Assessing the Reuse Value of Socially Created Metadata for Image Indexing

Beseki Stvilia
Florida State University
Besiki.Stvilia@cci.fsu.edu

Corinne Jörgensen
Florida State University
corinne.jorgensen@cci.fsu.edu

Shuheng Wu
Florida State University
sw09f@my.fsu.edu

Final Report
2010 OCLC/ALISE Library and Information Science Research Grant Project

February 15, 2012

Suggested citation:
Stvilia, Beseki, Corinne Jörgensen, and Shuheng Wu. 2012. "Assessing the Reuse Value of Socially Created Metadata for Image Indexing." 2010 OCLC/ALISE research grant report published electronically by OCLC Research. Available online at: http://oclc.org/research/grants/reports/2012/stvilia2012.pdf

Introduction

With increasing system flexibility now allowing end-users to add their own descriptive terms to items in a collection, a frequently asked question is what role (if any) these additional terms play in enhancing description and access. There have been ample suggestions in the literature that terms added to documents from Flickr and Wikipedia can complement traditional methods of indexing and controlled vocabularies. These terms are popularly called tags or referred to as metadata. At the same time, adding new metadata to existing metadata objects may not always add value to those objects. For the purposes of this research, we used the collective term "social terms" to group end-user contributed content (tags and metadata). This research is a step towards establishing a framework for evaluating the value of socially created metadata to enhance the quality of traditional knowledge organization systems (KOS). Because images have been particularly effective in stimulating user involvement in tagging, this paper evaluates the potential value of end-user-generated metadata from Flickr (tags) and the English Wikipedia (related article terms) to enhance traditional KOS such as the Thesaurus for Graphic Materials (TGM) and the Library of Congress Subject Headings (LCSH) by providing additional access points.

Furthermore, the usefulness and quality of metadata is recognized as contextual (Greenberg, 2001; Stvilia, Gasser, Twidale, Shreeves, Cole, 2004). To facilitate reuse of social metadata for image indexing it is essential to determine what terms are useful to the user and how to identify useful terms for a particular image inexpensively – that is, automatically (Jörgensen, 1995, 1998). This study examined relationships among categories of image tags, tag attributes (length and assignment order), and users' perception of usefulness of tags.

The outcomes of the project included (1) models and methods for assessing adding value of social terms to traditional KOS; (2) identifying distributions of grammatical and cognitive types is social terms; (3) measuring and comparing the quality and value of knowledge organization systems, including machine-generated and expert-created controlled vocabularies; (4) identifying the types of activities that might lead to metadata and knowledge creation in social content creation and sharing systems; (5) examination of relationships among user demographics, index term characteristics and user's perception of term usefulness in image indexing. The findings of this study can inform the design of controlled vocabularies, indexing processes, and retrieval systems for images. In particular, the findings of the study can advance the understanding of image tagging practices, tag facet/category distributions, relative usefulness and importance of these categories to the user, and inexpensive mechanisms for identifying important terms. The use of knowledge representation and organization tools (thesauri, taxonomies, ontologies) is ubiquitous among information professionals. Therefore, the outcomes of this study are relevant to and beneficial in any field in which indexing, thesaurus, or ontology construction and maintenance are routine activities.

Research Questions and Design

The original proposal included the following research questions:

1. Are the types of concepts and terms found in Flickr metadata different from the types of concepts and terms found in representative textual documents and reference sources such as Wikipedia? (i.e., Are the concepts medium specific?) What are the optimal ways of integrating concept and term data from these sources?

2. Will combining evidence from Flickr collection-level metadata (group titles and description) with item-level metadata (photo titles and tags) help generate a better quality thesaurus than using only item-level metadata?

3. *Can a thesaurus generated from Flickr and Wikipedia metadata be useful to users when they describe and search for images? Does integrating the concepts, terms, and relations (i.e., a thesaurus) extracted from the Flickr and Wikipedia metadata with TGM metadata add value to the TGM, and if so, what is the structure (i.e., types of concepts and relations) of that added value?*

However, after receiving reviews from the ALISE/OCLC grant review panel, we decided to follow the reviewers' suggestion and reduce the scope of the study and focus on the third group of research questions.

