This is a preprint of an article published in Library & Information Science Research: Stvilia, B., Jörgensen, C., & Wu, S. (2012). Establishing the value of socially created metadata to image indexing. *Library & Information Science Research*, 34(2), 99-109.

Establishing the Value of Socially Created Metadata to Image Indexing

Besiki Stvilia[1], Corinne Jörgensen, and Shuheng Wu
School of Library and Information Studies
College of Communication and Information, Florida State University
Tallahassee, FL 32306-2100, USA
{bstvilia, cjorgensen, sw09f}@fsu.edu

Abstract

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. At the same time, adding new metadata to existing metadata objects may not always add value to those objects. This research examines the potential added value of using user-contributed ("social") terms from Flickr and the English Wikipedia in image indexing compared with using two expert-created controlled vocabularies—the Thesaurus for Graphic Materials and the Library of Congress Subject Headings. Our experiments confirmed that the social terms did provide added value relative to terms from the controlled vocabularies. The median rating for the usefulness of social terms was significantly higher than the baseline rating but was lower than the ratings for the terms from the Thesaurus for Graphic Materials and the Library of Congress Subject Headings. Furthermore, complementing the controlled vocabulary terms with social terms more than doubled the average coverage of participants' terms for a photograph. The study also investigated the relationships between user demographics and users' perceptions of the value of terms, as well as the relationships between user demographics and indexing quality, as measured by the number of terms participants assigned to a photograph. It was found that the participants with more tagging and indexing experience assigned a greater number of tags than did the other participants.

1. Introduction

Image indexing is a complex socio-cognitive process that involves processing sensory input, classifying, abstracting, and mapping sensory data into concepts and entities often expressed through socially defined and culturally justified linguistic labels and identifiers (Heidorn, 1999). Unlike the content of text documents, raw image content is not linguistic. This makes image indexing more dependent on both human indexers and the use of image knowledge organization and representation systems (KOS) to provide content models that are understandable and searchable by humans (Heidorn, 1999; Rasmussen, 1997). The large-scale participation of the public in open social content creation

[1] Corresponding author

communities and systems (e.g., Wikipedia, Flickr) makes it possible to accomplish tasks that still require significant human involvement at a relatively low cost, including describing millions of photographs and generating new KOS (e.g., DBpedia, http://dbpedia.org). Questions remain, however, about the quality and reuse value of socially-generated metadata in general and for image indexing in particular.

Knowledge organization and representation systems (e.g., lists of terms, taxonomies, thesauri, ontologies) traditionally have been essential parts of the information organization and retrieval infrastructure in libraries and museums, and they have now become increasingly important on the Web to support entity and concept identification, semantic annotation, information retrieval, and question answering (e.g., Perez, 2009). Not surprisingly, considerable research has been conducted on controlled vocabulary and ontology construction, including research identifying quality index terms and research on automatic concept and relationship identification (e.g., Chen, Yim, Fye, & Schatz, 1995; Lancaster, 2000; National Information Standards Organization [NISO], 2005; Soergel, 1974). Nevertheless, the construction of quality KOS involves expensive knowledge engineering work. Furthermore, quality is being recognized as contextual and dynamic (Jörgensen, 1995b; Strong, Lee, & Wang, 1997; Stvilia, Gasser, Twidale, & Smith, 2007). With changes in domain culture, activity systems, knowledge and technology, and user expectations, the quality of these KOS systems can quickly become outdated and require regular intensive maintenance and upkeep. There is a need to identify sources of, and define methods and mechanisms for, inexpensive dynamic acquisition, evaluation, and integration of new knowledge into traditional KOS.

Greenberg (2002) suggested that collaboration between intellectual content creators and information organization experts may lead to a higher level of quality in metadata creation. Creators have intimate knowledge of their creations, whereas indexers and catalogers can use their knowledge of metadata schemas and classification systems to assist the creators . There has been an increase in efforts to acquire metadata for existing image and photo collections by deploying these collections in existing social tagging communities (e.g., Springer et al., 2008) or by bundling image collections with social tagging systems and mechanisms (e.g., gaming interfaces) to jump-start new communities and motivate users to contribute metadata (e.g., Trant, 2008; von Ahn & Dabbish, 2004). In addition, with the establishment and increase in popularity of social content creation and tagging systems such as Wikipedia and Flickr, researchers can gain access to large sets of nonexpert-created metadata (e.g., tags, classification strings). Previous research suggests that access to these data sets may help extend existing controlled vocabularies and ontologies, or help generate new ones (Agirre, Ansa, Hovy, & Martinez, 2000; Jörgensen, Stvilia, & Jörgensen, 2008; Matusiak, 2006; Medelyan & Milne, 2008; Stvilia and Jörgensen, 2010; Wetterstrom, 2008).

Metadata in social tagging systems are generated by different types of users in different contexts and for different purposes (Cunningham & Masoodian, 2006; Stvilia & Jörgensen, 2009). The terms that users select when they search for images may differ from the terms they use when describing and organizing their photos (Chung & Yoon, 2008). Furthermore, users may perceive the quality of suggested controlled vocabulary terms differently, which may affect the use of those terms (Fidel, 1991). The users' level of domain expertise and familiarity with the system have been found to influence their evaluation of the quality and usefulness of suggested vocabulary terms (Nelson, Johnston, & Humphreys, 2001; Shiri &

Revie, 2006; Vakkari, Pennanen, & Serola, 2003). Prior studies of Dublin Core metadata schema use have also revealed that quality is contextual. Local metadata providers may lack economic incentives to encode the metadata that are considered shared knowledge in their local communities. Metadata could be considered of high quality in a local context, but if they are moved and aggregated in a different context, their quality could be evaluated differently (Stvilia, Gasser, Twidale, Shreeves, & Cole, 2004). Hence, before extending an expert-constructed vocabulary with socially created metadata, it is important to assess the quality of the metadata and, more importantly, their added value to users of the vocabulary. Adding new metadata may not necessarily translate into a value increase or cost reduction for the activity (Stvilia & Gasser, 2008).

2. Problem Statement

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 use 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 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 the Thesaurus for Graphic Materials (TGM) and the Library of Congress Subject Headings (LCSH) by providing additional access points. 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.

3. Related Research

The unique characteristics of visual media have led to at least two different approaches to indexing images, based on the level of indexing and the technologies used: content based and concept based (Rasmussen, 1997). Content-based indexing, a computer method for algorithmically parsing images, can be done automatically and can be an inexpensive method for assigning certain kinds of metadata to large numbers of images. However, content-based indexing and retrieval systems can successfully "recognize" only pixel-level attributes (e.g. color, shape, texture). They are not very successful at translating low-level pixel-derived data into higher-level content metadata that users typically understand, describe, and use to search images. In concept-based indexing, which is primarily a manual approach, human indexers assign terms to images representing higher level concepts and semantic relationships that are unable to be parsed in the machine environment. However, concept-based indexing,

as a manual process, is expensive and is still both "information lossy" (i.e. assigned key words may not capture the full semantics of the image content) and context sensitive, because different indexers may use different sets of key words to describe the same semantic meaning, leading to the well-known "vocabulary problem" (Furnas, Landauer, Gomez, & Dumais, 1987; Smeulders et al., 2000).

Addressing the vocabulary problem—the problem of people using different words when describing or searching for the same concepts and entities, or alternatively, using the same words for different concepts and entities—has been one of the oldest problems in knowledge organization and continues today to be a very active area of research and practice (e.g., Buckland, 1999; Furnas et al., 1987; Klavans et al., 2009; Svenonius, 1986; Tan, Kan, & Lee, 2006). Controlled vocabularies (i.e., lists of terms, thesauri) are used to address the vocabulary problem between the user and the system by translating user terms into the indexing language used by the system. Museums and libraries have traditionally practiced concept-based image indexing and have invested heavily in sophisticated controlled vocabularies and ontologies, such as the TGM, the LCSH, and the Art and Architecture Thesaurus. To accomplish effective translation between the user and the system languages, however, a controlled vocabulary needs to align well with the information needs and language of the user (Soergel, 1974).

