Is Tagging Effective?—Overlapping Ratios with Other Metadata Fields

Wooseob Jeong
School of Information Studies
University of Wisconsin -
Milwaukee, USA
wjj8612@uwm.edu

Abstract

The potential advantages of tagging have been addressed in numerous literatures, however still the effectiveness of tagging in information retrieval has not confirmed yet along with continuous intensive debates between advocates of new tagging systems versus traditional controlled vocabulary metadata. Despite all the potential advantages of tagging, the overlapping ratios between tags and the words used in other metadata fields, such as title and description, are significant. In this study, with the data from Youtube.com videos, the degree of overlapping is examined among the fields of title, description and tag, with additional questions about tagging, such as changes in numbers of words in each metadata field over time and the difference between web site promotion videos and non-promotion videos. The findings include 1) the number of words in each metadata fields have increased over time; 2) web site promotional videos have more words in each metadata fields than non-promotional videos; 3) more than 50% of words are shared among metadata fields including the tag field; 4) as much as 25% of the videos have the exactly same words repeated among the metadata fields. More similar studies with data from other social tagging sites are suggested to verify these findings.

Keywords: tagging; effectiveness; overlapping ratios; metadata; Youtube.com.

1. Introduction

Tagging has gained much attention in the library and information science field in recent years and the majority of literature on tagging has been positive on its advantages and potentials. A wide range of applications for tagging have been made for libraries, including academic libraries (Arch, 2007), school libraries (Brooks, 2008) and public libraries (Spiteri, 2007; Rethlefsen, 2007) and other areas of public information access including health and heath care education (Boulos and Wheelert, 2007) and government information (Rokolj, 2008).

The potential advantages of tagging have been addressed in numerous venues including monographs and popular magazines as well as academic journals. Gene Smith (2008) argued that tagging matters because it is popular, multifaceted, flexible, and social. Tagging is considered as the way to organize the stuff you don't have time to organize (Fallows, 2007), and it may bridge the huge gap between users' vernacular and the controlled terms in bibliographic records or taxonomy (Fichter, 2006). It could also provide alternatives to indexers' inconsistency, allowing for flexibility of users' term choice, particularly for newly-created ones (Matusiak, 2006), and can facilitate social aspects of online communities by so-called "social tagging" (Furnas et al., 2006). Boast and others (2007) witnessed users' positive use of tagging despite the professionals' concerns of chaos and disorder resulting from its unstructured manner.

A much smaller number of studies have questioned the effectiveness of tagging, motivation of tagging, and the dominant "personal" as opposed to the "social" aspect of tagging (Sen et al., 2007), along with the reaffirmation of the value of traditional controlled vocabulary (Macgregor and McCulloch, 2006). Dvorak (2005) warned of a potentially massive spam through tagging, referring to "the worst form of public graffiti," and Sanders (2008) was concerned that despite its potential, the benefit of creating additional finding pathways by tagging can be easily lost if there is no active participation. In terms of sloppiness of tagging, Hedden (2008) questioned the

https://doi.org/10.23106/dcmi.952109539

effectiveness of social tagging, while emphasizing the benefit of the pure semantic tagging in searching. Guy and Tonkin found that 40 percent and 28 percent of tags were erroneous in Flickr and del.icio.us respectively. They also found 8 percent of Flickr tags and 11 percent of del.icio.us tags to be plural forms (Guy and Tonkin, 2006).

However, few pointed out that significant overlapping exists between tagging and other existing metadata fields, such as title and description, which makes tagging lose its effectiveness. In their empirical study of tagging with the data from del.licio.us, Heymann and others (2008) concluded that they could not verify that the impact of tagging was significant because there was no evidence that tagging actually provides enough additional information to improve searching. Since their study compared the actual text of each web page with its tags, the study is not really about the overlapping among metadata fields.

In this study, with a set of data collected from YouTube.com, the significant overlaps are identified to demonstrate the ineffectiveness of tagging.

