Margaret E. I. Kipp
University of Wisconsin-Milwaukee

Jihee Beak
University of Wisconsin-Milwaukee

Poster: Examining Studies Comparing Tags and Controlled Vocabularies

Abstract: Tags have been compared to controlled vocabulary terms and have been suggested as replacements or enhancements in indexing. This paper explores tagging and controlled vocabulary studies in the context of studies examining title and author keywords or user search terms to determine the methods used and impact of these studies.

Résumé: Les étiquettes ont été comparés avec les vedettes-matière et ont été suggérés comme remplacements ou comme addition à l'indexation. Cet article examine la recherche sur l'étiquetage et les vedettes-matière dans le cadre d'études portant sur les mots-clés de titre ou d'auteur ou les termes des requêtes d'usager pour déterminer les méthodes utilisées et l'impact de ces études.

1. Introduction

Social tagging has become increasingly popular since its beginnings on social bookmarking sites like delicious.com. Proponents and early research suggest that social tagging provides some measure of subject access to items which may otherwise lack such access (Mathes 2004; Kipp 2005). This paper examines the set of studies comparing controlled vocabularies, user tags and occasionally author keywords assigned to various documents to determine the usefulness of all these different terms for indexing or information retrieval.

2. Background

Social tagging is a new phenomenon, but earlier studies examined the use of title keywords, author keywords and user search terms for indexing. Research by Montgomery and Swanson (1962) demonstrated a high degree of concurrence between title keywords and subject headings (86%), but found 14% of articles were unindexable based solely on title. Later research showed that such concurrence of terms was field dependent (O'Connor 1964; Frost 1989). Schultz, Schultz and Orr (1965) found that author keywords matched subject terms more closely than title terms. More recent research determined that author keywords matched tags more closely than subject headings (Kipp 2005). Gross and Taylor (2005) used transaction logs to examine user search terms and determined that one third of keyword searches would fail without indexing due to differing terminologies. All of these studies compared terms assigned by different user groups to draw conclusions about the usability of these terms for information retrieval or information organisation.

4. Methodologies and Results

This study examined 33 published studies which compared tags and other vocabularies. Early analysis showed that each study used differing methods and data for analysis but drew similar conclusions based on the results of matches between tags and other vocabularies (Kipp 2011).

A majority of the studies examined terms assigned to documents in LibraryThing, Connotea and CiteULike (Table 1). The next most common source was Steve.museum and Flickr, which deal with images. Many studies used existing subject heading lists and thesauri for comparisons. LCSH was most commonly used followed by MeSH (Medical Subject Headings). A few studies used comparisons to author-assigned keywords instead of controlled vocabularies.

Tools	Frequency
LibraryThing	8
Connotea	6
CiteULike	5
Steve.museum	4
Flickr	3
Delicious	3

Table 1: Most common tools

Many studies examined general documents indexed on sites with no specific subject orientation, but some studies did examine specific subjects (Table 2). Studies of title and author keywords determined that match percentages varied widely depending on the field of study (O'Connor 1964).

Subjects	Frequency
General	11
Biology/Sciences	8
Museum Studies	5
Fiction and Nonfiction	2
Political and Social Sciences	2
Library and Information Studies	1

Table 2: Subject areas examined

While all the studies performed some form of match between terms, most used differing methods which were described in limited terms (Table 3). The majority of studies which did not use a form of partial syntactic match (e.g. stemming) or semantic match (synonyms, thesaurus categories) suggested that such matching would generate higher percentage matches (Kipp 2011).

Types of Match	Frequency
Syntactic Match - Exact	7
Syntactic Match - Partial	4
Semantic Match	3

Table 3: Type of match used by the studies

Some issues which all such studies need to examine specifically, but which many did not discuss in detail, are issues of how to generate the comparisons. Even studies which claimed to use exact match often relied on a limited amount of stemming to remove distinctions between plurals and on adjusting the case of terms or splitting terms joined by a hyphen or underscore in order to generate a high percentage of matches (e.g. splitting 'information_retrieval' in order to match Information Retrieval, a subject heading). Issues of variations in the formatting and spelling of terms are core issues in the development of controlled vocabularies and have had a negative impact on retrieval and indexing as users fail to find documents if they do not use the exact term and format as used in the indexing terms or in the document itself.

4. Discussion and Conclusions

This study examines tagging studies which have compared tags and other vocabularies. The goal of these studies is to examine the possibility of using social tagging to enhance indexing or retrieval of documents through subject access. However, studies examining types of tags show that tagging provides more than simply subject access and in order to determine the real utility of tagging, studies should expand their comparisons beyond controlled vocabularies to examine format terms, name authorities and other possible metadata from surrogate records.

5. References

Frost, C. O. 1989. Title Words as Entry Vocabulary to LCSH – Correlation Between Assigned LCSH Terms and Derived Terms from Titles in Bibliographic Records with Implications for Subject Access in Online Catalogs. *Cataloging & Classification Quarterly* 10(1):165–179.

Gross, T. and Taylor, A. G. 2005. What Have We Got to Lose? The Effect of Controlled Vocabulary on Keyword Searching Results. *College & Research Libraries* 66(3): 212–30.

Kipp, M. E. I. 2005. Complementary or Discrete Contexts in Online Indexing : A Comparison of User, Creator, and Intermediary Keywords. *Canadian Journal of Information and Library Science* 29(4):419–436.

Kipp, M. E. I. 2011. Controlled vocabularies and tags: An analysis of research methods. *North American Symposium on Knowledge Organization (NASKO), Toronto, June 15-16, 2011*.

Mathes, A. 2004. Folksonomies - Cooperative Classification and Communication Through Shared Metadata. http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

Montgomery, C. and Swanson, D. R. 1962. Machinelike indexing by people. *American Documentation* 13(4):359–366.

O'Connor, J. 1964. Correlation of indexing headings and title words in three medical indexing systems. *American Documentation* 15(2):96–104.

Schultz, C. K., Schultz, W. L., and Orr, R. H. 1965. Comparative indexing: Terms supplied by biomedical authors and by document titles. *American Documentation* 16(4):299–312.