A controlled vocabulary is defined as an explicitly enumerated and controlled list of terms with unambiguous and nonredundant 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 their hierarchical and associative relationships (Lancaster, 2000). A controlled vocabulary can serve multiple purposes: translation (addressing the vocabulary problem by translating end-user terms into the indexing vocabulary and language used by the system); consistency (promoting uniformity in the format and in the assignment of terms); an indicator of relationships (indicating semantic relationships among terms); labeling and browsing (providing consistent and clear hierarchies in a navigation system to help users locate the desired content objects), and retrieval (NISO, 2005). These objectives are achieved by the careful design and maintenance of controlled vocabularies, which include eliminating ambiguity, controlling synonyms, establishing relationships, and testing and validating terms to align these with user, organizational, and community vocabularies and activity needs (NISO, 2005; Soergel, 1974). The quality of a controlled vocabulary may change with (1) changes in the synonym–homonym structure (e.g., new synonyms may appear); (2) changes in the classificatory structure or hierarchy (e.g., new concepts may be added); or (3) changes in the indexing language (e.g., a new description may be introduced; changes may occur in the description or use of a descriptor or in descriptor relationships; Soergel, 1974, pp. 457–458). User needs for terms and relationships can be identified through search log analysis or laboratory experiments involving end-user searches (Soergel, 1974, p. 458; Stvilia, 2007). Alternatively, user terms and relationships can be harvested from end-user or community-created documents, data, and metadata (e.g., Chen et al., 1995; Sarjant et al., 2009).

Considerable prior research has been conducted on index and query term evaluation (Blair, 1996; Brooks, 1993; Cleverdon, 1997; Greenberg, 2001a; Hersh, Pentecost, & Hickam, 1996; Wacholder & Liu, 2006), and there is consensus that quality is contextual (Strong et al., 1997; Stvilia et al., 2007). A known entity or object identification task may require the use of highly specific metadata (e.g.,

identifiers), whereas for a general relevance-based ("aboutness") or other attribute-based (e.g., format) selection task, the use of a low-specificity term may suffice (Stvilia et al., 2004). Similarly, different indexing techniques may perform differently for different tasks and for different media. Automated term frequency statistics-based indexing may perform almost as well as human indexing for the textual document relevance (i.e., aboutness) identification task (Salton, 1986; Sparck Jones, 2005). For question answering, word sense disambiguation, and knowledge acquisition, however, more sophisticated natural language processing-based techniques and context-specific modeling and processing may be needed (e.g., at the paragraph or sentence level; see Agirre et al., 2000; Mani, Samuel, Concepcion, & Vogel, 2004; Urbain, Goharian, & Frieder, 2007). A large body of research has compared the effectiveness of indexing based on controlled vocabularies with full-text, algorithmic approaches to indexing for textual documents (e.g., Blair, 1986), and a review of this line of research can be found in Anderson and Perez-Carballo (2001a, 2001b).

Researchers have emphasized the importance of indexing documents based on end-user information needs and search queries (e.g., Soergel, 1974). It has been shown that different groups of end users can have different information need structures for and perspectives on the same document (Hjørland, 2008; Lancaster, 2000, p. 9). A significant amount of research has focused on identifying and categorizing nonexpert index terms for images (Jörgensen 1995a, 1995b, 1996, 1998; Jörgensen & Jörgensen, 2002; Lin et al., 2006; Mathes, 2004). Jörgensen (1995b) used different types of descriptive tasks with a wide range of participants and derived 10 broad classes related to descriptive image content from approximately 14,000 terms. Other studies have found that users group or categorize images by broad concepts, whereas they describe images by using more specific concepts and terms (Jörgensen, 1995a, Rorissa & Iyer, 2008). In addition, it has been suggested that the granularity of terms used to describe images may not be at the same level as the granularity of terms used in a search. Chung and Yoon (2008) sampled and classified Flickr tags and user image queries obtained from Excite 2001 query logs into the Shatford (1986) categories (abstract, generic, and specific). A comparison of the tag and query categories revealed that whereas most of the tags were of the generic type (63%), most of the queries were specific (51%).

In addition, research suggests large sets of nonexpert-created metadata from social content creation systems such Wikipedia and Flickr can be used to extend existing controlled vocabularies, or help generate new ones. For instance, Jörgensen, Stvilia, & Jörgensen (2008) matched a large sample of Flickr tags to a nonspecialist controlled vocabulary and found that Flickr terms could be helpful in completing the term list for the vocabulary. Wetterstrom (2008) compared user-assigned tags for books in a New Zealand library with the LCSH used by catalogers for the same books. The author found that 75% of tags did not have a match with the subject headings and that the majority of mismatches were popular terms. Matusiak (2006), who compared Flickr terms with index terms used for similar images in a library collection, found that Flickr terms could represent user vocabulary in multiple languages. In a study related to this one, Stvilia and Jörgensen (2010) compared the distribution of cognitive categories (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976) and the quality of tags in the Library of Congress (LoC) photostream on Flickr with the distributions of term characteristics in the TGM and LCSH. They found that 37% of the original tag set and 15.3% of the preprocessed set (after the removal of tags with fewer

than three characters and URLs) were invalid or misspelled terms. Valid nouns, named entity terms, and complex terms constituted approximately 77% of the preprocessed set. More than half of those terms were not found in the TGM and LCSH, one fourth of which were regular nouns and noun phrases, suggesting that the terms could be complementary to more traditional methods of indexing using controlled vocabularies.

Other studies have examined the possibility of automated thesaurus generation from tag co-occurrence data or folksonomies (Heyman & Garcia-Molina, 2006; Mika, 2007). A folksonomy is defined as an informal grouping of related tags. A thorough review of folksonomy-related research can be found in Trant (2009). In addition, attempts have been made to combine evidence from multiple sources to extend controlled vocabularies. Marchetti, Tesconi, Ronzano, Rosella, and Minutoli (2007) proposed architecture for a semantic tagging system that uses WordNet and Wikipedia to disambiguate tags. Klavans et al. (2009) developed a tool kit to support the image indexing workflow. The tool kit includes modules for semiautomatic metadata extraction from the textual description of the image and entity disambiguation and resolution using external thesauri and authority files.

The literature offers ample evidence that tags differ from controlled vocabulary terms. In addition, the preceding studies provide valuable insight into approaches toward and algorithms for automated thesaurus generation. Additional research, however, is needed to investigate the degree of reuse and the potential added value of terms and relationships found in socially created metadata compared with expert-constructed controlled vocabularies in general, and with controlled vocabularies used for image indexing in particular.

4. Research Questions

This study examined the efficacy of using Flickr and Wikipedia metadata to generate thesaurus enhancements for an established image thesaurus, the TGM, and the subject headings produced by the LoC, the LCSH. In particular, the study evaluated the following research questions:

- Does providing social terms from Flickr and the English Wikipedia have the potential to add value to the TGM and the LCSH by providing additional access points? What is the nature of this added value?

The study uses the phrase ‘social terms’ as a collective designator to refer to both tags directly assigned to photographs by Flickr members, as well as related Flickr tags and English Wikipedia terms selected algorithmically. To identify value, two aspects of the potential added value of social terms were evaluated. First, the study examined the *subjective* value of the social terms for an image description task, as perceived by participants in an experimental setting. Following that, the study examined the *objective* value of the terms, measured using an objective metric - a normalized fraction of participant-selected terms covered by those terms.

