Creating Context for User-Generated Tags: An Exploratory Study

Nicole D. Alemanne
Florida State University
142 Collegiate Loop
Tallahassee, FL 32306-2100
nalemanne@fsu.edu

Besiki Stvilia
Florida State University
142 Collegiate Loop
Tallahassee, FL 32306-2100
+1 850 645 7366
bstvilia@fsu.edu

Corinne Jörgensen
Florida State University
142 Collegiate Loop
Tallahassee, FL 32306-2100
+1 850 644 8116
cjorgensen@fsu.edu

ABSTRACT

This exploratory study investigates methods for enhancing Flickr tags as image metadata through the creation of context. Community generated tags from a sample of images in the Library of Congress's (LOC) Flickr photostream were harvested and compared to metadata from related Wikipedia articles. In addition, a content analysis of comments in the LOC photostream was conducted. This informs an exploration of methods of combining user-generated tags with other resources to create richer, contextual metadata for images. In addition, the LOC and Wikipedia subject terms were compared to subject headings from the Thesaurus for Graphic Materials (TGM) to determine whether socially created metadata can be used to enhance a current knowledge organization tool by suggesting new concepts, terms, and relationships.

Categories and Subject Descriptors

H.5.3 [**Information Systems**]: Group and Organization Interfaces – *collaborative computing.*

General Terms

Measurement, Documentation, Languages.

Keywords

Social Tagging, Metadata, Image Description, Cultural Heritage.

1. INTRODUCTION

Researchers have proposed employing user-generated tags to enhance the metadata and descriptions of cultural heritage resources. Research continues to be conducted to determine if such key words might enhance the description and discovery of resources by adding the users' perspective [1]. Another fruitful area of investigation is the relationship of user-generated vocabulary to currently used metadata schemas and ontologies [2].

However, questions remain about the efficacy of user-generated key words for resource discovery as freely developed tags lack a number of the characteristics of key words developed through the use of controlled vocabularies and thesauri. Tagging has a shallow learning curve [3]. There are few usage rules, and experimentation is easy [4]. However, inconsistent vocabulary usage can create problems for resource discovery [5, 6]. In addition, the creator's definition of a tag, the context for the tag, and disambiguation are not available.

In January, 2008, the Library of Congress (LOC) launched a pilot project in which it made two collections of approximately 3,000 historical photographs available on the photo sharing website Flickr and invited the public to interact with the collections through tagging and description. Many of the images included in the pilot lacked in-depth caption information. Flickr offers its users the ability to append tags, comments, and notes to photos in the collections. By the end of October 2008, more than ten million views had been recorded and 67,176 tags had been added by 2,518 unique Flickr accounts [7]. LOC has continued to add collections to the photostream, and there are over 7,500 images in the collections as of mid-November 2009.

2. SCOPE AND PURPOSE

Because user-generated tags do hold promise for description and discovery, research to determine methods for creating context and disambiguation is an essential component in the broader tag-related research efforts. This project investigates a small sample of images that have user-generated tags, and works with outside resources to attempt to establish methods for creating context and disambiguation.

3. RESEARCH DESIGN

This research was conducted through content analysis, "a method of transforming the symbolic content of a document . . . from a qualitative, unsystematic form into a quantitative, systematic form" [8]. The content is systematized through coding, a process through which the elements are placed in a limited number of categories [8]. In particular, this is a conceptual analysis, in which concepts are quantified, categorized, and [9]. Explicit terms were identified and categorized using manifest coding [8, 9].

The project employs several sources of the data: the LOC photostream in Flickr (http://www.flickr.com/photos/library_of_congress/), Wikipedia (http://en.wikipedia.org/wiki/Main_Page), and the Thesaurus for Graphic Materials (TGM) (http://www.loc.gov/rr/print/tgm1/). A purposive sample of ten LOC images was selected, and the researcher identified Wikipedia entries covering the main subject of these images. Images from seven of the nine collections were included (table 1).

Table 1. Distribution of images by LOC Flickr collection

Collection	# Sample Images
News in the 1910s	3
1930s-40s in Color	2
Abraham Lincoln (1809-1865)	1
Baseball Americana	1
Photochrom Travel Views	1
Women Striving Forward, 1910s-1940s	1
World War I Panoramas	1
FSA/OWI Favorites	0
Illustrated Newspaper Supplements	0

LOC tags were harvested for each image, as were keywords from the Wikipedia entries. The LOC tags and Wikipedia key words were compared to determine the incidence of similarity and difference, to determine the efficacy of combining user-generated tags and Wikipedia-generated key words in creating metadata for images, and to investigate if Wikipedia entries might be used to create context and disambiguation for user-generated key words. In addition, a content analysis of comments in the LOC Flickr photostream was employed to explore the idea of comments as a process of collective disambiguation. Finally, the LOC/Wikipedia subject terms were compared to TGM subject headings to determine whether a current controlled vocabulary might accommodate user-supplied metadata.

