Sam Grabus, Jane Greenberg, Peter Logan, and Joan Boone. 2019. Representing Aboutness: Automatically Indexing 19th- Century Encyclopedia Britannica Entries. NASKO, Vol. 7. pp. 138-148.

Sam Grabus —Drexel University
Jane Greenberg —Drexel University
Peter Logan —Temple University
Joan Boone —Drexel University

Representing Aboutness: Automatically Indexing 19th-Century Encyclopedia Britannica Entries

Abstract

Representing aboutness is a challenge for humanities documents, given the linguistic indeterminacy of the text. The challenge is even greater when applying automatic indexing to historical documents for a multidisciplinary collection, such as encyclopedias. The research presented in this paper explores this challenge with an automatic indexing comparative study examining topic relevance. The setting is the NEH-funded 19th-Century Knowledge Project, where researchers in the Digital Scholarship Center, Temple University, and the Metadata Research Center, Drexel University, are investigating the best way to index entries across four historical editions of the Encyclopedia Britannica (3rd, 7th, 9th, and 11th editions). Individual encyclopedia entry entries were processed using the Helping Interdisciplinary Vocabulary Engineering (HIVE) system, a linked-data, automatic indexing terminology application that uses controlled vocabularies. Comparative topic relevance evaluation was performed for three separate keyword extraction algorithms: RAKE, Maui, and Kea++. Results show that RAKE performed the best, with an average of 67% precision for RAKE, and 28% precision for both Maui and Kea++. Additionally, the highest-ranked HIVE results with both RAKE and Kea++ demonstrated relevance across all sample entries, while Maui's highest-ranked results returned zero relevant terms. This paper reports on background information, research objectives and methods, results, and future research prospects for further optimization of RAKE's algorithm parameters to accommodate for encyclopedia entries of different lengths, and evaluating the indexing impact of correcting the historical Long S.

Introduction

Aboutness, a fundamental concept in the field of Information Science, refers to the topicality, subject, or meaning of an information object (Hjørland 2001). The term is frequently used in reference to the topical relevance for categorization efforts (Saracevic 2007). Aboutness can be defined in two different ways: subjectively, such as the individualistic analysis that Moby Dick is about more than just a whale (Svenonius 2000, 49), and objectively, as a set of terms that can be agreed upon as useful for database information retrieval purposes (Svenonius 2001, 46). We use language to communicate a document's aboutness, but the process of determining even objective aboutness can be challenging. Humanities resources, in particular, are inherently linguistically complex and do not benefit from "linguistic determinacy" to the degree that scientific resources can (Svenonius 2001, 49). Despite this challenge, many scholars argue that controlled vocabularies and other standardized terminologies can help represent aboutness in a uniform way, to replace or complement natural language text representation (Moens 2000, 108). Controlled vocabularies can help avoid language idiosyncrasies, such as regional word usage and spellings, idioms, and abbreviations, which render simple document keyword extraction insufficient for digital humanities information retrieval purposes (Bair and Carlson 2008). Indexing a document

with controlled vocabulary terms also helps to ensure text analysis and representation consistency (Bueno-de-la- Fuente, Mateos, and Greenberg 2016). This may be accomplished through the mapping of automatically-extracted keywords to controlled vocabulary terms, such as the Library of Congress Subject Headings (LCSH).

Although there is extensive use of automatic indexing in many day-to-day information retrieval environments, research focusing on the aboutness layer of relevance is challenging (Cosijn 2010; Saracevic 2007). Many applications apply brute force of automatic and semi-automatic indexing, but do not consider the multidimensional aspects of topic relevance evaluation. Also, limited, are the number of large-scale automatic indexing studies using controlled vocabularies. These limitations present additional challenges in moving forward with automatic indexing of digital humanities corpora, which are increasingly being made available for research. Topic relevance evaluation is particularly challenging for multidisciplinary historical humanities documents, in which the language is not domain-specific, and is frequently expressive and linguistically complex. The research presented in this paper takes steps to addresses these challenges by examining automatic indexing for multidisciplinary, digital humanities text.