The study also asked:

- Are there relationships among user demographic characteristics and users' valuation of the usefulness of index terms?

- Are there relationships between participant characteristics and the number of terms participants used to describe and search for photographs?

5. Methodology

Within a broad conceptual approach, the study comprised several interconnected research activities: data collection, data analysis, model development, and evaluation. To carry out these activities, a mixed methodology was used (Bailey, 1994). 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). Taylor (1986) in his value-added model of information systems gives four interpretations of the concept of value which can be extended to library KOSs: (1) creation of wealth through production and distribution; (2) increase of usefulness; (3) exchange-value and (4) impact of information on the user. The amount of wealth created, in general, is determined by the amount of the resources spent, that is, by the cost of production. A marketplace establishes the exchange-value of a product. Determining the amount of increase in usefulness or the impact on the user requires, however, the knowledge of the immediate user context. In addition, the usefulness of a KOS can be evaluated as a function of activity success or failure (Stvilia & Gasser, 2008). One of the main purposes of controlled vocabularies is to translate user terms into the indexing language used by the system. To accomplish this successfully, however, a controlled vocabulary needs to align well with the information needs and language of the user (Soergel, 1974). If no mapping or match exists between the terms selected by the user for an information object and the terms used by the system index, then the system does not retrieve the object and the search fails.

The study adapted the experimental designs used by Jörgensen (1998) and Chen et al. (1995). Prior studies of end-user image-searching behavior, thesaurus use, and query expansion have shown that the search vocabulary of the user and the user's perceptions of thesaurus and metadata usefulness may vary with the type of task and user characteristics, such as domain knowledge and familiarity with the system (e.g., Choi, 2008; Cunningham & Masoodian, 2006; Efthimiadis, 2000; Greenberg, 2001b). Although one can expect that describing and understanding some of the Flickr photographs and tags might require knowledge of the historical or cultural context of the photograph, the study did not target any specialized knowledge domain and therefore did not include domain knowledge as a moderating variable in the research design. Similarly, the research design did not include system familiarity as a variable in the design because the participants self-selected for the experiments had no prior experience working with the experimental system. The study did, however, include task type, participant age, level of education, sex, and language as independent variables.

5.1. Experiments

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 first experiment (the description task) involved a group of 35 students (both graduates and undergraduates) and staff members 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 deployed on a local Web server and 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. In particular 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 (i.e., with the least amount of browsing and query revising). Twenty-five participants from the original group of 35 completed the search task.

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 predefined query types were based on the types of information needs for the photos, as identified by Cunningham and Masoodian (2006): specific needs (referring to a specific event, person, or activity), general nameable needs, general abstract needs (involving abstract concepts), and subjective needs (satisfying emotional responses). Findings from the analysis of exit interviews and diaries are reported elsewhere (Stvilia et al., in preparation).

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. In particular, the study matched the controlled

vocabulary and social terms pre-assigned to a photograph with the terms participants used in the experiments to describe and search for the same photograph. The degree of coverage of participant terms by pre-assigned terms was calculated as the Dice coefficient (van Rijsbergen, 1979). Both the pre-assigned and participant term sets were preprocessed before matching. Terms were stripped of plural suffixes as well as non-letter characters. Conjunctions and determiners (e.g., and, that, then) and terms shorter than three characters were also removed from the sets.

5.2. Sampled Photographs

The sample used in the experiments consisted of 10 photographs selected from a set of 7,192 photographs from the LoC Flickr photostream, downloaded on September 13, 2009. The LoC and other institutions of cultural heritage use Flickr to increase the use of their rich visual collections and the public's awareness of them, and this engages the Flickr community in contributing to, and potentially improving the quality of the metadata for these photos. A detailed description of the LoC photostream on Flickr, including analysis of the tags used in the photostream and member conversations captured in Flickr, was published in an earlier related work (Stvilia & Jörgensen, 2010). The sample of photographs used in the current study was a convenience sample. Because the sample was intended for use in controlled experiments, the following criteria were used to select the photographs and determine the size of the sample. First, the sample had to include photographs on different topics. Second, the size of the sample had to be moderate enough so that the term-rating part of the experiment could be completed by participants within one hour.

5.3. Pre-assigned Index Terms

To preindex a photograph for the description experiment, the study used 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 (see Figure 1 and Table 1).

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. In addition, the articles are interconnected, with networks of inner links and subject category trees. Comprehensiveness and currency of coverage have been the main virtues of the English Wikipedia, and recently there has been growing interest in mining Wikipedia for automated knowledge acquisition (e.g., DBpedia). Researchers have extended existing thesauri and identified new concepts, terms, and relationships by analyzing Wikipedia article content, link structures,

and category trees (e.g., Medelyan & Milne, 2008). Other studies have used Wikipedia content to disambiguate tags used for semantic annotation (e.g., Marchetti et al., 2007).

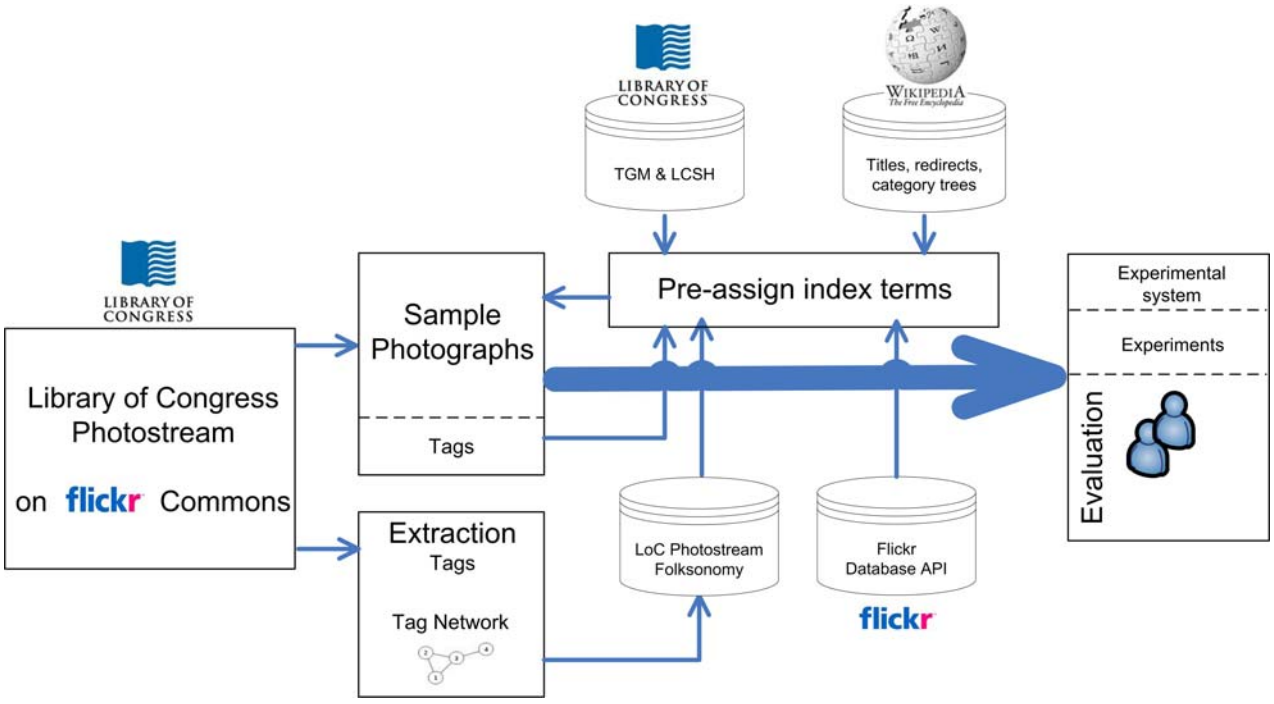

Fig. 1. Study design overview.
TGM = Thesaurus for Graphic Materials; LCSH = Library of Congress Subject Headings; LoC = Library of Congress; API = application programming interface.

