Exploring the Structure and form of Folksonomies: A case study of Marine Science Social Tags

Praveenkumar Vaidya
Research Scholar
Department of Studies in Library and Information Science,
University of Mysore
Mysuru
e-mail: vaidyapraveen@gmail.com

N. S. Harinarayana
Associate Professor
Department of Studies in Library and Information Science,
University of Mysore,
Mysuru
e-mail: ns.harinarayana@gmail.com

Abstract: *Social tags are proved to be important metadata for information retrieval. But due to arbitrary and haphazard nature of social tags, it becomes essential to categorise the tags in various forms. The structure and forms of social tags has to be studied for nature and behaviour of users. This research work is an attempt to understand the character and user behaviour of marine science users. For this research work 5150 marine science, scholarly articles listed in CiteULike were collected with 42,369 tags to examine and analyse. The results show that tags were found with various structure and forms to understand the character and behaviour of marine science taggers.*

Keywords: Folksonomies, Web 2.0, Social tagging, Marine Sciences, CiteULike

1. Introduction

Of late in context of Web 2.0, the user created metadata known as folksonomies or social tags became significant component of information retrieval. The folksonomy: the conflation of the words ‘folk’ and ‘taxonomy’ suggest a lightweight way of enhancing descriptions of online information resources (Trant, 2009; Wal, 2004). The tagging has been considered as a possible solution for enhanced information organisation system for a personalised use. The controlled vocabularies with hierarchical structure always played a vital role in information retrieval with precision. But introduction of social tags by users added another element of information retrieval even if the tags lacked the precision in information retrieval. Shirky (2005), asserts that “poly-hierarchy is essential to understanding the multi-faceted nature of meaning” of information objects and attributes to users-behaviour. Due to non-hierarchical structure of social tags, the user assigned tags come with various formats and configuration. Hence, the social tags must be studied about their structure and forms. The user-generated tags represent contextual information, subjective opinions and qualities or self-presentation and organisation aspects (Cantador, Konstas, & Jose, 2011). This study specifically explores the social tags assigned to scholarly journals listed in CiteULike by marine science researchers.

2. Review of Literature

There are many studies found which explore the structure and form of social tags. Spiteri (2007) examined how tags that constitute folksonomies are structured examine how the tags that constitute folksonomies are structured. The tags were evaluated against 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".

Guy and Tonkin (2006) studied the discrepancies like misspelt tags (e.g., libary, libray), badly encoded tags, such as unlikely, compound word groupings (e.g.,TimBernersLee), tags that do not follow convention in issues such as case and number, singular versus plural form (e.g., apple, apples), personal tags that are without meaning to the wider community (e.g., mydog) and single-use tags that appear only once in the database (e.g.,billybobsdog) of the tags extracted from social networking sites like flickr and del.icio.us and explained the difficulties of using tags for information organisation and retrieval.

Golder and Huberman (2006) identified the seven categories or kinds of tags and analysed the structure of collaborative tagging systems as well as their dynamical aspects. They also discovered regularities in user activities, tag frequencies and the kinds of tags used. The researchers presented a model which predicts the stable patterns of collaborative tagging.
Gupta, Li, Yin, & Han (2011), in their work, explained the categories of tags available in any tagset namely content-based tags, context-based tags, attribute tags, ownership tags, subjective tags, organizational tags, purpose tags, factual tags, personal tags, self-referential tags and tag bundles.

Kathuria (2011) in her study explores how the social tags are employed by users of LibraryThing. The full study analyzed 1231 social tags collected from 122 works on Asian women. Findings from this study showed that users construct a variety of tags in a very arbitrary way and usually assign tags according to their convenience. The researcher has access to plethora of tag construction available through these social tags.
Kipp and Campbell (2006) in their research work analyse "the tagging patterns exhibited by users of del.icio.us, to assess how collaborative tagging supports and enhances traditional ways of classifying and indexing documents. The authors discovered that tagging practices to some extent work in ways that are continuous with conventional indexing. However, the tags also indicated intriguing practices relating to time and task which suggest the presence of an extra dimension in classification and organization, a dimension which conventional systems are unable to facilitate".

The above works suggest the importance of tag structures available in tagsets rendered by users and necessity of study of such work in case of social tags extracted from scholarly social networking website namely CiteULike.