This paper reports results from a comparative topic relevance evaluation for automatically indexing a large corpus of multidisciplinary 19th-century Encyclopedia Britannica entries with LCSH, using an existing automatic indexing tool called HIVE. The research was conducted over a six-month period, from June 2018-December 2018. The work was motivated, in part, by the Library Education and Data Science for the National Digital Platform (LEADS-4-NDP), an IMLS-funded fellowship program. The work was continued through the "Developing the Data Set of Nineteenth-Century Knowledge" project, an NEH-funded collaboration between Temple University's Digital Scholarship Center and Drexel University's Metadata Research Center. The next section of this paper presents background covering Temple University's 19th-Century Knowledge Project, the automatic indexing application used for this project, and three keyword extraction algorithms integrated into the application. The background is followed by the research objectives and method supporting the topic relevance precision comparison. Next, the results are presented, comparing the precision scores for automatic indexing results with three different algorithms. The results are followed by a contextual discussion, and a conclusion that discusses key findings and identifies next steps.

Background

Indexing for the 19th-Century Knowledge Project

The data used in this research was obtained through the Nineteenth-Century Knowledge Project (Logan 2019), an NEH-funded initiative to digitize historic editions of the Encyclopedia Britannica, in order to study the construction of knowledge in the nineteenth century. Due to their comprehensiveness, scope, and ubiquity, these encyclopedia editions also document changes over time in the conceptualization of knowledge in the English-speaking world. Upon completion of the large-scale digitization project, this will be the first

accurate, standards-compliant textual dataset for this corpus. All data is encoded in TEI-XML and will be made freely available when completed sometime in 2020.

The complete data set will consist of all 100,000 entries from the four major editions of the Encyclopedia that spanned the nineteenth century:

- 3rd ed., 18 vols., 1797
- 7th ed., 21 vols., 1842
- 9th ed., 25 vols., 1889
- 11th ed., 29 vols., 1911

To enrich analysis of the corpus, researchers seek to append accurate subject metadata to each individual entry file, so that the corpus can be segmented and analyzed comparatively across segments. To address this need in a feasible, cost-effective way, metadata researchers engaged in this project are exploring automatic and semi-automatic indexing approaches.

The collection presents significant challenges for accurately representing encyclopedia entry aboutness in an automated way: (1) First, as a general reference source, it includes entries from multifarious domains of knowledge. Automatic indexing tailored to specific domains is more successful than multidisciplinary materials, which puts larger demands on the controlled vocabulary and indexing algorithm. (2) Second, the corpus spans 120 years, and during that time, entry topics were described in varying ways with different semantics, which the automated indexer has to resolve. (3) Finally, entries vary substantially in length, from a single sentence to over 150,000 words, so word count has to be accounted for in the indexing algorithm, regarding minimum word frequency. To address these challenges for the 19th-Century Knowledge Project, this project requires technology that can perform automatic indexing with a multidisciplinary controlled vocabulary. This particular corpus also requires that the tool can accommodate indexing entries of extreme lengths. The HIVE technology, described in the next section, has been selected for this project. The next two subsections will discuss HIVE and its use of controlled vocabularies, as well as the three keyword extraction algorithms tested for this research.

Helping Interdisciplinary Vocabulary Engineering (HIVE) and Controlled Vocabularies

HIVE is a linked-data automatic indexing technology that works with controlled vocabularies. The demonstration systems used in this research allows the user to select one or more controlled vocabularies, upload text or input a URI, and execute an automatic indexing sequence. The system then retrieves a set of terms, based on the underlying automatic indexing algorithm. One of the chief advantages of HIVE is the ability to incorporate and use controlled vocabularies. There are many sophisticated automatic indexing algorithms that can provide good indexing results. However, they do not link to controlled vocabularies, and thus, lack standard assignment of terminology when indexing. Due to its multidisciplinary nature, expressiveness, expansiveness, and continued ubiquity among digital library collections (Walsh 2011), we have selected the Library of Congress Subject Headings as our controlled vocabulary for this phase of the project.