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. 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).

The related social terms included in the preassigned term sets were obtained algorithmically by using a Java code developed by one of the researchers. To identify the set of related terms for a photograph from the LoC folksonomy, the folksonomy was searched by using each term from the first

two sets (the TGM and LCSH terms and the Flickr tags for the photographs). This procedure selected the terms with pairwise mutual information with a query term equal to or greater than the median mutual information value of the folksonomy (i.e., six). The results of the queries were aggregated into one set without removing duplicate entries. To determine the most relevant LoC folksonomy terms for the context of a photograph, the code used a term frequency-based "voting" approach suggested in the literature (Sigurbjörnsson & van Zwol, 2008). In particular, the code calculated term frequencies in the aggregate set and selected up to 30 of the most frequently occurring terms. The selected LoC folksonomy terms were then merged with the set of index terms for the photographs.

In the next step, the code identified related terms from the Flickr database by querying it with each term from the set of index terms for the photographs by using the "getRelated" procedure of the Flickr application programming interface. The code used the same procedure as with the LoC folksonomy terms to aggregate and promote terms returned by the queries into the index term set for the photograph.

In the last step of the algorithm, the code searched the English Wikipedia database to find a matching article as well as redirect terms, equivalent terms, and upper level category terms for each term in the set of index terms for the photographs. To be a match the article's title had to match the index term. Because redirects are often used by Wikipedia for spelling corrections (i.e., to connect commonly occurring misspelled forms of a term with the correct form), the number of redirects for each article can be significant and may include invalid terms. To mitigate this problem, the code selected only the redirects that were valid terms (i.e., not misspelled) and two upper level category terms. The resulting set of terms was then merged with the set of index terms for the photographs.

After the automated procedure had completed the identification of related social terms, the same two researchers independently examined the sets and removed irrelevant and invalid entries. In a separate indexing session, the researchers discussed and resolved indexing differences and developed final sets of index terms for each photograph (see Table 1).

This is a preprint of an article published in Library & Information Science Research: Stvilia, B., Jörgensen, C., & Wu, S. (2012). Establishing the value of socially created metadata to image indexing. *Library & Information Science Research*, 34(2), 99-109.

Table 1. The composition of preassigned term sets.

Photo ID	Index term						
	TGM	LCSH	Photo tags	LoC folksonomy	Flickr database	Wikipedia	Total
1	6	4	6	1	6	10	33
2	4	3	3	0	5	1	16
3	11	3	50	0	6	28	98
4	15	8	25	0	3	11	62
5	6	2	2	0	6	2	18
6	5	0	7	0	4	4	20
7	10	3	10	0	9	6	38
8	4	0	5	0	0	0	9
9	6	3	11	0	7	4	31
10	8	1	6	1	6	1	23
Mean	7.5	2.7	12.5	0.2	5.2	6.7	34.8
Median	6	3	6.5	0	6	4	27

TGM = Thesaurus for Graphic Materials; LCSH = Library of Congress Subject Headings; LoC folksonomy = folksonomy of the Library of Congress photostream on the Flickr. Note: duplicates and word variations were removed.

5.4. Experimental System

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.

6. Findings

The demographics of the participants in the three experiments were as follows. Almost half of the participants were undergraduate students (46%). The rest of the sample was distributed among doctoral students (28%), master's students (20%), and master's degree holders (6%). The age of the participants ranged from 19 to 59 years old. Forty-three percent of participants in the sample were female and 57% were male. Thirty-four percent indicated that they had some tagging experience and 11% identified themselves as having an intermediate knowledge of indexing. Finally, 29% were non-native speakers of the English language. The median number of tags that participants assigned to the photographs in the description task was five, whereas the median number of terms used in the search task queries was four.

The study investigated whether adding social terms from Flickr and the English Wikipedia to the TGM and LCSH added value to these controlled vocabularies. Two aspects of the value of social terms were evaluated. First, the study examined the subjective value of the terms in the image description task as perceived by participants. Following that, the study examined the objective value of the terms, measured as a normalized fraction of participant terms covered by those terms.

The study assessed the perceived value of the social terms by comparing the ratings of the usefulness of the social terms with a baseline that was set to a neutral rating (i.e., 3). The mean and median ratings for the usefulness of the social terms were 3.4 and 4.0, respectively. This median was significantly higher than the baseline, with $p = 0.0001$ for the one-sample median test ($z = 27.8$).

In addition, the Mann-Whitney test showed that the distribution of the ratings of social terms was significantly different from the distributions of the ratings of TGM and LCSH terms with the controlled vocabulary terms ranked higher than the social terms ($p = 0.0001$; $z = 13.2$) The mean and median values for the TGM and LCSH term ratings were 3.7 and 4.0, respectively.

The second facet of the added-value analysis measured the degree of additional coverage of participant terms provided by the social terms. The median percentage of added coverage of participant terms from the description task (i.e., tags) was 127%. The median percentage of the added coverage of query terms from the search task was also high, 108%. The added coverage was calculated as follows:

$$addedCoverage = \frac{\#of\_Matches\_of\_Participant\_Terms\_to\_Social\_Index\_Terms}{\#of\_Matches\_of\_Participant\_Terms\_to\_TGM\ \&\ LCSH\_Terms} \times 100$$

In addition to the above metric, the study used the Dice coefficient to evaluate set overlaps between the social and participant terms. Results of the analysis showed that the median set overlap of participant terms from the description task (i.e., tags) with the complete set of pre-assigned index terms was twice as high as the median overlap of participant terms with the TGM and LCSH terms only: 0.1 versus 0.05 (i.e., 10% vs. 5%). The median set overlaps of query terms used in the search task with the complete set of preassigned index terms and the controlled vocabulary terms were 0.19 vs. 0.12 (i.e., 19% vs. 12%). It is important to note that the median set overlap of query terms with the terms from the description task was even higher, 0.23 (i.e., 23%). The set overlaps were calculated as follows:

$$c = \frac{2|A \cap B|}{|A| + |B|},$$

where c stands for the Dice coefficient; A and B stand for the sets of terms and $|\ .\ |$ denotes the size of a term set (van Rijsbergen, 1979).

The study also examined whether participants' valuations of the pre-assigned terms were affected by their demographic characteristics. The study used an ordered logistic regression to regress term source and the demographic profiles of the participants into the term rating. In addition to term source, the regression model included education, age, sex, tagging experience, Flickr familiarity, indexing

experience, and language as independent variables. Because none of the participants was familiar with the Steve tagger software, familiarity with the Steve tagger was not included in the model. The definitions and codes of the variables are provided in Table 2. The regression analysis confirmed the findings of the Mann-Whitney test that the participants rated TGM and LCSH terms higher than social terms. In addition, the results of the regression analysis revealed that participant age and tagging experience were negatively related to term rating. The relationships of Flickr familiarity and indexing experience with term rating, however, were positive (see Table 2 and Figure 2). No significant relationship was observed between sex and term rating.

