

Towards Assessing Relative Value of User-Generated Tags to Knowledge Organization Systems

Abstract: This research analyzes user-generated tags for Flickr images and whether and how these can contribute to controlled vocabularies, e.g. the *Thesaurus for Graphic Materials*. Uncontrolled tags provide diverse and rich description not possible with a controlled vocabulary. Integrating tags into an existing controlled vocabulary provides synergy to description for access.

Résumé : Cette recherche analyse les étiquettes d'images générées par les utilisateurs de Flickr et cherche à déterminer si et comment celles-ci peuvent contribuer à l'élaboration d'un vocabulaire contrôlé, p. ex. le *Thesaurus for Graphic Materials*. Les étiquettes libres offre une diversité et une richesse descriptive non atteignable avec un vocabulaire contrôlé. L'intégration de ces étiquettes au vocabulaire contrôlé existant permet une synergie dans la description aux fins d'accès.

1. Introduction

This exploratory study examines the nature of the added value of image tags from the Library of Congress' photostream on Flickr to a specialized controlled vocabulary used for image indexing by the Library of Congress - the *Thesaurus of Graphic Materials* (TGM). The study is guided by an Information Quality (IQ) Assessment Framework (Stvilia, 2006, Stvilia et al. 2007). Activities that are part of the knowledge organization process frequently experience IQ problems, not the least of which is the well-known "vocabulary problem" (Furnas, et al., 1987). Information Quality is complex for several reasons: 1) It is multidimensional, requiring different metrics for different dimensions and an ability to evaluate and compare the products of these metrics; 2) It is context-dependent, with quality judgments varying on differing genres of information and the nature of the specific task; and 3) It is dependent on the collection, with both the collection and the individual items being affected by their aggregate nature.

Traditionally, knowledge organization and representation systems (e.g., lists of terms, taxonomies, thesauri, ontologies) have been essential parts of the information organization and retrieval infrastructure in libraries and museums. Now, they have become increasingly important on the Web as well to support information retrieval, entity and concept identification, semantic annotation, and question answering. At the same time, knowledge organization systems can become quickly outdated and require continuous quality maintenance and alignment with the dynamic needs of end-users and use contexts in general; this can involve updating by expensive knowledge acquisition and engineering work. Inexpensive methods for this are sought, but at the same time these methods must conform to the requirements of IQ.

The current study explores the relative value that tags from the Library of Congress' photostream on Flickr could contribute to an existing knowledge organization system. This initial study is establishing improved methodologies for this process and exploring the results within an IQ framework. The study is also investigating the grammatical, functional, and cognitive nature of the tags.

2. Method

A controlled vocabulary is defined as an explicitly enumerated and controlled list of terms with unambiguous and non-redundant definitions (NISO, 2005). A controlled vocabulary can be as simple as a list of terms and as complex as a thesaurus containing concepts and entities along with related index terms, and hierarchical and associative relationships (Lancaster, 2000). A widely used general definition of quality is “fitness for use” (Juran, 1992); however, as noted, IQ is context dependent. IQ measurements can be grouped into three categories: intrinsic, relational, and reputational (Stvilia, 2006, Stvilia et al. 2007). *Intrinsic* quality measures the internal characteristics of a thesaurus itself in relation to some general reference standard in a given culture, such as WordNet (a general purpose comprehensive ontology of word senses; <http://wordnet.princeton.edu/>). *Relational* or contextual quality measures relationships between the thesaurus and some aspects of its usage context. Indeed, some of the quality dimensions (e.g., completeness) can be evaluated only in relation to the needs of a specific activity or action. Finally, *reputational* quality is the position of the thesaurus in a cultural or activity structure (e.g., a trust network), often determined by its origin and record of mediation. One of the functions of a controlled vocabulary is to translate end-user terms into the indexing vocabulary and language used by the system (NISO, 2005). To accomplish this task, the vocabulary has to be well aligned with the end-user vocabulary. Hence, one of the ways of assessing the relational quality of the TGM is by measuring its overlap with an end-user vocabulary represented by the tags assigned to photos of the Library of Congress’s collection on Flickr. Similarly, the relative value of a set of tags from Flickr to the TGM can be assessed based on the fraction of noun tags from the tag set that do not have a match in the TGM and could be added to it.

