Malone, C.K. (2007). When more is better: A counter-narrative regarding keyword and subject retrieval in digitized diaries. 18th Annual ASIS SIG/CR Classification Research Workshop doi: 10.7152/acro.v18i1.12870

# When More is Better:
# A Counter-Narrative Regarding Keyword and Subject Retrieval in Digitized Diaries

Cheryl Knott Malone
University of Arizona
School of Information Resources & Library Science
1515 East First Street
Tucson, AZ 85719
ckmalone@u.Arizona.edu

## Abstract

Many commercial full-text databases and digital libraries provide keyword and preferred-term (subject) indexing, but few allow participatory tagging of content by users or provide ontologies in support of natural language information retrieval. Consequently, keyword and subject searching strategies still matter. But keyword searching, because it can yield results high in recall and low in precision, is often seen as a beginner's strategy best replaced by subject searching using authoritative headings and descriptors. In certain circumstance explored in this essay, keyword searching may be quite effective in and of itself for retrieving digitized primary sources for the study of history.

## Background and Purpose of the Study

The title "when more is better" refers to two different aspects of information organization for effective discovery and retrieval. First, it refers to instances where using keywords for greater recall (and less precision) may be better than identifying and using the corresponding controlled vocabulary terms. Second, it refers to a trend toward more in information organization, particularly as more relates to social tagging and the development of ontologies.  Some early research on social tagging suggests that more tags by more individuals means more accuracy and authority, particularly when coupled with algorithms and architectures that supplement and support tagging (Winget 2006). Other research related to the development of ontologies suggests that more linkages to existing language tools such as subject thesauri and taxonomies will enhance the effectiveness of ontologies for natural language processing and the machine learning integral to such processing (Khoo et al., 2007). Campbell (2006) has suggested that social tagging and ontologies (and other formal knowledge organization tools) are interrelated through "complementary purposes," and that there is a place for both approaches (p. 14). Despite these recent trends, a suggestion that more is better counters the conventional wisdom in information retrieval that privileges less, especially when less means greater precision (and thus lower recall). The customary narrative is perhaps most noticeable in searching instructions that urge database users to deploy the thesaurus of preferred terms for their chosen topics. But it also appears to be assumed in much of the research into information seeking behavior.

ISSN: 2324-9773

The narrative is essentially this:

> A person who is using library catalogs and electronic databases for researching a topic typically uses keywords labeling that topic to discover and retrieve the desired relevant information. A keyword search is likely to yield results that match the query literally but are nevertheless irrelevant to the user as keyword searching falls victim to the vagaries of natural language with its synonyms and variant spellings and double meanings. Consequently, the information seeker should find and learn to use the system's controlled vocabulary, whether subject headings in a library catalog or descriptors in a commercial database, to retrieve fewer but more relevant results.

In the age of Google, the last sentence in this narrative might be replaced with this one:

> Nevertheless, with the right ranking algorithm, such concerns are minimized as the system takes into account factors beyond the literal keywords input in the search box to ensure that the links presented on the first screen of results are relevant to the query.

Google continues to tweak its algorithms, and information organization experts continue to consider other alternatives as evidenced by the current interest in social tagging as useful metadata and by the development of elaborate, sophisticated ontologies. But for the time being, hundreds of databases accessible via library portals neither invite participatory tagging of resources nor offer ontology-supported natural language processing. Consequently, the discipline-specific thesaurus, when available in a commercial database, does provide useful assistance not only by revealing the preferred terms for topics but also by indicating the relationships among terms in an effort to guide users to relevant material. As Walker et al. (1999) point out in their textbook on information retrieval, assignation of thesaurus terms is especially helpful in databases where what's being searched is the record representing the item rather than the full text of the item. Assignation of thesaurus terms adds to the searchable metadata contained in surrogates that can lead to the discovery and retrieval of pertinent items.

Yet in some databases, the surrogate record includes not only the metadata but also the data, the full text, with each substantive keyword indexed and searchable. And in some full-text databases, there is a list of preferred terms, a controlled vocabulary that does not reach the exalted level of thesaurus because it is merely a list without any indications of the relationships among terms included. Such lists can be useful for disambiguating synonyms and variant spellings, but, depending on their design and application, their usefulness is limited. The use of such terms instead of the seeker's own thought-up keywords as labels for a research topic may in fact impede discovery and retrieval. In such cases, the preferred terms may be less than ideal because they exhibit the wisdom of a few experts rather than collective efforts of participating users in a social network. And the keyword indexing, while increasing recall, may be less than ideal because it lacks the semantic and contextual components ontologies supporting natural language processing can supply. Critiques of social tagging as metadata reveal continuing discomfort with the ambiguities inherent in dynamic natural language. Similarly, interest in developing ontologies reveals a drive to disambiguate. In some contexts, however, keyword searching can lead to exactly the right sort of discovery and retrieval. One such context is examined here -- an electronic database of full-text primary sources for the study of history.