Table 2. Results of the regression analyses

		Coefficient (SE)		
Variable	Codes and definitions	A	B	C
Education	1 - Some college or college graduate 2 - Master's student 3 - Master's degree 4 - Ph.D. student 5 - Ph.D. degree	–0.03 (0.03)	–0.35 (0.25)	–0.17 (0.03)**
Age	Age of the participant	–0.004 (0.002)*	–0.15 (0.01)**	0.06 (0.001)**
Sex	0 - Female 1 - Male	–0.06 (0.04)	–0.87 (0.27)**	0.67 (0.05)**
Tagging experience	0 - No 1 - Yes	–0.44 (0.05)**	1.17 (0.29)**	1 (0.07)**
Flickr familiarity	0 - Never used 1 - Other 2 - Searching and browsing 3 - Added tags and comments 4 - Storing and sharing photographs 5 - Research and work	0.07 (0.02)**	–0.42 (0.10)**	0.11 (0.02)**
Indexing experience	0 - Novice 1 - Intermediate 2 - Expert	0.19 (0.06)**	8.38 (0.47)**	–0.61 (0.06)**
Native English speaker	0 - No 1 - Yes	0.55 (0.06)**	–1.14 (0.51)*	0.50 (0.05)**
Source	0 - TGM and LCSH 1 - Social	–0.48 (0.04)**	NA	NA

A = results of the ordered logistic regression regressing participant demographics and term source into term ratings (model fit likelihood ratio: $\chi^2 = 554.2$, $p < 0.0001$, number of observations = 12,180, pseudo $R^2 = 0.02$); B = results of the 0.75 quantile regression regressing participant demographics into the number of tags assigned to a photograph in the description task (number of observations = 350, pseudo $R^2 = 0.18$); C = results of the 0.75 quantile regression regressing participant demographics into the number of search terms used in the search task (number of observations = 250, pseudo $R^2 = 0.09$). TGM = Thesaurus for Graphic Materials; LCSH = Library of Congress Subject Headings; NA = not analyzed.
*$p < 0.05$. **$p < 0.005$.

The study also investigated the relationships between participant demographics and the number of tags assigned to the photographs in the description task to identify some of the characteristics of prolific indexers. The Wilks-Shapiro normality test showed that the number of tags was not normally distributed. Hence, the researchers used a 0.75 quantile regression analysis to regress the demographic variables into the number of tags. The analysis revealed that the variables age and Flickr familiarity were negatively

related to number of tags. Likewise, male and native-English-speaking participants assigned fewer tags. The relationships of indexing experience and tagging experience with the number of tags, however, were positive (see Table 2 and Figure 3).

Table 3. Correlation table of the participant demographic variables.

Variable	Education	Sex	Native speaker of English	Tagging experience	Flickr experience	Indexing experience
Age	0.70**	–0.24**	–0.46**		0.19**	–0.06**
Education		–0.24**	–0.69**	0.05**	0.31**	–0.14**
Sex			–0.04**	–0.35**	–0.32**	–0.23**
Native speaker of English				0.06**	0.04**	0.03**
Tagging experience					0.54**	0.31**
Flickr experience						0.07**

Spearman's rank correlation. Only statistically significant relationships are included.

**$p < 0.005$.

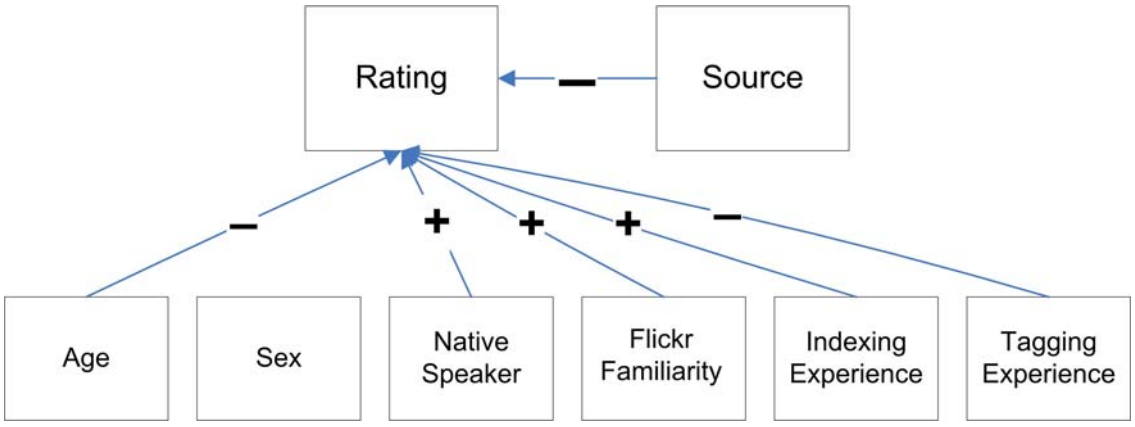

Figure 2. Relationships between participant demographic characteristics and term ratings from the description task.

Finally, the study examined the relationship between participant demographic characteristics and the number of query terms used in the search task (see Table 2 and Figure 3). The analysis found that tagging experience, Flickr familiarity, and age were related positively with the number of terms.

Nonnative speakers, female participants, and participants with no indexing experience used fewer terms in their queries. In addition, somewhat surprisingly, a negative relationship was observed between participant education and the number of query terms used.

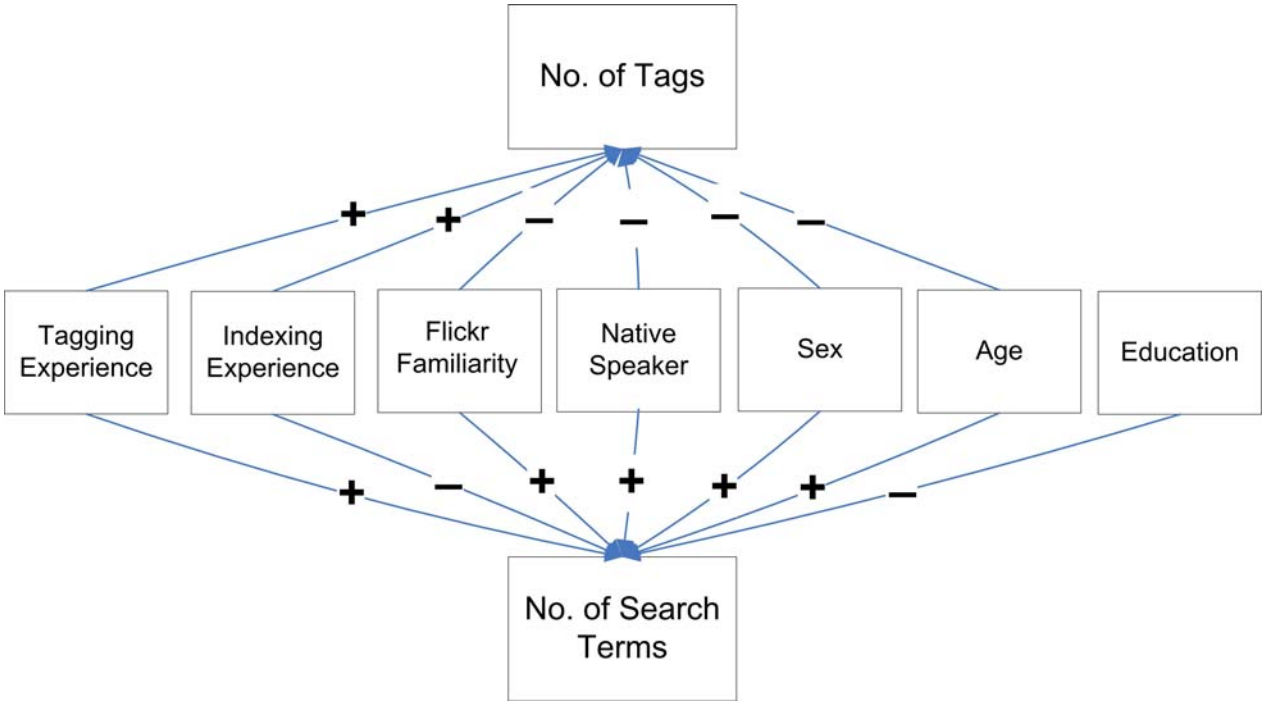

Figure 3. Relationships of participant demographic characteristics with the number of tags used in the description task and the number of terms used in the search queries.