3. Objectives of the study

- To find out the configuration and morphology of users assigned tags to documents
- To find out character and behaviour users who assign the tags to documents
- To find out the nature of content-based tags

4. Research Methodology

For this work, the social tags are extracted from social networking site of CiteULike where the scholarly articles listed by various users and also assigned tags to marine science journal articles. "CiteULike is a Web-based tool to help scientists, researchers and academics store, organise, share and discover links to academic research papers. It has been available as a free Web service" (Emamy & Cameron, 2007). "Oceanography also known as oceanology is the study of the physical and biological aspects of the ocean. It is an Earth science, which covers a wide range of topics, including ecosystem dynamics; ocean currents, waves, and geophysical fluid dynamics; plate tectonics and the geology of the sea floor; and fluxes of various chemical substances and physical properties within the ocean and across its boundaries. These diverse topics reflect multiple disciplines that oceanographers blend to further knowledge of the World Ocean and understanding of processes within astronomy, biology, chemistry, climatology, geography, geology, hydrology, meteorology and physics. Paleoceanography studies the history of the oceans in the geologic past" (Wikipedia, 2018).

4.1 Research sample dataset

For this work, the researcher did collect a representative dataset from CiteULike social tagging system and study was conducted to categorise the tagging data. The dataset included 42,369 tags from 5,150 journal articles from 356 marine science journals listed in ASFA aquatic Science and Fisheries Abstract) journal's list. It was observed that the selected with tags were articles published during 1954 to 2015. The extracted tags were systematically transposed to Microsoft Excel worksheet to manipulate the data for requisite analysis. Table 1 shows glimpse of the data collected for this research work.

Table 1
Data collection statistics

Data collection	CiteULike Tags
Total tags	42369
Articles	5150
Number per article	8.23

Table 2
Top 20 journals with user tags

Journal Titles	Journal Frequency	No. of Tags in journal	Tags per article
Marine Ecology Progress Series	313	2965	9.47
Journal of Physical Oceanography	272	884	3.25
Limnology and Oceanography	242	904	3.74
ICES Journal of Marine Science / Journal du Conseil	238	1457	6.12

Progress in Oceanography	166	1848	11.13
Marine Policy	157	984	6.27
Deep Sea Research Part II: Topical Studies in Oceanography	122	2911	23.86
Journal of Marine Systems	122	2029	16.63
Hydrobiologia	109	1011	9.28
Fisheries Research	103	720	6.99
Deep Sea Research Part I: Oceanographic Research Papers	96	1161	12.09
Estuarine, Coastal and Shelf Science	92	913	9.92
Journal of Geophysical Research	87	206	2.37
Paleoceanography	86	396	4.60
Journal of Fish Biology	81	551	6.80
Marine Biology	78	738	9.46
Ocean and Coastal Management	77	647	8.40
Canadian Journal of Fisheries and Aquatic Sciences	77	443	5.75
Marine Pollution Bulletin	76	837	11.01
Journal of Geophysical Research: Oceans	75	187	2.49
Fish and Fisheries	72	444	6.17

Table 2 shows the top 20 journals where the tags were extracted from ASFA listed journals. By observing the Table it is understood that the ‘Marine Ecology Progress Series’ is top journal listed (313 articles) by majority of marine scientists which did attract 2965 tags with 9.47 tags per article. Interestingly, the journal ‘Deep Sea Research Part II: Topical Studies in Oceanography’ was listed less than the top journal (122 articles) but which provided 2911 tags with 23.86 tags per article. It can be inferred that the articles from the journal ‘Deep Sea Research Part II: Topical Studies in Oceanography’ attracted more tags from marine scientists and this journal may comprise of useful articles for the users.

5. Analysis and discussion

5.1 Single and Multi-word tags

It is very usual practice in folksonomies that the idea may be represented in an article either with single tag or multiword tag for any scholarly article. In case of CiteULike, the system prohibits users to assign multi-worded tags. In absence of such provision to assign multi-word tags to scholarly articles, the users tend to insert ‘hyphen’ or ‘underscore’ in between two or more concepts. Such tags with ‘hyphen’ or ‘underscore’ inserted between two concepts, generally reveal the importance of the article. For example, if three concepts are combined as ‘acid-base-balance’ or ‘acid_base_balance’, this is considered as three words for analysis when separator is removed. On the other hand, the term or tag ‘‘acidbasebalance’’ without any separator in between is considered as a single word for analysis purpose.