HIVE was initially launched in 2009 as an IMLS (Institute of Library and Museum Studies) project, coordinated with the Dryad data repository (Dryad Digital Repository 2019). This first instance of the HIVE tool (referred to from here on as HIVE 1), integrates Maui and Kea++ keyword extraction algorithms. HIVE has been enhanced over the years with new vocabularies, the RAKE keyword extraction algorithm, and most recently, scaling and display options. These latest enhancements are accessible in HIVE's most current web interface (referred to from here on as HIVE 2), which uses RAKE. HIVE 2 enables a user to specify the number of desired terms retrieved, minimum word frequency, maximum words per phrase, minimum characters per word, and display options (e.g., relevance view, alphabetical view). HIVE 1 has also been maintained fairly well, and allows the user to select either the Kea++ or Maui algorithm. This groundwork provides a good setting for investigating HIVE and its applicability for the 19th-Century Knowledge Project, and more specifically, comparing the efficacy of each algorithm integrated into the platform. The next section discusses the three keyword extraction algorithms integrated into HIVE.

Three keyword extraction algorithms

Maui and Kea++ are both machine-learning algorithms that use training data to improve the algorithm's performance. RAKE is not a machine-learning algorithm, and operates instead with the adjustment of key parameters, including minimum characters per word, maximum words per phrase, and minimum word frequencies. Testing reported here compares results using these algorithms. Features of the three algorithms are briefly outlined in the table below (Table 1).

Table 1: Features of the three keyword extraction algorithms compared in this study

	KEA++	MAUI	RAKE
Features	• TFxIDF • First occurrence • Keyphraseness (keyphrase frequency) (Medelyan and Whitten 2008)	• TF-IDF • First occurrence • Keyphraseness • Length • Node degree • Wikipedia-based keyphraseness • Spread • Semantic relatedness • Inverse Wikipedia linkage (Medelyan, Perrone, and Witten 2010)	• Word frequency • Word degree • Ratio of degree to frequency • Co-occurrences • Stop words • Adjustable parameters: -Minimum characters per word -Maximum words per phrase -Minimum word frequency (Rose et al. 2010)
Machine-Learning	Yes	Yes	No

Research Objectives

The principal goal of this work is to perform comparative topic relevance evaluation for an existing automatic indexing tool (HIVE), using entries across four 19th-century Encyclopedia Britannica editions. This analysis was performed to determine which of the three keyword extraction algorithms in HIVE yields more precise results when automatically indexing entries. The next section reports methods and steps taken to address these objectives.

Methods

We approached this research by performing a baseline experiment on the Knowledge Project's Encyclopedia entries to refine the parameters of HIVE's algorithms, then perform a topic relevance precision analysis to determine which algorithm produces the highest quantity of relevant results. The overriding goal of this project was to optimize the automatic generation of controlled vocabulary terms for integration into the TEI- XML headers for the individual encyclopedia entry files.

The protocol for performing this research involved the following steps:

1. Compiling a sample for testing:
   1.1 At the onset of this project, digitization efforts on the letter R encyclopedia entries across four editions were complete. Because we exclusively had access to the letter R encyclopedia entries, we eliminated "blind" entries from our sample: entries which simply redirect the reader to another entry (e.g., 9th edition entry on Refraction: "See Light and Optics").
   1.2 For comparative purposes across editions, we intersected a full list of entry terms from all four digitized editions of the Encyclopedia Britannica to determine which terms appear in all four editions. The end result was a list of 50 encyclopedia terms, with four entries each (across all four editions), for a total of 200 total entries. We created .TXT files for each entry in our sample.
   1.3 Due to the extensive length of some entries, a baseline sample of 10 entries was randomly selected from our research data collection of 200. The entry lengths range between 174 and 20,158 words.
2. Uploading entries to HIVE: We uploaded each entry in our sample to the HIVE platform, selected Library of Congress Subject Headings as our controlled vocabulary, and processed the text to generate automatic indexing results.
3. Improving HIVE's interface: After a first-phase relevance evaluation for the preliminary results, HIVE's Senior Research Analyst developed user-interface features for adjusting the RAKE keyword extraction algorithm parameters.
4. Re-testing and optimization: We re-tested our sample terms with HIVE 2, adjusting algorithm parameters to optimize our indexing results.

5. Relevance Analysis: Our indexing results were used for comparing the performance for each of the three keyword extraction algorithms. Precision analysis of topical relevance was guided by methods for automatic indexing evaluation by Golub et al., 2016. For each of the three algorithms tested, we performed a first-pass precision analysis to determine the proportion of HIVE results that were relevant results. Since HIVE ranks the results according to relevance, we also determined what proportion of HIVE's highest-ranked results for each algorithm are relevant results. This analysis required either reading the encyclopedia entries in full, or extensive scanning for long entries.