7. Discussion

This study investigated the added value of social terms relative to the TGM and LCSH through 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 study 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. 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.

These findings confirm the suggestions in earlier studies, including our own, that socially created metadata can be useful for supplementing and extending expert-created controlled vocabularies (Jörgensen, Stvilia, & Jörgensen, 2008; Rorissa, 2010; Stvilia & Jörgensen, 2009, 2010; Yi & Chan, 2009; Yoon, 2009). This study is an extension of an earlier study in which the authors examined the intrinsic and relational quality of the LoC photostream folksonomy relative to the TGM and LCSH (Stvilia & Jorgensen, 2010). Results of that study showed that although 15.3% of terms in the preprocessed set (after the removal of tags with fewer than three characters and URLs) were invalid or misspelled, more than half of the preprocessed photostream tags were not found in the TGM and LCSH, and more than one fourth of those terms were regular nouns and noun phrases, suggesting that those terms could be complementary to more traditional methods of indexing using controlled vocabularies. The current study confirmed that suggestion with the results of the controlled user experiments.

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. These results also provide evidence supporting the view in the literature (e.g. Schwartz, 2008) that folksonomies and other sources of social terms have value in extending and enhancing expert-created KOS by providing additional descriptors and access points, rather than by substituting for the KOS.

In addition, this study found that query terms from the search experiments were best covered by the terms participants used in the description experiments. Although this finding had been expected, it still provides another indication of the potential value of social metadata. It is important to note that the same group of participants was involved in both types of experiments, even though there was a 2-week interval between the description and search experiments and the participants did not have access to the sets of terms they had used in the description experiment when completing the search experiment. In addition, after assigning tags to a photograph, participants were presented with pre-assigned terms for the same photograph and were asked to rate those terms. Hence, any possible effects of participants remembering tags on the finding would be mitigated by similar effects of participants remembering the pre-assigned terms. Nevertheless, it would be interesting to examine whether this finding would still hold if the experiments were replicated with larger and different groups of participants.

The study also explored the relationships between participant demographics and term ratings. Results of the analysis showed that older participants with tagging experience rated the terms lower than did the other participants. On the other hand, native English speakers with a greater familiarity with Flickr and with some indexing experience provided high ratings. It was interesting to observe the negative relationship between tagging experience and term rating versus the positive relationship between indexing experience and term rating, suggesting that the participants with greater tagging experience were less favorable toward the terms than were the participants with indexing experience. Some of this differences could be attributed to the sample's characteristic. In particular there was a positive correlation between education and tagging experience and the negative correlation between education and indexing experience (see Table 3). A higher number of the participants with advanced degrees had more tagging experience and no indexing experience than otherwise. In addition, the regression analysis showed that the relationships between demographic variables and term rating accounted for only 2% of term rating's variance. One may hypothesize that a considerable portion of the unexplained variance in term rating

could be related to term properties such as cognitive categories (Rosch et al, 1976) and importance to the content of the image. Future work may involve exploring possible relationships between the cognitive categories of terms and term ratings.

The analysis revealed a positive relationship between participant indexing and tagging experience and the number of terms used in the description task. In addition, older, male native speakers and participants with greater Flickr familiarity assigned fewer terms. The amount of variance of the number of index terms explained by these relationships was small (18%) pointing to the presence of other factors that may contribute to this variance. One can theorize that a substantial portion of the unaccounted variance could be caused by differences in image content complexity (Hastings, 1999). A fruitful future research direction could be to explore how to model and quantify image content complexity in a systematic and inexpensive way.

Prior studies of end-user image-searching behavior, thesaurus use, and query expansion have shown that the search vocabularies of users and their perceptions of thesaurus and metadata usefulness may vary with the task type as well as by user characteristics, such as domain knowledge and familiarity with the system (e.g., Choi, 2008; Cunningham & Masoodian, 2006; Efthimiadis, 2000; Greenberg, 2001b; Hastings, 1994, 1999). Furthermore, there is a significant body of literature on the quality of indexing. The measures of document- and term-level indexing quality found in the literature include completeness and precision (Rolling, 1981). Completeness measures the fraction of the relevant index terms assigned, whereas precision refers to the fraction of index terms that are relevant. These measures are related to the measures of precision and recall used in evaluating the effectiveness of an information retrieval system (Cleverdon, 1997; Salton & McGill, 1982). A more complete set of indexing quality measures was defined by Soergel (1994): completeness (the fraction of the set of all relevant terms used), purity (the fraction of correctly rejected irrelevant terms), exhaustiveness (the extent to which the concepts are covered by the terms), specificity (the generic level of concepts expressed by the terms), and structure of the terms (e.g., the hierarchy of terms that supports indexing and inclusive searching, the role of indicators that specify the role of terms, and links that state the relationships between the terms). Wacholder and Liu (2006), on the other hand, defined two properties of index term quality for textual items: coherence and specificity. They found that users preferred coherent (i.e., meaningful) and specific terms. The latter property was often correlated with term length or complexity. The findings of this study provide interesting insights into the relationship between user demographics and user perception of index term usefulness, as well as the relationships between user characteristics and user "prolificacy" or quality of indexing, as measured by the number of tags assigned to the photograph. It is important to note that the number of terms used is a simple and shallow metric of indexing completeness. It assumes that all the assigned tags are valid and relevant terms. That is, the metric does not evaluate the intrinsic quality (e.g., spelling accuracy) or the relational quality (e.g., their relevance and importance to the content and context of the photographs) of individual index terms in the set. Future work might involve replicating the analysis with a deeper and more comprehensive metric(s) that would measure additional dimensions of image indexing quality (e.g., precision, specificity) to see if the relationships identified in this study would hold. Furthermore, it would be interesting to examine whether the index and query term quality metrics defined for textual documents are directly applicable to visual media as well.

The analysis of the relationships between participant demographics and the number of query terms used in a search task revealed that older, native-English-speaking male participants and participants with greater tagging experience used longer queries. However, a negative relationship was found between indexing experience and the number of terms used in queries, which differed from the positive relationship between indexing experience and the number tags used in the description task. The negative relationship between education level and the numbers of query terms and tags is also intriguing and calls for further research.

8. Conclusion

Ample suggestions have been made in the literature that social metadata can complement traditional methods of indexing and using controlled vocabularies. To our knowledge, this is the first study to confirm those suggestions with the findings of controlled end-user experiments. The outcomes of the study include a framework and methods for measuring the added value of socially created metadata to a KOS. The use of knowledge representation and organization tools (thesauri, taxonomies, ontologies) is ubiquitous. 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. In addition, the study explored the relationships between user demographics and users' perceptions of the value of terms, as well as the relationship between user demographics and indexing quality, as measured by the number of tags used. Identifying the characteristics of a high-quality indexer is essential not only for educating and training future indexers and catalogers, but also for making metadata generation processes in social tagging and content creation systems more effective by predicting the quality of contributions based on the characteristics of the members, thus enabling intelligent task routing.

The study has several limitations. The subjects were self-selected from a single academic department. In addition, although participants were free to use any term to describe or search for a photograph, the formal experimental settings of the study could have an effect on their decision-making about the applicability and usefulness of a term and might discourage or alternatively encourage the use of certain terms.