2. Research Questions

This study examines whether there is a significant redundancy of tagging against already established access points, such as title and description with an empirical data set. Unlike the majority of current research on tagging, which supports tagging's potential, this study questions the effectiveness of it, because of the redundancy with additional findings from the data.

Specifically the following research questions attempt to answer:

1. Do the numbers of unique words in metadata increase over time?
2. Are there any differences between videos with web site promotion and those without such promotion with regards to the numbers of unique words in metadata?
3. Are there any changes over time in the ratio of overlapping between unique words among metadata fields?

3. Methodology

The words in the metadata fields for the videos of Youtube.com were collected and analyzed. Youtube.com is considered as the most popular video sharing site on the web currently. A total of 17,130 videos' metadata were collected. The videos in the data were each uploaded by different members of Youtube.com. In other words, each video represents a unique member and there are no multiple videos from the same member. To filter aggressive spammers with lots of irrelevant web addresses, videos with only one or no URL included in their description field were selected.

For each video, the number of "unique" words in the fields of title, description and tag were counted. "Uniqueness" was used, because there are numerous occasions where the same words are repeatedly used among fields, which may confound the data analysis. For example, if a title has "baby, baby, baby …," while its tag has "baby crying," it is counted as title having "baby" and tag having "baby crying." Plurals and other varied forms of a word were counted as separate words, so "books" and "book" are different words. Special characters were removed and replaced by a space before word counting, so "father's" became "father" and "s." This may generate multiple meaningless words, but the effect was minimized by the uniqueness described before. Once all the numbers were counted, outliers were excluded due to the extreme numbers of words in the title, description, or tag field (Standard Deviation >2.5). As a result, 16,084 videos remained in the data set. The year each video was uploaded was also recorded.

Based on the numbers in the refined data set, the percentages of overlapping words between two fields among the three fields (title, description and tag) were calculated with different bases (6 combinations). In addition, after combining and extract unique words from the fields of title and description, the overlapping ratios between the combined word set and the field of tag were calculated (total of 8 combinations).

https://doi.org/10.23106/dcmi.952109539

TABLE 1. Shows the abbreviation of each ratio, which will be used in the data analysis

RTD	the percentage of words from the description in the title
RDT	the percentage of words from the title used in the description
RDG	the percentage of words from the tags used in the description
RGD	the percentage of words from the description used in the tag
RTG	the percentage of words from the tags used in the title
RGT	the percentage of words from the title used in the tag
RCG	the percentage of words from the tag used in the title-description combination
RGC	the percentage of words from the title-description combination used in the tag

4. Data Analysis

4.1. Numbers in Metadata Fields Over Time

The data shows the number of unique words in a field has increased for all three metadata fields. This may support the argument that the importance or usefulness of tagging was gained by the members of Youtube.com over time, while the usefulness is still questionable. It may imply self-promotion or in extreme cases, "spam" with more words to increase visibility in search. Table 2 and Figure 1 show the numbers and the trend.

TABLE 2. The average number of words in metadata fields over time

	Title	Description	Tag	n[1]
2005	3.52	12.47	4.73	62
2006	4.15	14.43	5.95	2642
2007	4.50	14.10	6.68	6889
2008	4.82	15.63	7.75	5277
2009	5.12	17.53	8.92	1214
Total	4.59	14.91	7.07	16084

[1] There aren't too many videos uploaded in 2005 at Youtube.com. The number of videos for 2009 is small because the data was collected in early 2009.