For practical reasons, this preliminary testing was performed with one evaluator. Due to the experimental, non-formalizable, and subjective nature of evaluating topic relevance of a particular sample for a specific user community (Sebastiani 2002 as cited in Golub et al. 2016), the problem of intra-indexer inconsistency in evaluating topic relevance is well-documented (Olson and Wolfram 2008). We subsequently determined that the addition of multiple indexers was not practical or necessary for this phase of relevance evaluation.

For each of the three algorithm results, the evaluator indicated whether each subject heading result was a relevant (R), partially relevant (PR), or non-relevant (NR) result. Some researchers have implemented the measure of partial relevance as a means to indicate user uncertainty regarding the information object's degree of relevance, but it may also be used to signify the degree of relevance in relation to an information goal (Hjørland and Christensen 2002).These measures were selected to account for a few factors specific to indexing this unique historical and multidisciplinary corpus: (1) Early results with Maui and Kea++ made it clear that the obvious non-relevant results were far easier to identify. This was due to the inability of the algorithms to disambiguate between different semantic meanings of a word (e.g., 11th edition entry on Rum LCSH HIVE results: Rummy (Game) and Spirits (Islam)), as well as the inclusion of time-inappropriate subject headings (e.g., 11th edition entry on Rifle, LCSH HIVE result: ZSU-23-4 (Antiaircraft gun)); (2) The inherent subjectivity of evaluating encyclopedia entry aboutness, and the level of certainty with which an evaluator can make an absolute judgement based on reading or scanning the entries; 3) The level of subject heading granularity, which may leave the evaluator feeling that while a particular subject heading may not be obviously incorrect, the low level of granularity renders the result only partially relevant for the digital humanities user community (e.g., 7th edition entry on Raleigh (Sir Walter) LCSH HIVE result: Time). These examples demonstrate that while it could be obvious to the evaluator that the different semantic meanings and time-inappropriate terms render results non-relevant, the relevance and usefulness of the broad results may depend on the user community and the context in which it is used throughout the entry.

Results

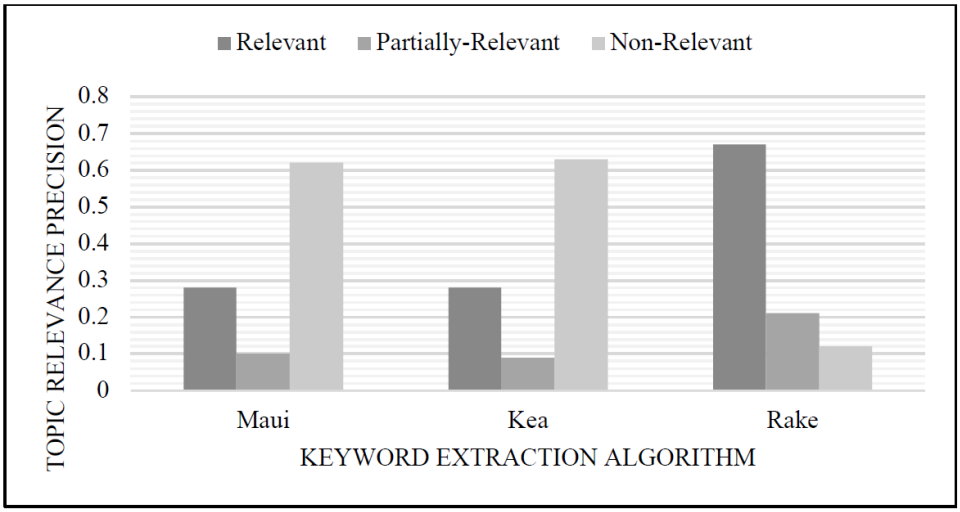
For each of the 10 sample entries processed with HIVE, a preset number of 10 subject headings were returned, for a total of 100 subject headings for each of the three algorithms (300 total).