## Methods

In this study, I evaluate specific examples of keyword and subject searching of a single database, *British and Irish Women's Letters and Diaries* (BIWLD) in order to suggest that in some cases more results may be better than fewer results, even when those additional results represent a burden for the researcher. BIWLD is an electronic database totaling approximately 100,000 digitized pages. It contains the full text of 294 previously published diaries and collections of letters written by women between the mid-sixteenth and mid-twentieth centuries. The publisher also included about 4,000 pages of previously unpublished material (Alexander Street Press, 2005a).

This study focuses on the diaries; the database search system allows for searching of only the diaries included in BIWLD. The database does not have a record for each separate daily entry from a diary. Instead, the database records are "diary documents," each of which is the keyword-indexed full text of one month's worth of diary entries. Human indexers also assign preferred terms from a controlled-vocabulary list to diary-document records. Diary documents may include any number of daily entries, depending on how often a woman wrote in her diary in any given month. Of the 294 sources included in the collection, 131 include diary results. The database has 4,568 full-text diary-month records. Only about 10 percent of the diary documents are from previously unpublished resources. Since the collection of diaries is relatively small and includes relatively little previously unpublished work, the key value-added feature of the electronic database is the provision of access points, namely keywords and subject terms, that lead searchers into the full texts. And Alexander Street Press (2005b) highlights this feature in its marketing material.

I performed a keyword search and a correlated subject search, limiting all results to diaries only. The keyword search was for *magazine*, truncated with the * device to include the plural *magazines*. The correlating subject search for the controlled-vocabulary subject heading *magazines* yielded many fewer results, as would be expected. This search relates to my interest in the history of print culture and was designed to discover under what circumstances women mentioned magazines in their diaries and which magazines they mentioned. My objective was to identify and analyze retrieval issues apparent in the results of corresponding keyword and subject searches. I did not attempt to measure recall and precision of the sets of results because I did not know how many relevant documents were in the database for each search and because, as Harter (1992; 1996) has shown, relevance judgments vary according to the stage of the search and the searcher's perceptions.

## Magazines in Diaries

A search for keyword *magazine** in only the diaries yields 125 results. Of those, 14 are from previously unpublished diaries. Since I was interested in how women discussed magazines in their diaries, it was necessary to take a closer look at the 125 results to eliminate any mentions other than those in diary entries. As Table 1 shows, of the 125 results, 49 (39%) are from the editorial notes appended to the published diaries rather than diary entries penned by the women themselves; 76 (61%) might be assumed to be relevant in the sense that they are mentions of the word *magazine* or *magazines* within a diary entry. However, 34 of the occurrences of the keyword *magazine** are used in the sense of storehouses rather than periodicals and 2 of the occurrences are references to "magazine suiting," a type of fabric. Of the 76 diary months mentioning *magazine* or *magazines* in diary entries rather than editorial notes, 36 are false drops. Consequently, 40 (32%) of 125 results in the set are relevant to the purpose of our query. In these 40, the keyword is used either in a proper name of a periodical or to denote an unnamed magazine or the category of magazines.

Table 1. Contexts of Keyword *Magazine** Results

| Context | Number | Percentage |
|---|---|---|
| Editorial matter | 49 | 39 |
| Meaning other than periodicals | 36 | 29 |
| Meaning "magazines" | 40 | 32 |
| Total | 125 | 100 |

In an earlier study, Malone (2007) searched for the keyword *library* and found far fewer false drops: 90 (19%) out of 467. It's not clear what proportion of false to relevant hits is optimum or even acceptable, but when two-thirds of a results set is irrelevant, most searchers probably would fault the search strategy and reformulate it. One way to do that would be to eliminate certain types of results, such as editorial comments. The search system in this particular database does not allow the use of the Boolean *not* to eliminate editorial matter from a set of diary entries. Another way to eliminate homographs such as *magazine* is to use a controlled vocabulary term that captures the meaning the searcher intended. That is an option in this database.

A subject search for the subject heading *magazines* yields only three results.
Two are from the diary of Melesina Chenevix St. George Trench (1862a; 1862b). In January 1819, she wrote disapprovingly about an article in the *Anti-Jacobin* regarding the writer and intellectual Madame de Staël. In May of 1826, she wrote approvingly about an article discussing translations of Goethe in the *Quarterly Review.* Each passage is six lines long and is as much about the article as about the magazine, so it is not entirely clear why the subject heading *magazines* was assigned. But one reason may have been that neither *magazine* nor *magazines* appears in these two passages. Consequently, they are not included in the 125 results for the earlier keyword search. The same is true of the third result in the subject search set.