https://doi.org/10.23106/dcmi.952109539

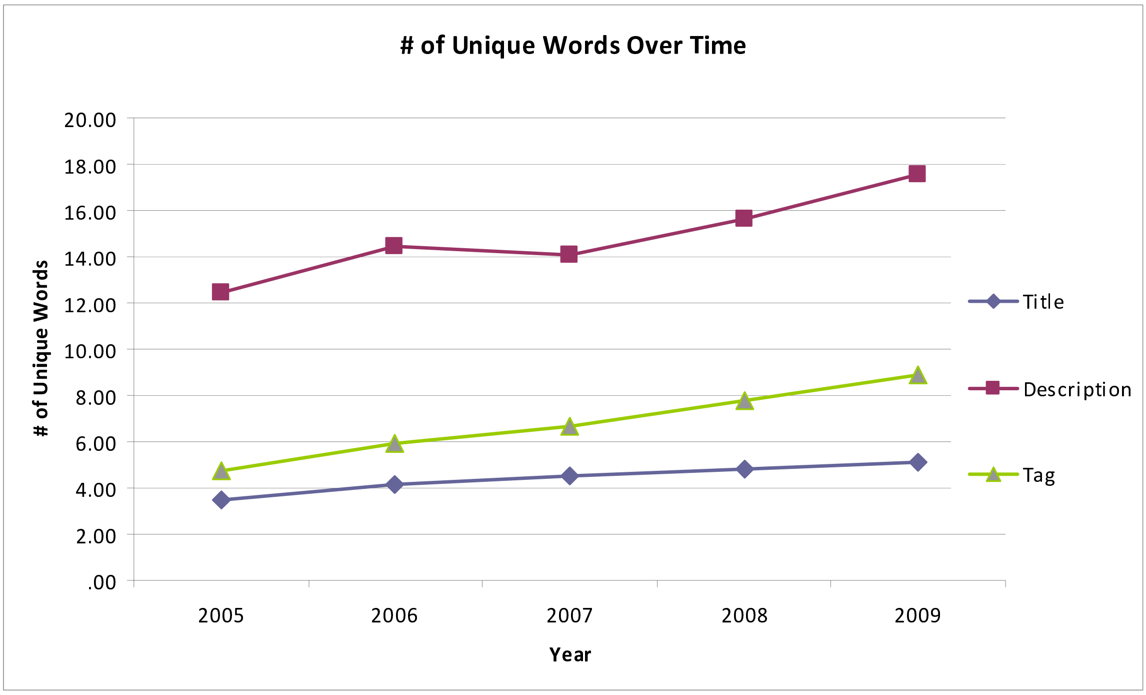

FIG. 1. The average number of words in metadata fields over time.

4.2. Difference between Videos Promoting a Web Site and Those with No Web Site

Since there are numerous self-promoting videos which work like advertisements, a question arises: Do taggers put more words for self-promoting videos (with a web address in its description field? The data confirmed that is indeed the case. Table 3 and Figure 2 show the numbers and the differences.

TABLE 3. The average numbers of unique words in metadata fields with vs. without web address

	Title	Description	Tag
Non-Promotion (without web address)	4.53	14.24	6.84
Promotion (with web address)	5.29	22.48	9.68