Study design

The data in this study consisted of tags and comments associated with photos from the Library of Congress photostream in Flickr. A total of 28,303 unique tags and 43,152 comments associated with 7,192 photos were downloaded on September 13, 2009, from the Library of Congress Flickr photostream.

The harvested Flickr tags were preprocessed before matching with the TGM and LCSH. In particular, multiterm concatenated tag sets were recursively split into individual terms based on the terms and inflections from the WordList (http://wordlist.sourceforge.net/). In addition, the set was cleaned of all URLs and tags with fewer than 3 characters. This reduced the number of tags in a set to 20,946. Finally, both the Flickr tags and the controlled vocabulary terms from the TGM and LCSH were stemmed using the Porter Stemmer algorithm (Porter, 1980). To collect, preprocess, match, and analyze the data sets, the study used a Flickr Java application programming interface (http://sourceforge.net/projects/flickrj/), the Atlas.ti and Stata software, and Java codes developed by one of the researchers.

Evaluation of the intrinsic and relational quality of the tags was guided by a previously developed framework of information quality assessment (Stvilia, Gasser, Twidale, & Smith, 2007). Intrinsic quality measures the internal characteristics of a tag itself in relation to some general reference standards in a given culture, such as the WordNet (a general-purpose comprehensive ontology of word senses; http://wordnet.princeton.edu/). In contrast, relational quality measures relationships among the tag and some aspects of its usage context, and the reference source of that usage context. The relational quality of the photostream's folksonomy was evaluated relative to the context of maintaining traditional controlled vocabularies. In particular, the study assessed the suitability of the folksonomy as a source of terms for the TGM and LCSH.

The intrinsic quality of the folksonomy was evaluated as the ratio of valid terms, whereas the relational quality was evaluated as the ratios of valid terms not present in the baseline controlled vocabularies (the TGM and LCSH). The complete set of tags from the Flickr photostream (20,946 tags) was matched to the terms from the TGM (13,317 terms, including both subject and genre terms; http://www.loc.gov/rr/print/tgm1/), the LCSH (696,358 preferred and alternative terms or key phrases; http://id.loc.gov/authorities), and also the WordNet ontology (version 2.1; 207,016 terms) to identify the scale and the nature of overlaps and differences between the tags and the terms from these KOSs. The study used a subsumption operator to match the Flickr tags to the controlled vocabularies. To be a match, a Flickr tag had to be a subset of the vocabulary term.

To obtain a more nuanced description of the intrinsic quality of the tags and their suitability in updating or enhancing the controlled vocabularies, the study content analyzed random samples of 300 terms from each set: the folksonomy, the TGM, and the LCSH. The samples were coded for grammatical types and the cognitive categories of nouns of Rosch, Mervis, Gray, Johnson, and Boyes-Braem (1976): Basic, Subordinate, and Superordinate. The term type distributions from the folksonomy were then compared with and contrasted to the term type distributions in the TGM and LCSH samples. The important characteristic of the Basic category is that it contains most commonly occurring or commonly used terms for concepts. According to the theory of controlled vocabulary construction (Lancaster, 2000; Soergel, 1974), preferred terms for concepts in a controlled vocabulary usually are selected from the most frequently used terms. Hence, Basic category nouns can be a valuable source of preferred terms in controlled vocabulary construction and maintenance.