The third result was written by Sydney Owenson (Lady Morgan) in November of 1827. A brief passage that criticizes four periodicals by name, it seems sufficiently about magazines to be assigned that subject heading despite the fact that the keyword *magazine* or *magazines* does not appear in this diary document. In these three cases, the human indexer appears to have recognized that a searcher interested in women's diary entries mentioning magazines would want to discover passages that mention periodical by title without mentioning the keyword *magazine* or *magazines.* Because there is no overlap between the keyword and subject searches in this study, the only factor guiding the assignment of the subject heading *magazines* to a diary document appears to be the absence of the keyword in the document.

## Findings and Discussion

Malone, C.K. (2007). When more is better: A counter-narrative regarding keyword and subject retrieval in digitized diaries. 18th Annual ASIS SIG/CR Classification Research Workshop doi: 10.7152/acro.v18i1.12870

Our notions of "aboutness" may be a problem here if we think of it as the subject matter of a work. A diary entry is necessarily different from other kinds of works, such as print and electronic monographs; journal, magazine, and newspaper articles; and films. A diary entry, or even a month's worth of diary entries, resembles stream-of-consciousness expression as it jumps from one topic to another often without the logical progression of an argument that readers would expect to find in a formal essay on a particular theme or issue. A diary entry, or a month's worth of diary entries, may mention directly or indirectly many different topics, making it difficult to determine which of the topics warrants an aboutness determination and thus the assignment of an authoritative subject heading or descriptor. Warrant suggests that there is a topical tipping point that pushes a passage or a work into aboutness territory. But warrant is from the perspective of the professional cataloger or indexer focused on describing the work itself. What is difficult to account for is the user's intention. In this case, would the user consider any mention of magazines (in the periodicals sense) to be sufficiently about magazines for the purposes of her research? The user's intention regarding the topic being researched -- or we might say the user's conceptualization of the topic in the particular context in which the user is operating -- cannot be captured and represented at the level of the professional indexer because it, like relevance judgments, is dynamic and idiosyncratic. One of the potential strengths of participatory tagging is that it can, at least theoretically, suggest how the creator or user of an item conceptualizes or contextualizes it by how the creator or user labels it.

The research literature on information seeking and use has verified that both aspiring and established historians rely on secondary materials such as monographs and articles as well as on primary texts, including the eyewitness contemporaneous accounts recorded in diaries and letters. (Delgadillo & Lynch, 1999; Tibbo, 2003) BIWLD is marketed to academic libraries and is intended for the use of college and university faculty and students researching historical topics broadly construed. Academic librarians in turn market the database to their clientele by offering instruction for history and women's studies courses and by listing the database on subject guides to history. MARC records are available for each whole work (not each diary document) in the database, making it possible for potential users to discover these works and the fact that digitized versions are available in BIWLD by searching OCLC's WorldCat or their local academic library catalog if the library has purchased access to the database and loaded the MARC records. A spot check of a few titles suggests that the bibliographic records include only a single Library of Congress subject heading. It seems likely that university students completing course assignments are the ones most likely to find and use the BIWLD database. This is true not only because of the way in which libraries are marketing the product but also because of the sources and methods favored by established historians. Helen Tibbo's (2003) work documents that academic historians value unpublished rather than the kinds of previously published diaries and letters that predominate in BIWLD (p. 19) and that they tend not to search large online union catalogs such as WorldCat where they would be able to discover bibliographic records representing the works in the BIWLD collection (pp. 21-22). Consequently, it seems likely that BIWLD is used mostly by undergraduate and graduate students. Recent literature has documented that students rely heavily on Internet searching (which must mean they use keyword searching almost exclusively), especially in the absence of outside intervention. (Tenopir, 2003; Greenberg, 2004). So, a full-text database of primary historical materials that is keyword indexed may fit the search styles of today's students quite well. This raises at least two questions: 1) Are students patient or motivated enough to go through the relatively high numbers of results that can be retrieved with simple keyword searches to find the best ones for their research projects? 2) Are students able to evaluate the results critically enough to eliminate the false drops and to select the best ones for their purposes?

Depending on how subject analysis is performed in a particular database, keyword search results--despite the messiness of natural language--may facilitate a researcher's discovery of novel material more effectively than subject search results. In a database of keyword-searchable digitized historical documents, keyword searching may lead to the discovery of relevant material that is not indexed by subject descriptors. It is ironic that a sophisticated researcher accustomed to subject searching may not discover relevant material unless she tries the presumably less sophisticated technique of keyword searching.