Also, the study's findings are based on an analysis of data collected from only two tasks: a general image description task and a known image search task. One may expect that in need-specific and unknown image seeking the user may use different sets of terms. Future work related to this study will analyze participant diaries and interviews to gain additional insight into the value structure of the user for index and query terms in different contexts and for different types of image-seeking needs.

Finally, the study used the same group of participants in the experimental tasks. Although the description and search tasks were completed more than two weeks apart from each other, and participants did not have access to the terms they used in the description task, one might expect that participants could still remember some of the terms they used in the description task. Replicating the experiment with ,a different, more representative sample of participants would be desirable and would add strength to the findings of the study.

Acknowledgements

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

References

Agirre, E., Ansa, O., Hovy, E., & Martinez, D. (2000). Enriching very large ontologies using the WWW. In *Proceedings of the ECAI Ontology Learning Workshop*. Berlin, Germany. Retrieved January 31, 2011, from http://arxiv.org/abs/cs/0010026

Anderson J., & Perez-Carballo, J. (2001a). The nature of indexing: How humans and machines analyze messages and texts for retrieval—Part I: Research, and the nature of human indexing. *Information Processing and Management, 37,* 231–254.

Anderson J., & Perez-Carballo, J. (2001b). The nature of indexing: How humans and machines analyze messages and texts for retrieval—Part II: Machine indexing, and the allocation of human versus machine effort. *Information Processing and Management, 37,* 255–277.

Bailey, K. (1994). *Methods of social research* (4th ed.). New York, NY: The Free Press.

Blair, D. C. (1986). Indeterminacy in the subject access to documents. *Information Processing and Management, 22,* 229–241.

Blair, D. C. (1996). STAIRS redux: Thoughts on the STAIRS evaluation. *Journal of the American Society for Information Science, 47*(1), 4–22.

Brooks, T. A. (1993). All the right descriptors: A test of the strategy of unlimited aliasing. *Journal of the American Society for Information Science, 44*, 137–147.

Buckland, M. (1999). Vocabulary as a central concept in library and information science. In T. Arpanac et al. (Ed.), *Proceedings of the Third International Conference on Conceptions of Library and Information Science* (pp. 3–12). Zagreb, Croatia: Lokve. Retrieved June 28, 2004, from http://www.sims.berkeley.edu/~buckland/colisvoc.htm

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*, 175–193.

Choi, Y. (2008). Analyzing image searching on the Web: How do undergraduates search and use visual information? In *Final Report, 2008 OCLC/ALISE Library and Information Science Research Grant*

Project. Dublin, OH: OCLC Research. Retrieved August 17, 2009, from http://library.oclc.org/cdm4/item_viewer.php?CISOROOT=/p267701coll27&CISOPTR=276

Chung, E., & Yoon, J. (2008). A categorical comparison between user-supplied tags and web search queries for images. *Proceedings of the American Society for Information Science and Technology, 45*(1), 1-3. Retrieved January 31, 2011, from http://dx.doi.org/10.1002/meet.2008.1450450392

Cleverdon, C. (1997). The Cranfield Tests on index language devices. In K. Sparck Jones & P. Willet (Eds.), *Readings in information retrieval. Morgan Kaufmann multimedia information and systems series* (pp. 47–59). San Francisco, CA: Morgan Kaufmann.

Cover, T., & Thomas, J. (1991). *Elements of information theory*. New York, NY: Wiley.

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, NY: ACM.

Cunningham, S., & Masoodian, M. (2006). Looking for a picture: An analysis of everyday image information searching. In G. Marchionini & M. Nelson (Eds.), *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 198–199). New York, NY: ACM.

Efthimiadis, E. N. (2000). Interactive query expansion: A user-based evaluation in a relevance feedback environment. *Journal of the American Society for Information Science, 51*, 989–1003.

Fidel, R. (1991). Searchers' selection of search keys: II. Controlled vocabulary or free-text searching. *Journal of the American Society for Information Science and Technology, 42*, 501–514.

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

Greenberg, J. (2001a). Optimal query expansion (QE) processing methods with semantically encoded structures thesauri terminology. *Journal of the American Society for Information Science and Technology, 52*, 487–498.

Greenberg, J. (2001b). Quantitative categorical analysis of metadata elements in image applicable metadata schemas. *Journal of the American Society for Information Science and Technology, 52,* 917–914.

Greenberg, J. (2002). Semantic web construction: An inquiry of authors' views on collaborative metadata generation. In *Proceedings of the International Conference on Dublin Core and Metadata for e-Communities* (pp. 45–52). Firenze, Italy: Firenze University Press.

Hastings, S. (1994). Query categories in a study of intellectual access to digitized art images. *Proceedings of the 58th Annual Meeting of the American Society for Information Science*, 32, 3-8.

Hastings, S. (1999). Evaluation of image retrieval systems: Role of user feedback. *Library Trends*, *48*, 438–452.

Heidorn, B. (1999). Image retrieval as linguistic and nonlinguistic visual model matching. *Library Trends, 48*, 303–325.

Hersh, W., Pentecost, J., & Hickam, D. (1996). A task-oriented approach to information retrieval evaluation. *Journal of the American Society for Information Science, 47*(1), 50–56.

Heyman, P., & Garcia-Molina, H. (2006). *Collaborative creation of communal hierarchical taxonomies in social tagging systems* (Tech. Rep.). Palo Alto, CA: Stanford University. Retrieved August 7, 2009, from http://dbpubs.stanford.edu/pub/showDoc.Fulltext?lang=en&doc=2006-10&format=pdf&compression=&name=2006-10.pdf

Hjørland, B. (2008). What is knowledge organization (KO)? *Knowledge Organization, 35*(2/3), 86–101.

Jörgensen, C. (1995a). Classifying images: Criteria for grouping as revealed in a sorting task. In R. Schwartz (Ed.), *Advances in classification research: Vol. 6. ASIS monograph series: Proceedings of the 6th ASIS SIG/CR Classification Research Workshop* (pp. 45–64).

Jörgensen, C. (1995b). *Image attributes: An investigation*. Unpublished Ph.D. thesis, Syracuse University, Syracuse, NY.

Jörgensen, C. (1996, October). *Indexing images: Testing an image description template*. Paper delivered at the 59th Annual Meeting of the American Society for Information Science, Baltimore, MD.

Jörgensen, C. (1997). Challenges in image classification. In E. Efthimiadis (Ed.), *Advances in classification research: Vol. 8. ASIS monograph series: Proceedings of the 8th ASIS SIG/CR Classification Research Workshop* (pp. 92–98).

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

Jörgensen, C., & P. Jörgensen (2002, January). Testing a vocabulary for image indexing and ground truthing. In G. B. Beretta & R. Schettini (Eds.), *Internet Imaging III: Vol. 4672. Proceedings of SPIE* (pp. 212–215).

Jörgensen, C., Stvilia, B., & Jörgensen, P. (2008). Is there a role for controlled vocabulary in taming tags? In J. Lussky (Ed.), *Proceedings of the 19th Workshop of the American Society for Information Science and Technology Special Interest Group in Classification Research* (pp. 000–000). Columbus, OH.

Klavans, J. L., Sheffield, C., Abels, E., Lin, J., Passonneau, R., Sidhu, T., & Soergel, D. (2009). Computational linguistics for metadata building (CLiMB): Using text mining for the automatic identification, categorization, and disambiguation of subject terms for image metadata. *Multimedia Tools and Applications 42*, 115–138.