4. DATA COLLECTION

Two types of data were collected and coded for the project: user-generated tags were harvested from the ten Flickr LOC images, and subject terms were harvested from the ten connected Wikipedia entries. All tags from the Flickr images were retained and used, with the exception of the "Library of Congress" tag attached to each image (this was considered an administrative tag). Wikipedia terms were harvested from the body and from the information box for each entry; all unique terms were retained. In addition, the comments sections of the LOC Flickr images were downloaded for analysis.

The harvested terms were coded by the researcher in order to investigate the incidence of similarity and difference between LOC Flickr tags and Wikipedia terms and the question of whether Wikipedia might entries be used to create context and disambiguation for user-generated tags. Two sets of codes were used—one for the analysis of similarities and differences in the Flickr and Wikipedia terms, and the other to create categories through which to analyze whether Wikipedia terms might be used to create context and disambiguation for user-generated tags.

4.1 Similarities and Differences

To determine the incidence of similarity and difference between LOC Flickr tags and Wikipedia terms, the researcher developed a coding scheme to differentiate between like and different Flickr and Wikipedia terms. Three categories were used for the similarities/differences coding:

- Flickr tags: Terms that are unique to the Flickr user-generated tags—these terms only appear in the Flickr tags, and not in the Wikipedia entries.
- Wikipedia terms: The complementary category to the first—terms that appear in the Wikipedia entries but were not used by Flickr taggers.
- Similar: Terms that appear in both the Flickr tag lists and the Wikipedia links list. Due to the lack of controlled vocabulary, terms that appeared to be similar were included in this category. For example, the terms 'America', 'United States', and 'USA' were considered to be similar terms.

For this process, each image was coded individually. For each image, the researcher started with the first Flickr tag, and compared it to the list of Wikipedia terms to determine the coding, with all Flickr-only terms coded in the Flickr list and similar terms coded in both lists. After the Flickr tags were fully coded, the Wikipedia links that had not been coded as 'similar' were coded as Wikipedia only terms.

4.2 Term Categories

To investigate whether Wikipedia entries might be used to create context and disambiguation for user-generated tags, the researcher developed a set of codes to categorize the terms. For this process, each image was coded individually. The categories were developed to create facets that represent the possible range of ways that users might describe the subject of images.

The terms were coded into four categories:

- Location: Any term that represents a georeferenceable location or a URL. Georeferenceable locations are those that can be established in terms of map projections or coordinate systems [10].
- Name: Any term that is a proper name but is not georeferenceable.
- Time: Any term that represents an individual point in time or a range of dates or times.
- Description: Any term that does not fit into the above categories.

4.3 Other Tasks

Two other tasks were completed to increase contextual understanding of the results. The LOC Flickr photostream includes a comments section that is used to greater and lesser degrees across images. These comments were analyzed to explore the idea of comments as a process of collective disambiguation. The interaction of the commenters was of specific interest for this task. In addition, LOC/Wikipedia subject terms were compared to TGM subject headings to determine whether a current controlled might accommodate user-supplied metadata.

5. REFERENCES

[1] Trant, J. 2009. Studying social tagging and folksonomy: A review and framework. Journal of Digital Information 10, 1 (2009). http://journals.tdl.org/jodi/issue/view/65.

[2] Stvilia, B. and Jörgensen, C. 2009. User-generated collection-level metadata in an online photo-sharing system. Library & Information Science Research 31, 1 (Jan. 2009), 54-65.

[3] Marchetti, A., Tesconi, M., Ronzano, F., Rosella, M., & Minutoli, S. 2007. SemKey: a semantic collaborative tagging system. In Proceedings of the 16th International Worldwide Web Conference (Banff, Alberta, Canada, May 08-12, 2007). http://www2007.org/workshops/paper_45.pdf.

[4] Beckett, D. 2006. Semantics through the tag. In Proceedings of XTech 2006: Building Web 2.0 (Amsterdam, The Netherlands, May 16-19, 2006). http://xtech06.usefulinc.com/schedule/paper/135.

[5] Golder, S. A., & Huberman, B. A. 2006. Usage patterns of collaborative tagging systems. Journal of Information Science 32, 2 (Apr. 2006), 198-208.

[6] Guy, M. and Tonkin, E. 2006. Folksonomies: tidying up tags? D-Lib Magazine 12, 1 (Jan. 2006). http://www.dlib.org/dlib/january06/guy/01guy.html.

[7] Library of Congress. 2008. For the common good: the Library of Congress Flickr pilot project. Library of Congress. http://www.loc.gov/rr/print/flickr_report_final_summary.pdf.

[8] Monette, D. R., Sullivan, T. J., & DeJong, C. R. 2008. Applied social research: a tool for the human services (7th ed.). Brooks/Cole.

[9] Colorado State University. 2009. Writing guides: content Analysis. Colorado State University. http://writing.colostate.edu/guides/research/content/index.cfm.

[10] Hill, L. 2006. Georeferencing. MIT Press.