## Conclusion

Our assumption in information organization for discovery and retrieval -- and for that matter the assumption underlying Google's search engine -- has been that people do not want to discover all the mentions of a word no matter the context. At some point, more is deemed to be too much. Under some conditions, though, more can be better, even when it puts a greater burden on the researcher to process the results. The example of keyword and subject searching for *magazines* in old diaries suggests a few of the possible conditions when more is better: 1) a relatively small corpus of full text; 2) a cohesive collection of items selected for their similarities (gender of author, nationality of author, language, document type); and a user group (college students) interested in finding particular passages containing their keywords rather than browsing a text in the traditional linear way. Conclusions cannot be drawn from the single two-part search exemplar used in this report. But the results do suggest that, in the absence of participatory tagging and elaborate ontologies for natural language searching, it pays to understand the nature of the material being searched and the indexing practices applied to it. And it helps to be prepared to spend time experimenting with search strategies and reading through more results than one might enjoy in order to understand the material and the indexing. If it does nothing else, this essay reminds us how inefficient information seeking can be and suggests why we look to approaches such as social tagging and ontologies as alternatives.

## References

Alexander Street Press. (2005a). Products: British and Irish Women's Letters and Diaries. Retrieved June 19, 2007, from http://alexanderstreet.com/products/bwld.htm.

Alexander Street Press. (2005b). Information about British and Irish Women's Letters and Diaries, Release 4. Accessible from within the database.

Alexander Street Press. (2005a). British and Irish Women's Letters and Diaries. Opening screen. Accessible from within the database.

Alexander Street Press. (n.d.). Help. Accessible from within the database.

Campbell, D.G. (2006). A phenomenological framework for the relationship between the Semantic Web and user-centered tagging systems. In J. Furner and J.T. Tennis (Eds.), *Advances in classification research, vol. 17: Proceedings of the ASIS&T SIG/CR Classification Research Workshop (Austin, TX, November 4, 2006).* Retrieved September 30, 2007, http://dlist.sir.arizona.edu/1838/.

Dalton, M.S. & Charnigo, L. (2004). Historians and their information sources. *College & Research Libraries, 65,* 400-425.

Delgadillo, R. & Lynch, B.P. (1999). Future historians: Their quest for information. *College & Research Libraries 60,* 245-259.

Greenberg, J. (2004). User comprehension and searching with information retrieval thesauri. *Cataloging & Classification Quarterly 37,* 103-120.

Harter, S.P. (1992). Psychological relevance and information science. Journal of the American Society for Information Science, 43(9). 602-15.

Harter, S.P. (1996). Variations in relevance assessments and the measurement of retrieval effectiveness. Journal of the American Society for Information Science, 47(1). 37-49.

Khoo, C.S.G., Wang, J.N.W. & Chan, S. (2007). Modeling cancer treatment information with an ontology. In K.S. Raghavan (Ed.), *International Conference on Future of Knowledge Organization in the Networked Environment, 3-5 September 2007*, 155-268. Bangalore: Indian Statistical Institute.

Malone, C.K. (2007). Electronic access to women's diaries and letters: An analysis of keyword vs. subject searching. In K.S. Raghavan (Ed.), *International Conference on Future of Knowledge Organization in the Networked Environment, 3-5 September 2007*, 237-254. Bangalore: Indian Statistical Institute.

Owenson, S.L.M. (1862). Diary of Sydney Owenson, Lady Morgan, November, 1827. In *Lady Morgan's Memoirs, Autobiography, Diaries and Correspondence, vol. 2,* 242-243. London: William H. Allan & Co.

Tenopir, C. (2003). *Use and users of electronic library resources: An overview and analysis of recent research studies*. Washington, DC: Council on Library and Information Resources. Retrieved October 11, 2007, http://www.clir.org/pubs/reports/pub120/pub120.pdf.

Tibbo, H.R. (2003). Primarily history in America: How U.S. historians search for primary materials at the dawn of the digital age. *American Archivist 66*, 9-50.

Trench, M.C.S.G. (1862a). Diary of Melesina Chenevix St. George Trench, January, 1819. In R.C. Trench (Ed.), *The Remains of the late Mrs. Richard Trench, being selections from her journals, letters, and other papers,* 393-394. London: Parker, Son & Bourn.

Trench, M.C.S.G. (1862b). Diary of Melesina Chenevix St. George Trench, May, 1826. In R.C. Trench (Ed.), *The Remains of the late Mrs. Richard Trench, being selections from her journals, letters, and other papers,* 515. London: Parker, Son & Bourn.

Walker, G., Janes, J. & Tenopir, C. (1999). *Information retrieval: A dialogure of theory and practice.* 2nd. ed. Westport, CT: Libraries Unlimited.

Winget, M. (2006). User-defined classification on the online photo sharing site Flickr . . . Or, how I learned to stop worrying and love the million typing monkeys. In J. Furner and J.T. Tennis (Eds.), *Advances in classification research, vol. 17: Proceedings of the ASIS&T SIG/CR Classification Research Workshop (Austin, TX, November 4, 2006).* Retrieved September 30, 2007, from http://dlist.sir.arizona.edu/1854/