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

Lin, X., Beaudoin, J. E., Bui, Y., & Desai, K. (2006). Exploring characteristics of social classification. In J. Furner & J. T. Tennis (Eds.), *Advances in classification research: Vol. 17. Proceedings of the 17th ASIS&T Classification Research Workshop*. Retrieved August 17, 2010, from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.91.3743

Mani, I., Samuel, S., Concepcion, K., & Vogel, D. (2004). Automatically inducing ontologies from corpora. In *Proceedings of CompuTerm 2004: 3rd International Workshop on Computational Terminology, COLING '2004*. Retrieved August 17, 2010, from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.120.3159

Marchetti, A., Tesconi, M., Ronzano, F., Rosella, M., & Minutoli, S. (2007). SemKey: A semantic collaborative tagging system. In *Proceedings of 16th International World Wide Web Conference* (*WWW2007*; pp. 000–000). Retrieved August 17, 2010, from http://www2007.org/workshops/paper_45.pdf

Mathes, A. (2004). *Folksonomies: Cooperative classification and communication through shared metadata*. Retrieved August 17, 2009, from http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

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

Medelyan, O., & Milne, D. (2008). Augmenting domain-specific thesauri with knowledge from Wikipedia. In *Proceedings of the NZ Computer Science Research Student Conference* (*NZCSRSC 2008*). Retrieved January 31, 2011, from http://www.cs.waikato.ac.nz/~olena/publications/nzcsrsc08_medelyan_milne.pdf

Mika, P. (2007). Ontologies are us: A unified model of social networks and semantics. *Journal of Web Semantics, 5*(1), 5–15.

Milne, D. & Witten, I. (2008). Learning to link with Wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management* (*CIKM '08*; pp. 509–518). New York, NY, ACM.

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.

Nelson, S., Johnston, D., & Humphreys, B. (2001). Relationships in medical subject headings. In A. Bean & R. Green (Eds.), *Relationships in the organization of knowledge* (pp. 171–184). New York, NY: Kluwer Academic.

Perez, J. (2009). Google rolls out semantic search capabilities. *PCWorld.* Retrieved August 17, 2009, from http://www.pcworld.com/businesscenter/article/161869/google_rolls_out_semantic_search_capabilities.html

Rasmussen, E. M. (1997). Indexing images. In M. E. Williams (Ed.), *Annual review of information science and technology* (Vol. 32, pp. 169–196). Medford, NJ: Learned Information.

Rolling, L. (1981). Indexing consistency, quality and efficiency. *Information Processing & Management, 17*(2), 69–76.

Rorissa, A. (2010). A comparative study of flickr tags and index terms in a general image collection. *Journal of the American Society for Information Science and Technology, 61*(11), 2230 - 2242.

Rorissa, A., & Iyer, H. (2008). Theories of cognition and image categorization: What category labels reveal about basic level theory. *Journal of the American Society for Information Science and Technology, 59*, 1383–1392.

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

Salton, G. (1986). Another look at automatic text-retrieval systems. *Communications of the ACM, 29*, 648–656.

Salton, G., & McGill, M. (1982). *Introduction to modern information retrieval.* New York, NY: McGraw-Hill.

Sarjant, S., Legg, C., Robinson, M., & Medelyan, O. (2009). "All You Can Eat" Ontology-building: feeding Wikipedia to Cyc. In *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01* (WI-IAT '09), Vol. 1. IEEE Computer Society, Washington, DC, 341-348.

Schwartz, C. (2008). Thesauri and facets and tags, oh my! A look at three decades in subject analysis. *Library Trends, 56*, 830–842.

Shatford, S. (1986). Analyzing the subject of a picture: A theoretical approach. *Cataloging and Classification Quarterly, 6*(3), 39–62.

Shiri, A., & Revie, C. (2006). Query expansion behavior within a thesaurus-enhanced search environment: A user-centered evaluation. *Journal of the American Society for Information Science and Technology*, *57*, 462–478.

Sigurbjörnsson, B., & van Zwol, R. (2008). Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th international conference on World Wide Web* (WWW '08). ACM, New York, NY, 327-336.

Smeulders, A., Gupta, A., & Jain, R. (2000). Content based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*, 1349−1380.

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

Soergel, D. (1994). Indexing and retrieval performance: The logical evidence. *Journal of the American Society for Information Science, 45*, 589–599.

Sparck Jones, K. (2005). Meta-reflections on TREC. In E. M. Voorhees & D. K. Harman (Eds.), *TREC: Experiment and evaluation in information retrieval* (pp. 421–448), Cambridge, MA: MIT Press.

Springer, M., Dulabahn, B., Michel, P., Natanson, B., Reser, D., Woodward, D., & Zinkham, H. (2008). *For the common good: The Library of Congress Flickr pilot project.* Retrieved August 19, 2009, from http://www.loc.gov/rr/print/flickr_report_final.pdf

Strong, D., Lee, Y., & Wang, R. (1997). Data quality in context. *Communications of the ACM, 40*, 103–110.

Stvilia, B. (2007). A model for ontology quality evaluation. *First Monday, 12*(12). Retrieved January 31, 2011, from http://frodo.lib.uic.edu/newjournals/index.php/FM/article/viewArticle/453

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

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

Stvilia, B., & Jörgensen, C. (2010). Member activities and quality of tags in a collection of historical photographs in Flickr. *Journal of the American Society for Information Science and Technology, 61*, 2477–2489.

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*, 1720–1733.

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.

Stvilia, Jörgensen, & Wu (in preparation). What makes a tag useful: the user perspective.

Svenonius, E. (1986). Unanswered questions in the design of controlled vocabularies. *Journal of the American Society for Information Science, 37*, 331–340.

Tan, Y., Kan, M., & Lee, D. (2006). Search engine driven author disambiguation. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries* (JCDL '06). ACM, New York, NY, 314-315..

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

Trant, J. (2008). *Keynote address*. Paper presented at the Dublin Core Annual Meeting, Humboldt-University, Berlin, Germany. Slides retrieved August 17, 2009, from http://conference.archimuse.com/blog/jtrant/tagging_and_folksonomy_keynote_dc2008

Trant, J. (2009). Studying social tagging and folksonomies: A review and framework. *The Journal of Digital Information, 10*(1). Retrieved August 17, 2009, from http://dlist.sir.arizona.edu/2595/

Urbain, J., Goharian, N., & Frieder, O. (2007). Combining semantics, context, and statistical evidence in genomics literature search. In *IEEE 7th International Symposium on BioInformatics and BioEngineering* (pp. 1313–1317). Retrieved August 19, 2009, from http://www.ir.iit.edu/publications/downloads/bibe07UrbainX5.pdf

Vakkari, P., Pennanen, M., & Serola, S. (2003). Changes of search terms and tactics while writing a research proposal: A longitudinal case study. *Information Processing and Management, 39*, 445–463.

van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworths, London. Retrieved, April 30, 3011, from http://frodo.lib.uic.edu/newjournals/index.php/FM/article/viewArticle/453

von Ahn, L., & Dabbish, L. (2004). Labeling images with a computer game. In *Proceedings of the CHI 2004* (pp. 319–326). New York, NY: ACM.

Wacholder, N., & Liu, L. (2006). User preference: A measure of query-term quality. *Journal of the American Society for Information Science and Technology, 57*, 1566–1580.

Wetterstrom, M. (2008). The complementarity of tags and LCSH—A tagging experiment and investigation into added value in a New Zealand library context. *The New Zealand Library and Information Management Journal, 50*, 296–310.

Yi, K., & Chan, L. M. (2009). Linking folksonomy to Library of Congress Subject Headings: An exploratory Study. *Journal of Documentation, 65*(6), 872-900.

Yoon, J. (2009). Towards a user-oriented thesaurus for non-domain-specific image collections. *Information Processing and Management, 45*(4), 452–468