https://doi.org/10.23106/dcmi.952109539

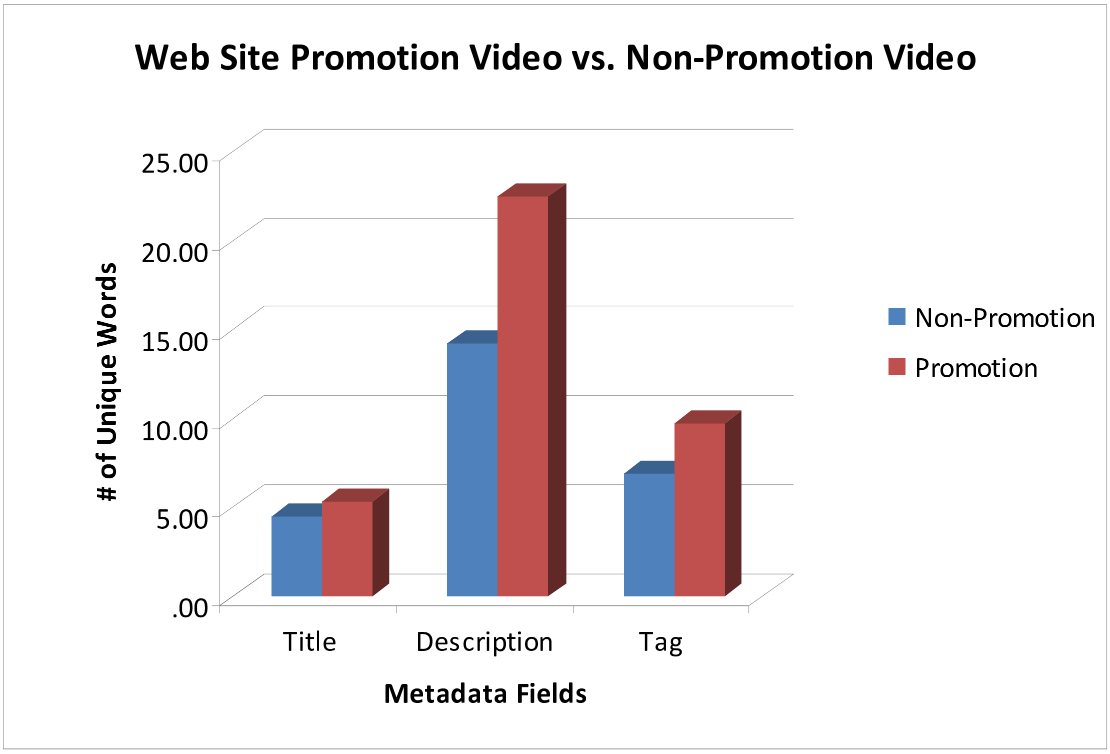

FIG. 2. The average numbers of unique words in metadata fields with web site promotion

4.3. Overlapping Ratios among Metadata

In terms of the overall overlapping ratios, the data show that 54.97 percent of the words from the title-description combination appear in the tag field as well, 52.93 percent from the title appear in the tag field, and 49.11 percent from the title appear in the description field (Figure 3). More than half of the words are shared among fields, which means taggers typed in the same information repeatedly. It is worthwhile to note that 39.77 percent of the words from the tag also appear in the title and 39.25 percent from the tag reappear in the description. The title field appears to play the most important role in terms of the key words.