Table 3
Single and Multi-word Tags

Naming convention	**Number in dataset**	**Percentage**
Single-word tags		
Tags with only one word	33576	79.24
Single-word tags with two or words	95	0.22
Multi-Word tags		
Tags with hyphen as word separator	7290	17.2
Tags with underscore as word separator	1202	2.85
Compound tags	206	0.49
Total	N=42369	100

Table 3 indicates the presence of single and multi-word tags in the dataset. It is observed that 79.24% of tags were of single words. The general observation is the marine scientists preferred to assign tags with single words to most of the articles which they listed for their reference. But, 20.05% of tags were either inserted with either 'hyphen' or 'underscore'. It is understood that tags with 'hyphen' or 'underscore' represent multiple concepts present in the article. In this dataset the longest word with hyphen as separator comprised of nine words. The word string was '*chelonia-mydas-gas-exchange-body-temperature-diving-behavior-green*'. Similarly, the longest word with underscore as a separator was National Oceanic and Atmospheric Administration.

In the same Table 3, it was found that 206 (0.49%) tags were compound tags. Few users did not try to insert either 'hyphen' or 'underscore' as word separator but ended up creating 'compound tags' which included multiple concepts in one word. For example, For example, the tag *'accidents aircraft british '* included three words *'accident', 'aircraft'* and '*british*', but the user opted to assign this tag without inserting either 'hyphen' or 'underscore' as word separator. But such compound words are visibly cumbersome and disorganised to understand.

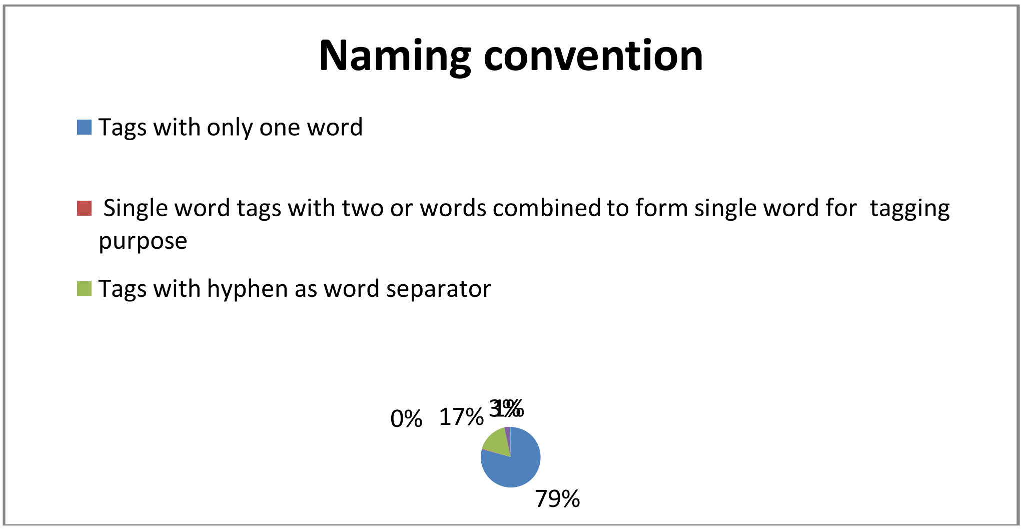

Figure 1: Single and multi-word tags present in the dataset

Many times it becomes very essential to add two words to complete the concept. In that case, as CiteULike does not provide an opportunity two represent in the form of tag. In such circumstances, the user will be compelled to assign two tags but by inserting a separator either 'hyphen' or 'underscore'. For example 'aquatic ecosystem' is a two-term concept and if the user just assign only one word as a tag 'aquatic', this does not give any clarity about the concept. But, when the word is associated with 'ecosystem', then 'aquatic ecosystem' term becomes a meaningful concept. Hence the tag has to be assigned with two terms with a separator of 'hyphen' or 'underscore'.

5.2 Subjective tags

Table 4 shows the subjective tags present in the dataset. The subjective tags are referred as the self-referential tags chosen by users.