In addition, the research designed and carried out controlled experiments. To evaluate the added value of social terms the study was guided by Taylor's (1986) value added model of information systems and an a model of metadata value (Stvilia & Gasser, 2008). The study adapted the experimental designs used by Jörgensen (1998) and Chen et al. (1995). Three separate experiments were used, a *description* task, a *search* task, and a *query-development* task in which participants documented their information seeking processes. The experiments involved 35 participants recruited from the College of Communication and Information at Florida State University. The participants were given 10 sampled photographs and asked to describe each photograph spontaneously by assigning tags. A copy of the modified Steve tagger software (http://sourceforge.net/projects/steve-museum/) was used by the participants to record the tags for each photograph. Next, to evaluate the perceived value of social terms, the subjects were presented with a set of pre-assigned index terms, including terms from the Flickr and Wikipedia, and were asked to rate each individual term on its usefulness for the task of describing the content of the photographs. In particular, the task instrument asked the participants whether they agreed that a particular term was useful in describing the content of the photograph. The participants had to answer the question on a five-point Likert scale (i.e., 'strongly disagree', 'disagree', 'neutral', 'agree', 'strongly agree'). At the end of each description experiment, post-session semi-structured interviews were conducted with participants to elicit additional information about their perceptions regarding the concepts of index term usefulness and value, and the decision-making process involved with rating the usefulness of pre-assigned index terms in the description task. Two weeks after completing the description experiments, the same group of participants was asked to complete the second task, the search task. The participants were given the same 10 photographs used in the description experiment and asked to approximate a search for a known photograph. The photographs were shown in a sequence, and for each photograph, the subjects were asked to formulate a query that, in their opinion, would allow them to locate the photograph with a hypothetical search engine with the least effort. Finally, the participants were asked to write autoethnographies (Cunningham & Jones, 2005). In this task, participants were asked to develop four queries of predefined types to find a relevant image or images using their favorite search engine. The participants were asked to document the information-seeking processes that led to the search queries in concise autoethnographies.

The study evaluated two facets of the added value of social terms. First, the study assessed the subjective or perceived value of the social terms by comparing the participant ratings of the usefulness of the social terms with a baseline that was set to a neutral rating (i.e., 3). In addition, the researchers evaluated the added value of social terms objectively by measuring the degree of additional match or coverage of user terms the social terms provided.

The experimental setup used in the study included a sample of 10 photographs selected from a set of 7,192 photographs from the LoC Flickr photostream, downloaded on September 13, 2009. The photographs were pre-indexed by the researchers using the following sources: the TGM, the LCSH, the set of tags Flickr members assigned to the photograph, the folksonomy of the LoC photostream on Flickr (LoC folksonomy), the complete Flickr database exposed through the "relatedTags" procedure of the Flickr application programming interface (http://www.flickr.com/services/api/), and the English Wikipedia. The LoC Photostream folksonomy was constructed from the complete set of LoC Flickr photostream tags (20,946 tags as of September 13, 2009) and their (i.e., tag) pairwise co-occurrence information in the photographs of the photostream as of September 2009. The tag co-occurrence information was used to determine the strength of a semantic relationship between a pair of tags, represented as a mutual information score (Cover & Thomas, 1991).

To obtain the English Wikipedia terms, the researchers used the Wikipedia Miner code libraries (Milne & Witten, 2008) with a July 2009 copy/dump of the English Wikipedia database. The English Wikipedia is the largest community-constructed and community-maintained encyclopedia; as of July 30, 2009, it contained approximately 3 million articles and more than 17 million pages in total, including disambiguation, redirect, and discussion pages. To preindex the sampled photographs for the experiments, the researchers used a snowball approach (see Figure 1). The sample was first preindexed with terms from the TGM and LCSH by two researchers independently. When assembling a set of controlled vocabulary terms for a photograph the researchers took into consideration index terms assigned to the photograph by the LoC, if any, as well as seeking and assigning any additional preferred, alternative or broader terms relevant to the photograph's content. To identify relevant terms the researchers used both the LoC's Web interface (http://id.loc.gov/) and local copies of the controlled vocabularies downloaded from the LoC's Website. After independently assigning controlled vocabulary terms, the researchers compared their completed sets, resolved indexing differences and determined final sets of controlled vocabulary terms for each photograph in the sample. These controlled vocabulary terms, combined with the tags the Flickr members assigned to the sampled photographs, were then used as a "seed" to iteratively obtain additional related terms algorithmically. These were first obtained from more contextual sources (the LoC folksonomy and the complete Flickr database) and then from a more general source (the English Wikipedia). A detail description of the procedure used to preindex the sample can be found in (Stvilia, Jörgensen, & Wu, in press).