https://doi.org/10.23106/dcmi.952109539

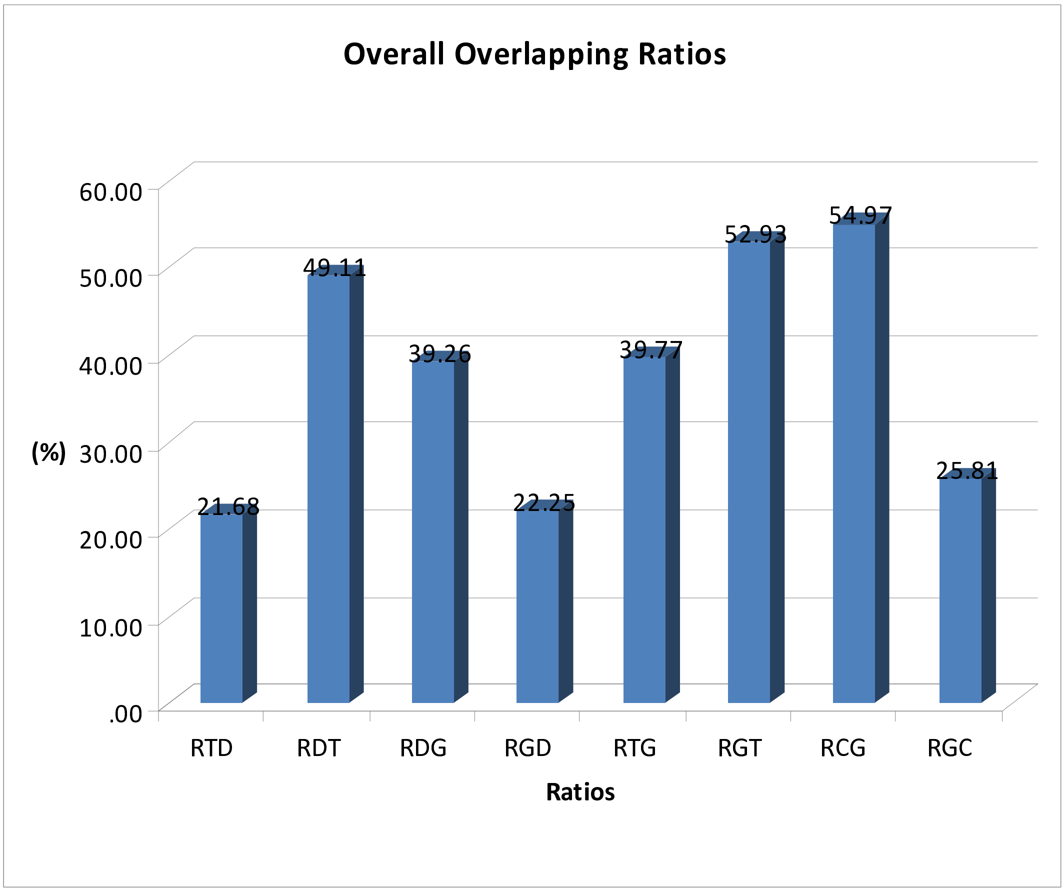

FIG. 3. Overall overlapping ratios

The trends or changes of the overlapping ratios over time might be interesting as well. The data show that the ratios generally became a little bit smaller each year, resulting from the significant growth in the number of words for each field as seen in the previous section.

TABLE 3. Overlapping ratios by year

	RTD	RDT	RDG	RGD	RTG	RGT	RCG	RGC
2005	23.94	59.69	43.69	19.87	45.85	51.35	59.48	23.31
2006	22.04	50.86	42.03	21.45	42.54	52.59	57.89	24.41
2007	21.91	48.58	40.37	22.65	41.19	53.20	56.90	26.39
2008	21.84	49.35	37.51	22.73	37.47	52.86	52.19	26.19
2009	18.83	46.76	34.31	19.73	35.38	52.56	49.58	24.09
Total	21.68	49.11	39.26	22.25	39.77	52.93	54.97	25.81