The data used by the study consisted of a set of 28,303 tags downloaded on September 13, 2009 from the Library of Congress’ Flickr photostream (http://www.flickr.com/photos/library_of_congress/) and associated with 7,192 photos. The tags then were matched to TGM terms (13,317 terms, including both subject and genre terms), and terms from the WordNet ontology (version 2.1; 207,016 terms). The TGM has been developed by the Library of Congress Prints and Photographs Division to support its image indexing needs following the principle of “library warrant,” with new terms and relationships added only if needed in the library’s operations. The TGM development team does not strive to construct comprehensive conceptual hierarchies of terms. This property makes the TGM an excellent baseline to assess the relative value of user generated terms and relations.

Since one of the goals of the study was to assess the value of the flickr tags as the source of new vocabulary terms and concepts, the study used a subsumption operator in matching the Flickr tags to the TGM. To be a match, a Flickr tag had to be a subset of the vocabulary’s term. To estimate an overlap between the Flickr tags and the vocabularies, however, one could also consider other matching functions such as the exact match between the terms, the Jaccard Similarity Coefficient, or the Euclidean Distance Measure. The Flickr tags were preprocessed before matching. In particular, multi-term concatenated tags in the Flickr set were recursively split into sets of terms based on the terms and inflections from the WordList (<http://wordlist.sourceforge.net/>). In addition, the set was cleaned of all URLs and the tags with less than 3 characters. This reduced the number of tags in the Flickr set to 20,946. Finally, both the Flickr tags and TGM terms were stemmed using the Porter Stemmer algorithm.

3. Findings

The analysis found 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). To understand the nature of the differences between the tags and terms from the TGM, the researchers manually analyzed a random sample of 300 tags with no match in the TGM. As was expected, most of the non-matches (64%) were associated with proper nouns and combined terms. The analysis also found regular nouns (8%) with no match in the TGM. Interestingly all of these nouns were present in WordNet.

Furthermore, only 6% of these nouns (e.g., “rods,” “gauge”) belonged to what is termed the “basic category” (Rosch et al, 1973). An important property of basic category terms is that they contain the most frequently occurring terms for concepts and can therefore be a valuable source for preferred terms for controlled vocabularies.

The analysis also found terms that were not found in the TGM, nor in WordNet nor Wikipedia. As expected, because of the historical nature of the Library’s collection, the terms represented objects and concepts of the past such as “Playograph,” “Grid Graph,” and “Aerocar.”

The researchers also examined general characteristics of the tags by manually analyzing a random sample of 300 tags from the complete Flickr set (including both matches and non-matches). The analysis found that more than half of these noun tags (56%) belonged to the basic category. Also, most of the tags (85%) could be categorized as subject metadata, and represented objects (30%), locations (26%), and people (17%). These results are largely similar to those of Jørgensen (1995).

4. Conclusions

The findings of the study suggest that tags from the Library of Congress’s photostream on Flickr can be a valuable source of concept terms, particularly of legacy/historical concepts. Future research may involve testing whether Flickr metadata can be a source of more contemporary, present-day concepts and terms as well by matching more general and time neutral samples of tags to the TGM and the LCSH.

5. References

Furnas, G., Landauer, T. K., Gomez, L. M., & Dumais, S. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11), 964–971.

Jørgensen, Corinne. “Image Attributes: An Investigation.” Ph.D. thesis, Syracuse University, 1995

Juran, J. (1992). *Juran on quality by design*. New York: The Free Press.

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

National Information Standards Organization (NISO). (2005). *Guidelines for the construction, format, and management of monolingual controlled vocabularies: An American national standard developed*. Bethesda, MD: NISO Press.

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

Stvilia, B. (2006). *Measuring information quality*. Thesis (Ph. D.) - University of Illinois at Urbana - Champaign, Urbana, 2006.

Stvilia, B., Gasser, L., Twidale M., B., Smith L. C. (2007). A framework for information quality Assessment. *JASIST*, 58(12), 1720-1733.