Table 4

Subjective tags number and percentage in dataset

Subjective tags	**Number in dataset**	**Percentage**
Self-referential code tags	225	0.53
Personal task organization	127	0.23
Other tags	42003	99.24
Total	42369	100

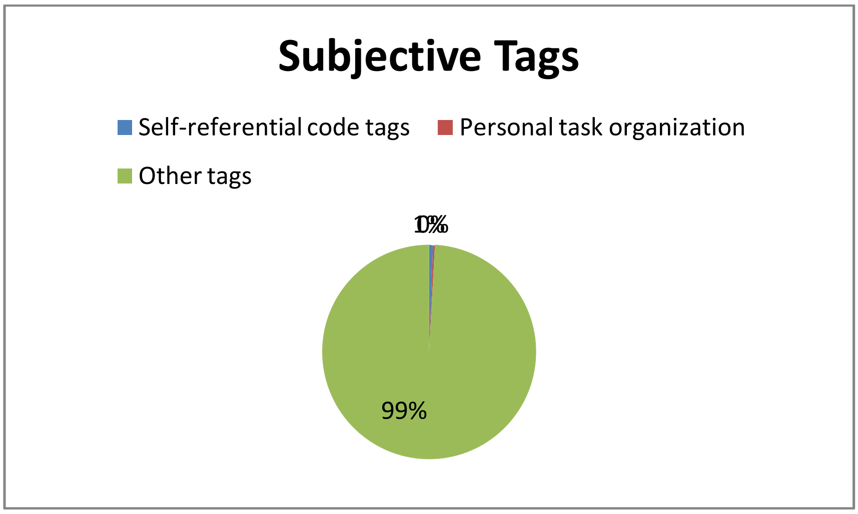

Figure 2. Pie-chart of Subjective tags present in dataset

In social tagging, users generally assign tags to resources of their choice, it may be a text or numerical. But these personal tags may sometime be meaningful to the users only not for readers. The user may assign tags to the scholarly articles to recall them when necessary for the reference. Such tags may be useful or meaningful for tag owners, but these tags were not for any analysis due to subjective nature. Table 4 indicates, out of 42,369 tags 225 (0.53%) tags are of personal or subjective in nature. Similarly, 127 (0.23%) tags were marked as personal task organisation tags in the dataset. It was very minuscule portion in comparison with the number of tags collected for this study.

For example, the tags like, *'1-toread'* and *'to do-mendeley'* suggest that the user had marked these scholarly articles for self-reference. The tag *'to read'* may not make any sense to others, but that tag indicates the task of user towards the marked article. Similarly, even the self-referential tags for example, *'16s-rdna'* may indicate *'Ribosomal DNA (rDNA)'*. Hence, such tags may provide underlying information about the tags.

5.3 Inconsistent or Factional tags

Due to arbitrary or haphazard nature of social tags, it is very common to find the inconsistent or factional tags in the dataset. The users generally assign tags which come to their mind instantly. The democratic property of tags provides the freedom to users to assign tags which can be recalled them whenever necessary.

Table 5 explains the various kinds of inconsistent or factional tags present in the dataset. It was found that 496 (1.17%) tags were idiosyncratic in nature. The idiosyncratic tags may be useful for tag owners but these tags were of no use for analysis because they did not indicate any meaning to other users. For example, the tags like '*aap','aatsr', 'adcirc'* did not make any sense to users, but tag owner had something in the mind to assign such tags.

Table 5

Inconsistent or Factional tags number and percentage in dataset

Inconsistent or Factional tags	Number in dataset	Percentage
Idiosyncratic tags	496	1.17
Stemmed tags	401	0.95
Foreign language tags	70	0.17
Typo error tags	14	0.033
Other tags	41408	97.73
Total	42369	100

Table 5 presents the existence of 401 (0.95%) of stemmed tags in the dataset, which is one of the common feature of folksonomies. Generally, such tags belong to technical or scientific terms, in which user may prefer to assign tags in stemmed manner. For example, the tag found in dataset *'pathol'*, the correct word may be 'pathology', but the user had stemmed the word and assigned this as *'pathol'*. Such user behaviour prevails in tagging system.

In the same tag set, the researcher found 70 (0.17%) foreign language tags, which were also considered as inconstant tags. The foreign language tags may be useful for tag assigner, but may not actually play any meaningful role to other users. As most of the listed articles in ASFA belong to English language, hence there was very little presence of foreign language tags in the dataset. For example, *'biologie', 'biologique', 'californie'* tags which belong to other than English language did not make any sense for analysis.

Table 5 also indicates the presence of 14 (0.033%) typo error tags. The CiteULike tag specific feature of assigning tags with correct spelling format. Hence, obviously, there was negligible portion of tags with spelling mistakes or typo errors. For example, *'Thermista', 'coatal', 'glacialis'* were some of the typo error tags which were removed from the dataset for further analysis.