To conduct the experiments, the study used modified Steve tagger software. New functionalities were added to the software by one of the researchers to load the index terms pre-assigned to the photographs and to match those terms with the tags provided by participants in the experiment. The software also allowed the participants to rate the pre-assigned index terms on their usefulness for the experimental task (i.e., describing the content of a photograph). A separate set of Java codes was developed to preprocess and match term sets and to calculate set overlap scores. The project used Stata software (StataCorp LP, College Station, TX) for statistical analysis and modeling. Qualtrics survey software (Qualtrics, Provo, UT) was used to collect data from the search experiments. Finally, a pilot study was conducted with three subjects to test and refine the design of the experiments and the exit interview protocol (Stvilia, Jörgensen, & Wu, in press).

Findings

To address the research questions the study started with assessing the intrinsic and relational quality of the Flickr folksonomy tags as well as their grammatical and cognitive types. A manual analysis of the random sample found that a full 15.3% of the Flickr tags were misspelled or invalid terms, including foreign words. Noun terms constituted 32.2% of the set, of which 26.3% were of the Basic type.

A comparison of the complete sets revealed that only 21% of the user-generated tags (4,477 tags) had a match in the TGM. The degree of overlap with WordNet was slightly higher, 33% of the tags (6,824 tags). The degree of overlap with the LCSH, however, was 45% (9,575 tags).

In addition, the distribution of the term types in the folksonomy differed from the term type distributions in the TGM and the LCSH. In particular, chi-square tests on the aggregated set showed that the distributions of the grammatical types were significantly dependent on the source (i.e., Flickr, TGM, LCSH). In particular, the analysis showed that the chances of complex terms occurring were higher in the TGM than in the folksonomy ($p < 0.001$) but were still lower than in the LCSH ($p < 0.001$). The folksonomy, however, contained a greater share of named entities than the TGM ($p < 0.002$) but still contained fewer than in the LCSH ($p < 0.001$).

Chi-square tests of the noun categories on source (i.e., Flickr, TGM, LCSH) showed significant dependencies for the Basic and Subordinate categories ($\chi^2 = 46.8$; $p < 0.001$; $\chi^2 = 97.7$; $p < 0.001$), but not for the Superordinate category. In particular, the term being of the Basic type increased the odds of the term being from the Flickr set relative to the TGM and the LCSH. The term being of the Subordinate or Superordinate types, however, increased the odds of its source being from the TGM or the LCSH rather than the folksonomy. A comparison of the TGM and LCSH samples showed that there were no statistically significant differences in the noun type distributions between these vocabularies, except for the Subordinate type. The odds of a Subordinate noun being from the LCSH were significantly higher compared with the odds of it being from the TGM.

Thus, in general, the photostream's folksonomy contained a relatively larger share of Basic nouns and smaller shares of Subordinate or Superordinate nouns than the TGM and the LCSH. In addition, nouns in the TGM were more general than those in the LCSH.

In addition, the researchers manually analyzed random samples of 300 tags with no match in the controlled vocabularies. As expected, most of the no-matches (64%) to the TGM were proper nouns (named entities) and complex terms. The analysis also revealed regular nouns (8%) with no match in the TGM. Approximately 2% of those nouns (e.g., rods, pause, gauge) belonged to the Basic category. The set of no-matches to the LCSH included 2% noun terms, none of which belonged to the Basic category. The distribution of grammatical types of the no-matches to the LCSH was very similar to that of the no-matches to the TGM.

The analysis identified valid terms in the Flickr set that had no match not only in the TGM and LCSH, but also in the English Wikipedia. As expected, because of the historical nature of the Library of Congress collection, the terms represented entities and concepts from the past, such as Playograph, Grid Graph, and Aerocar (Stvilia & Jörgensen, 2010).

Next, the study examined the added value of social terms relative to the TGM and LCSH using controlled experiments. The sets of index terms assigned to the sampled photographs from these controlled vocabularies were supplemented with related terms from Flickr and the English Wikipedia. The aggregate sets of terms were then used to assess the subjective (perceived) and objective value of the social terms for image indexing. The participants in the controlled experiments were asked to assess the usefulness of these terms for the task of describing the content of the photographs. The study also measured the degree of added coverage of participant terms provided by the addition of social terms.