https://doi.org/10.23106/dcmi.952109539

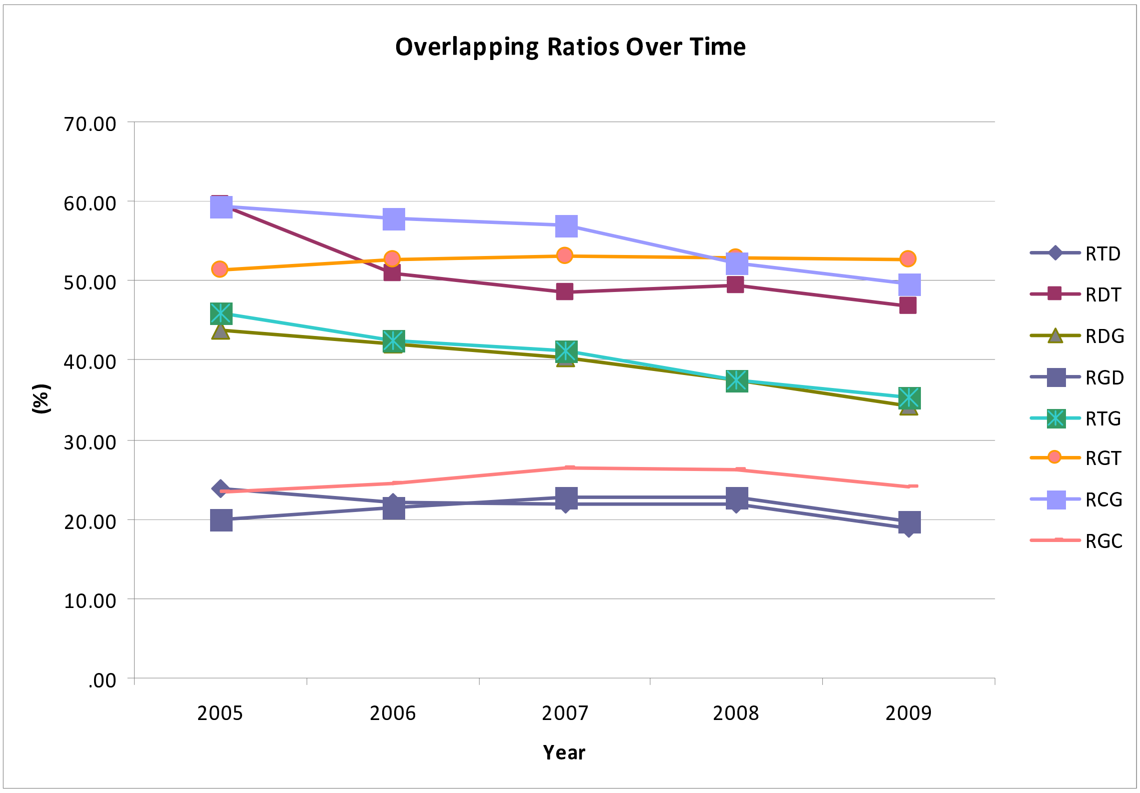

FIG. 4. Overlapping ratios by year

4.4. Absolute Overlapping Ratios

As much as 25% of the videos have the exactly same words repeated among the metadata fields, such as in the ratio between the description and the title and between the tag and the title. Table 4 shows the number of videos which have 100% overlap ratio among the fields.

TABLE 4. 100% overlap cases among the metadata fields (n=16084)

RTD	748	5%
RDT	3953	25%
RDG	2016	13%
RGD	550	3%
RTG	2110	13%
RGT	3726	23%
RCG	3514	22%
RGC	415	3%

5. Discussion

Significant overlapping ratios were found among the fields of title, description and tag for videos at Youtube.com. With more aggressive word counting, such as collapsing plurals and tenses into a single word, the overlapping percentage could be higher.

In fact, tagging is not a new concept. Many journals, conference proceedings, and even dissertations have required keywords from authors to improve their information retrieval

https://doi.org/10.23106/dcmi.952109539

performances in databases for years. Unfortunately, their efforts in the concept of keywords do not seem successful so far, and the "new" concept of creators tagging their work does not seem to show much improvement, due to its significant overlapping with the existing metadata fields like title and description.

The tag field plays a role between the title field and the description field. If that is the case, a suggestion would be that more words should be added to the title field or the description field, not creating the third field for tagging. Much richer information in both fields would help the search performance without the overlapped words in the tag field. If the connection function in tagging is the main benefit, the same mechanism should be used for the words in title or description.

The technical aspect of tagging is rather disappointing, especially for not allowing multi-word tags. As a result, many unnecessary variations of terms have been generated. This kind of utility can be easily implemented, even though the current system may be favorable for "lazy" users who do not use additional delimiters, such as commas.

Requiring multiple fields of metadata can be a burden to authors, submitters, and indexers. Personal experiences in the process of submission for ACM (Association for Computing Machinery) conferences easily confirm this cumbersomeness, especially with significant overlapping ratios among fields.

During the data collection, interesting observations were made. Educational videos (news from National Geographic or documentaries) have long descriptions but few tags, while personal videos (cats, children, etc) tend to have few of both. Also, a lot of descriptions are just song lyrics. Some videos have had the same words in multiple languages as their tags. There are lots about babies and animals in personal videos and personal videos tend to have tags without any sincerity. In the worst cases, even personal videos have totally irrelevant tags solely because the members who uploaded those videos want more visibility and exposure, to increase numbers of visitors.