Considering the huge tagset (42,369) used for this study, the presence of just 981 (2.32%) inconsistent or factional tags is very negligible.

6.2 5.4 Content-based tags

The content-based tags are centre of the folksonomies, in which many studies are conducted as reference point for information retrieval. The user-generated tags loaded with content can prove different set of metadata for effective retrieval. Table 7 signifies the presence of content-based tags in the CiteULike social tagset. Table 7 shows existence of 30,727 (72.52%) of content tags in tagset and 1,237 (2.92%) tags with abbreviation. The abbreviated tags are also considered as the content tags as the tags were assigned in the form of abbreviation. The abbreviation and acronyms are part of any subject indicators.

Table 7

Content-based tags

Content-based tags	**Number in dataset**	**Percentage**
Subject tags	30727	72.52
Abbreviation tags	1237	2.92
Other tags	10405	24.56
Total	42369	100

It is a surprise that the marine science literature does contain a lot of acronyms and abbreviations, but the results show otherwise. The tags in abbreviation were also considered as subject tags because 'it is a shortened form of a word or phrase' of the content used in the resource. The subject-based tags were used as resources by many researchers for their work because of the content value for the articles tagged in CiteULike. For example, *'calcification', 'cannibalism', 'decomposition'* is core subject tags and *'aqcx', 'asfa', 'baci'* are abbreviated tags. Such tags play an important role in information retrieval.

6. Conclusion

The social tags are new phenomenon in information retrieval which play vital role for many users. The tag categorisation helps to understand the nature of tags and user behaviour. In fact, the tags were assigned by users for their own reference, but they also help others to refer the resources listed in CiteULike. Hence, such user-generated tags help to enhance metadata and information retrieval. The character of marine science social tags of scholarly articles and user behaviour of marine scientists was studies in the research work. The future study may be conducted in the area of grammatical structure of tags, tag frequency and unique tag distribution.

References:

1. Cantador, I., Konstas, I., & Jose, J. (2011). Categorising social tags to improve folksonomy-based recommendations. Web Semantics: Science, Services and Agents on the World Wide Web, 9(1), 1–15. https://doi.org/10.1016/j.websem.2010.10.001

2. Emamy, K., & Cameron, R. (2007). Citeulike: A Researcher's Social Bookmarking Service | Ariadne: Web Magazine for Information Professionals. Retrieved October 5, 2012, from http://www.ariadne.ac.uk/issue51/emamy-cameron

3. Golder, S., & Huberman, B. (2006). Usage patterns of collaborative tagging systems. Journal of Information Science, 32(2), 198–208. https://doi.org/10.1177/0165551506062337

4. Gupta, M., Li, R., Yin, Z., & Han, J. (2011). An Overview of Social Tagging and Applications. In C. C. Aggarwal (Ed.), Social Network Data Analytics (pp. 447–497). Springer US. https://doi.org/10.1007/978-1-4419-8462-3_16

5. Guy, M., & Tonkin, E. (2006). Tidying up tags? D-Lib Magazine, 12(1), 1082–9873.

6. Kathuria, S. (2011). Content Analysis of Social Tags on Intersectionality for Works on Asian Women: An Exploratory Study of LibraryThing. Masters Theses. Retrieved from http://trace.tennessee.edu/utk_gradthes/988

7. Kipp, M., & Campbell, G. (2006). Patterns and Inconsistencies in Collaborative Tagging Systems: An Examination of Tagging Practices. Proc. Am. Soc. Info. Sci. Tech., 43(1), 1–18. https://doi.org/10.1002/meet.14504301178

8. Shirky, C. (2005). Ontology is Overrated: Categories, Links, and Tags. Retrieved from http://www.shirky.com/writings/ontology_overrated.html

9. Spiteri, L. (2007). Structure and form of folksonomy tags: The road to the public library catalogue. Webology, 4(2).

10. Trant, J. (2009). Studying social tagging and folksonomy: A review and framework. Journal of Digital Information, @prism.startingPage|virtual.citation.startpage@-@prism.endingPage|virtual.citation.endpage@.

11. Wal, V. T. (2004). You Down with Folksonomy? :: Off the Top :: vanderwal.net. Retrieved April 17, 2012, from
http://www.vanderwal.net/random/entrysel.php?blog=1529

12. Wikipedia. (2018). Oceanography. In Wikipedia. Retrieved from https://en.wikipedia.org/w/index.php?title=Oceanography&oldid=830590628