The results of the experiments showed that the social terms added value to the controlled vocabularies in the context of image indexing and retrieval. Participants perceived the social terms as generally useful. The median rating for the social terms was significantly higher than the baseline rating (i.e., the 'neutral' rating). Furthermore, the addition of social terms resulted in double the coverage of participant terms on average compared with the coverage of participant terms provided by the controlled vocabulary terms alone.

The results of the experiments showed that the participants valued the controlled vocabulary terms more highly than they valued the social terms, suggesting that the TGM and LCSH capture the most important and preferred terms. In addition, this study found that query terms from the search experiments were best covered by the terms participants used in the description experiments (Stvilia, Jörgensen, & Wu, in press).

To investigate the structure of the added value of social terms in image indexing, the pre-assigned terms rated by participants were coded by two of the researchers according to Jörgensen's coding scheme of broad categories composed of several types of attributes (Jörgensen 1995; 1998). For the evaluation task participants rated the pre-assigned index terms for each of the 10 photographs on a five-point Likert scale according to their usefulness for describing the content of the image. These results were interesting in that they demonstrated that frequency of a code may not necessarily correlate with the usefulness ratings that participants assigned. For instance, while Abstract category terms accounted for barely over 1% of the terms, by percentage, they were rated third highest in usefulness. Similarly, while Objects were consistently among the most described attributes, they were ranked lower than other categories such as People and people-related attributes and Description terms. (see Jörgensen, Stvilia, & Wu (2011) for more detail.)

Future Research

Future research related to the current study will investigate the first two groups of research questions of the original proposal. In particular, the future research will examine and compare different approaches and algorithms of constructing a thesaurus from Flickr and Wikipedia metadata. This will include an investigation of whether combining evidence from Flickr collection-level metadata (group titles and description) with item-level metadata (photo titles and tags) helps generate a better quality thesaurus than using item-level metadata only.

Project Publications and Presentations

As of December 2011, the project has produced the following publications and presentations:

- Jörgensen, C., Stvilia, B., & Wu, S. (accepted). Relationships among perceptions of term utility, category semantics, and term length and order in a social content creation system. Poster presentation to be given at *iConference 2012*, Toronto, Canada: iSchools.

- Jörgensen, C., Stvilia, B., & Wu, S. (presented 2011, October). Assessing the quality of socially created metadata to image indexing. Poster presentation at *ASIS&T 2011 Annual Meeting*, New Orleans, LA: ASIS&T.

- Stvilia, B., Jörgensen, C., & Wu, S. (in press, 2012). Establishing the value of socially created metadata to image indexing. *Library & Information Science Research*.

- Stvilia, B., Jörgensen, C., & Wu, S. (presented 2011, January). Social metadata and image knowledge organization systems. Poster presentation at *ALISE 2011 Annual Conference*, San Diego, CA: ALISE.
- Stvilia, B., & Jörgensen, C. (presented 2011, January). *Assessing the reuse value of socially created metadata for image indexing*. Presentation at *ALISE 2011*, San Diego, CA: ALISE.
- Stvilia, B., Jörgensen, C. (2010). Member activities and quality of tags in a collection of historical photographs in Flickr. *JASIST, 61*(12), 2477–2489.
- Stvilia, B. & Jörgensen, C. (2010). Towards assessing relative value of user-generated tags to knowledge organization systems. In: *Proceedings of the 38th Annual CAIS/ACSI conference*. Montreal, Quebec, Canada.

Acknowledgements

The research described in this report is partially supported by an OCLC/ALISE Research Grant, 2010. The report reflects the findings, and conclusions of the authors, and do not necessarily reflect the views of the OCLC or the ALISE.