6. Conclusion

This study shows that significant overlapping ratios exist among the metadata fields of title, description, and tag in videos of Youtube.com, questioning the effectiveness of tagging for organization and retrieval of information. It also shows additional pattern of tagging, such as increased numbers of words used for each field, difference between web site promotion videos and other videos, and the changes in overlapping ratios among fields over time.

Many previous studies have looked at the body of tag texts only, to explore why and how taggers do tagging. Further studies will conduct surveys and interviews with taggers to learn more about their intention and behavior in tagging. Also in further studies, the overlapping ratios issue should be examined with other social tagging data, such as tags from Flickr.com, to verify the findings of this study.

Acknowledgements

The author wishes to acknowledge the invaluable assistance of Soohyung Joo, Hohyon Ryu, Jaehyun Park, Jihee Beak, and Jeanette C Jordan, Graduate Assistants, in the data collection and the data analysis for this paper.

References

Arch, X. (2007). "Creating the academic library folksonomy: put social tagging to work at your institution." College & Research Libraries News, February: 80-81.

Boast, R., Bravo, M., and Srinivasan, R. (2007). "Return to Babel: emergent diversity, digital resources, and local knowledge." The Information Society, 23: 395-403.

Boulos, M and Wheelert, S. (2007). "The emerging Web 2.0 social software: an enabling suite of social technologies in health and health care education." Health Information and Libraries Journal, 24: 2-23.

https://doi.org/10.23106/dcmi.952109539

Brooks, L. (2008). " "Old school" meet school library 2.0: bump you media program into an innovative model for teaching and learning." Library Media Connection, Apr/May: 14-16.

Dvorak, J. (2005). "To tag or not to tag, that is the question." PC Magazine, May 24, 2005. (http://www.pcmag.com/article2/0,1759,1819101,00.asp)

Fallows, J. (2007). "Tag teams." The Atlantic, Jan/Feb: 163-165.

Fichter, D. (2006). "Intranet applications for tagging and folksonomies." Online, 30(3): 43-45.

Furnas, G., Fake, C., Ahn, L., Schachter, J., Golder, S., Fox, K., Davis, M., Marlow C., and Naaman, M. (2006). "Why do tagging systems work?" CHI '06 extended abstracts on Human factors in computing systems: 36-39. Montreal, Canada.

Guy, M. and Tonkin, E. (2006). "Folksonomie: tidying up tags?" D-Lib Magazine, 12(1). Available at: http://www.dlib.org/dlib/january06/guy/01guy.html.

Hedden, H. (2008). "How semantic tagging increases findability." EContent, 31(8): 38-43.

Heymann, P., Koutrika, G., and Gracia-Molina, H. (2008). "Can social bookmarking improve web search?" Proceedings of the international conference on Web search and web data mining: 195-206. Palo Alto, California, USA.

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

Matusiak, K. (2006). "Towards user-centered indexing in digital image collections." OCLC Systems & Services: International Digital Library Perspectives, 22(4): 283-298.

Rethlefsen, M. (2007). "Tags help make libraries del.icio.us: social book marking and tagging boost participation." Library Journal, Sep. 15: 26-28.

Rokolj, T. (2008). "Social bookmarking and folksonomies: possibilities for government information?" Documents to the People, 36(2): 21-24.

Sanders, D. (2008) "Tag – You're it!" American Libraries, 39(11): 52-54.

Sen S., Harper F., LaPitz A., and Riedl J. (2007). "The quest for quality tags." GROUP '07: Proceedings of the 2007 international ACM conference on Conference on supporting group work: 361-370. November 4-7, Sanibel Island, Florida.

Smith, G. (2008). Tagging: people-powered metadata for the social web. New Riders: Berkeley, CA.

Spiteri, L. (2007). "The structure and form of folksonomy tags: the road to the public library catalog." Information Technology and Libraries, 26(3): 13-25.

https://doi.org/10.23106/dcmi.952109539