. present in the UDC, 2. can be part of UNIMARC format or 3. cannot be part of catalogue record

Present in UDC	Present in UNIMARC	Not part of catalogue record
Topic	Series	Evaluation
Name	Related work	Award
Genre	Carrier	Plan / action
Form	Edition	Ownership
Audience	Accessibility	Gift
Place	Audience	Occasion
Time		Experience
Carrier		

Attributes which could not be expressed with a UDC number constitute 24% of all tags in the samples. They represent awards, series or collection, edition, evaluation, experience, action, occasion, and purpose, availability, ownership and related work. Some of these categories form part of a bibliographic description (6% of all tags in the sample). However, none of the analysed sites adopted ISBD and as a consequence their bibliographic information is not complete. It seems that this information is actually important to users for information objects like books, sound and video recordings.

Among the remaining tags, which could not be represented by UDC numbers, evaluation alone holds a 9% share. It is unlikely that the evaluation would become part of a bibliographic or subject description. However, if this information was included among users' tags, it would probably be helpful to some library users.

Form of document and medium (carrier) are two attributes related to each other that are both usually part of the bibliographic description and can also be expressed with UDC numbers. The question is whether repetition of the same data in the bibliographic and subject description part of the catalogue record is reasonable and economically justifiable.
Repetition of the title proper of the work, which occurred in about 10% of cases in Amazon and Library Thing, seems particularly troublesome. Trant and Bearman (2008) reported a similar finding. When they compared user assigned tags to museum documentation, primary title was assigned in 25% of cases, creator in 7% and creation date in 2% of cases.

We were surprised by the high rank of evaluation tags for Library Thing and Amazon. Librarians carefully avoid such categorisation because it is subjective and could be offensive to the user. It also cannot be expressed with UDC numbers. Another surprise was the large number of names among tags for traditional library material in Library Thing and Amazon. Names can be added to the UDC numbers. However, the question is whether all the associated different roles names could and should be expressed in the UDC.

Ambiguity is a frequent complaint about folksonomies. Neologisms may be culturally biased. There may be some terms among the tags that we do not understand but that are part of the user's everyday vocabulary. This is why we did not name our category "neologisms," but rather "unclear terms". 1% of tags was unclear for Del.icio.us, 6% for Library Thing, and 4% for Amazon. Unclear tags represented 8% of the total tags for *Cold Mountain* in Amazon in 2007, according to an earlier study (Demšar et al., 2009; Matoh & Koželj, 2009). Spiteri (2007) does not analyse concepts that she is unable to understand, but found 10% of tags in Del.icio.us were neologisms, slang or jargon. We do not know whether these percentages are high or low. It can be expected that they would cause difficulties in searching and browsing. It takes time to include new concepts in established indexing languages.

Conclusion and suggestions

Golub and colleagues (2009) presented a project, where user tagging was enhanced by traditional indexing languages. They found that users like to utilise the assistance offered by those indexing languages. We would take their suggestion further in the direction of Smith's (2008) observations of structured reports in Buzzillions.com and Mefeedia.com. These services identified the most frequently used facets among tags and structured their input by these facets. We propose that the user is offered a structured form for adding his or her tags.

The structure would separate personal and geographic names (not distinguishing real and imaginary persons and places).

Authority files for personal and corporate names, and thesauri of geographical names could be offered here to help the user in selecting the appropriate form of name. UDC could be offered to help users in selecting topic, genre, form and medium. When recording time, users should be offered examples of standardized forms of reporting time. It should not be too difficult to link data from the ISBD area 6 or the UNIMARC field for collection. Suggestions could also be offered regarding awards. On the other hand, evaluation, action, purpose, and experience should remain entirely free from suggestions. We also believe that users should not be forced to use only the suggested terms. They should be able to either use the suggested term or write their own. It would also be appropriate to ask users to suggest similar works. The form should provide space for entering any other terms a user wishes to enter that are not appropriate to prescribed fields on the form. Our suggestion for a more structured interface with the user is also based on Markey's (2007 a, b) observation that people are generally inclined to work on the principle of least effort but, are likely to be quite persistent if the system supports them during their work, search and exploration.

Our finding that a larger proportion of tags used in Del.icio.us and 43 Things can be found in the UDC compared to Library Thing and Amazon tells us that the UDC is indeed universal and could support not only library catalogues but diverse social networks and digital repositories as well.

References

All URL valid as of 07 July 2009

Demšar, B.; Globočnik, M.; Hrast, Z.; Nestič, A.; Samobor Gerl, M.; Štaut, M.; Zupan, S.; Šauperl, A. (2009) Folksonomije = Folksonomies. *Šolska knjižnica* 19, pp. 15-23.

Functional requirements for bibliographic records: Final report. 1998. München: K.G. Saur.

Golub, K.; Moon, J.; Tudhope, D.; Nielsen, M. L. (2009) Enhancing social tagging with a knowledge organization system. Available at: http://www.ukoln.ac.uk/projects/enhanced-tagging/dissemination/entag-ifla-v3-final.pdf.

Markey, K. (2007a) Twenty-five years of end-user searching, Part I: Research findings. *Journal of the American Society for Information Science and Technology,* 58, pp.1071-1081.

Markey, K. (2007b) Twenty-five years of end-user searching, Part II: Future research directions. *Journal of the American Society for Information Science and Technology,* 58, pp. 1123-1130.

Matoh, R.; Koželj, A. (2009) Predstavitev in analiza dveh folksonomij [Introduction and analysis of two folksonomies]. *Knjižničarske novice*, no. 3. Available at: http://www.nuk.uni-lj.si/knjiznicarskenovice.

Spiteri, L. F (2007) The structure and form of folksonomy tags: the road to the public library catalogue. *Webology* 4: article no.41. Available at: http://webology.ir/2007/v4n2/a41.html.

Trant, J.; Bearman, D. (2008) Public and professional vocabularies: comparing user tagging with museum documents and documentation. In *The 7th European Networked Knowledge Organization Systems (NKOS) Workshop : Workshop at the 12th ECDL Conference, Aarhus, Denmark*, Sept. 19, 2008. Available at: http://www.comp.glam.ac.uk/pages/research/hypermedia/nkos/nkos2008/programme.html.