References

1. Chen, H., Yim, T., Fye, D., & Schatz, B. (1995). Automatic thesaurus generation for an electronic community system. *Journal the American Society for Information Science, 46*(3), 175–193.
2. Cover, T. & Thomas, J. (1991). *Elements of Information Theory*. New York, NY: Wiley.
3. Cunningham, S., & Jones, M. (2005). Autoethnography: A tool for practice and education. In B. Plimmer (Ed.), *Proceedings of the 6th ACM SIGCHI New Zealand Chapter's International Conference on Computer–Human Interaction: Making CHI Natural* (pp. 1–8). New York: ACM.
4. Greenberg, J. (2001). Quantitative categorical analysis of metadata elements in image applicable metadata schemas. *Journal of the American Society for Information Science and Technology, 52,* 917–914.
5. Jörgensen, C. (1995). *Image attributes: An investigation*. Unpublished Ph.D. thesis, Syracuse University, Syracuse, NY.

6. Jörgensen, C. (1998). Image attributes in describing tasks: An investigation. *Information Processing and Management*, *34*(2/3), 161–174.

7. Jörgensen, C. (1999). "Retrieving the unretrievable: Art, aesthetics, and emotion in image retrieval systems." *Proceedings SPIE (International Society for Optical Engineering) Vol. 3644*, p. 348–355, Human Vision and Electronic Imaging IV, Bernice E. Rogowitz, Thrasyvoulos N. Pappas; eds.

8. Jörgensen, C., Stvilia, B., & Wu, S. (accepted). Relationships among perceptions of term utility, category semantics, and term length and order in a social content creation system. Poster presentation to be given at *iConference 2012,* Toronto, Canada: iSchools.

9. Jörgensen, C., Stvilia, B., & Wu, S. (presented 2011, October). Assessing the quality of socially created metadata to image indexing. Poster presentation at *ASIS&T 2011 Annual Meeting*, New Orleans, LA: ASIS&T.

10. Lancaster, F. (2000). *Indexing and abstracting in theory and practice.* Champaign: University of Illinois, Graduate School of Library and Information Science.

11. Milne, D., & Witten, I. H. (2008). An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceedings of the first AAAI Workshop on Wikipedia and Artificial Intelligence (WIKIAI'08).*

12. Porter, M. (1980). An algorithm for suffix stripping. *Program, 14*, 130–137.

13. Rosch, E., Mervis, C., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*(3), 382–439.

14. Soergel, D. (1974). *Indexing languages and thesauri: Construction and maintenance*. Los Angeles: Wiley.

15. Stvilia, B. & Jörgensen, C. (2010). Towards assessing relative value of user-generated tags to knowledge organization systems. In: Proceedings of *the 38th Annual CAIS/ACSI conference*. Montreal, Quebec, Canada.

16. Stvilia, B., & Gasser, L. (2008). Value based metadata quality assessment. *Library and Information Science Research, 30*(1), 67–74.

17. Stvilia, B., Gasser, L., Twidale M. B., & Smith, L. C. (2007). A framework for information quality Assessment. *Journal of the American Society for Information Science and Technology, 58*(12), 1720–1733.

18. Stvilia, B., Gasser, L., Twidale, M., Shreeves, S., & Cole, T. (2004). Metadata quality for federated collections. In S. Chengulur-Smith, L. Raschid, J. Long, & C. Seko (Eds.), *Proceedings of the International Conference on Information Quality—ICIQ 2004* (pp. 111–125). Cambridge, MA: MITIQ.

19. Stvilia, B., Jörgensen, C. (2010). Member activities and quality of tags in a collection of historical photographs in Flickr. *JASIST, 61*(12), 2477–2489.

20. Stvilia, B., Jörgensen, C., & Wu, S. (in press, 2011). Establishing the value of socially created metadata to image indexing. *Library & Information Science Research*.

21. Stvilia, B., Jörgensen, C., & Wu, S. (in press, 2011). Establishing the value of socially created metadata to image indexing. *Library & Information Science Research*.

22. Stvilia, B., Jörgensen, C., & Wu, S. (presented 2011, January). Social metadata and image knowledge organization systems. Poster presentation at *ALISE 2011 Annual Conference*, San Diego, CA: ALISE.

23. Stvilia, B., Twidale, M., Smith, L. C., & Gasser, L. (2008). Information quality work organization in Wikipedia. *Journal of the American Society for Information Science and Technology, 59*(6), 983–1001.

24. Taylor, R. (1986). *Value-added processes in information systems*. Norwood, NJ: Ablex Publishing.