Table 2: Topic relevance precision scores for three algorithms tested in HIVE
* Precision score percentages out of the 100 total subject headings for each algorithm

Relevance Measure*	Maui	Kea++	RAKE
Relevant	28%	28%	67%
Partially Relevant	10%	9%	21%
Non-Relevant	62%	63%	12%
HIVE top-ranked as true relevant	0%	100%	100%

These results (Table 2) show a clear indication of RAKE's higher topical relevance precision, as well as improved performance with HIVE in assessing the highest relevance-ranked result. The bar graph (Figure 1) visualizes these precision scores to demonstrate the significant increase in precision, with 67% of RAKE's HIVE results being relevant results. Maui and Kea performed at almost identical precision rates, with the one exception that Kea++ performed better than Maui when the evaluator judged the highest relevant result against the source material.

Figure 1: Topic relevance precision scores for three algorithms tested in HIVE



Discussion

The topic relevance evaluation above presented a comparison between keyword extraction algorithms integrated into the HIVE controlled vocabulary server and automatic indexing application. This research demonstrated a 39% increase in true relevant results when indexing with the RAKE keyword extraction algorithm, when compared to the results obtained with Maui and Kea++. The results also demonstrated that the highest-ranked results with RAKE and Kea++ were more likely to be relevant results than with Maui. While none of Maui's highest-ranked results were relevant results, all of Kea++ and RAKE's highest-ranked results were relevant results.

This study identified several common indexing errors that determined relevance ranking. These indexing errors reflect those identified by Information Retrieval researchers (Table 3, rows 1-4) (Lancaster 2003 as cited in Golub et al. 2016), but also reflect the complexities of using a controlled vocabulary, as well as the historical nature of this corpus (Table 3, rows 5-6):

Table 3: Examples of common indexing errors as found in the research on this corpus *Italics indicate comments added by researchers for clarity

INDEXING ERROR	EXAMPLE
1. Too many or too few results	• HIVE Version(s): 1 • Algorithm(s): Maui • Entry: 3rd edition, "Rhetoric" • Results: Zero results
2. Inappropriate levels of granularity (too broad)	• HIVE Version(s): Both • Algorithm(s): All • Entry: 11th edition, "Rameses" *(the city)** • Broad Result: "Names"
3. The absence of essential subjects	• HIVE Version(s): 1 • Algorithm(s): Maui & Kea++ • Entry: 9th edition, "Rice" • Missing Result: "Rice"
4. Presence of obviously incorrect subject headings	• HIVE Version(s): 2 • Algorithm(s): Kea • Entry: 11th edition, "Rose" *(the flower)** • LCSH Result: Heterosexual teachers
5. Different semantic meanings of a word	• HIVE Version(s): 1 • Algorithm(s): Maui • Entry: 11 edition, "Rum" *(the liquor)** • Results: "Rummy (Game)," "Spirits (Islam)"
6. Time-Inappropriate Subject Heading	• HIVE Version(s): 2 • Algorithm: RAKE • Entry: 11th edition, "Rifle" • Result: "ZSU-23-4 (Antiaircraft gun)"

As discussed by researchers, these errors may be the result of controlled vs. uncontrolled vocabularies, the level of specificity in the vocabulary, the linguistic characteristics of the text and subject matter, limitations of the available indexing tool, and the length of the text (Lancaster 2003 as cited in Golub et al. 2016). The most pervasive indexing errors were inappropriate specificity levels, and the presence of obviously-incorrect subject headings. Most of the obviously-incorrect subject headings were a result of the inability of automatic indexing tools to disambiguate between different meanings of a word.

Early HIVE 1 results with the Maui and Kea++ algorithms identified additional challenges inherent in indexing 19th-century documents with current-day controlled vocabularies, such as the misidentification of computer-related subject headings (e.g., Using Maui to index the 3rd and 7th edition entries on the painter Raphael returned the subject heading "Leo computer"). Subsequent testing with RAKE did not return as many time-inappropriate subject headings, but the example in Table 3, row 6 demonstrates that the issue still appears with RAKE throughout our testing on this corpus.

Due to the practical limitations of this experimental and necessarily iterative research, this first-phase relevance study was designed with one evaluator. However, the prevalence of obviously-incorrect subject headings rendered multiple evaluators unnecessary for this phase of the project. Since the evaluator read each entry before ranking topical relevance, this study was performed on a small sample, to account for large variations in encyclopedia entry length. This may affect the relevance scores reported in this research, but due to the pervasiveness of obviously-incorrect subject headings when indexing with Maui and Kea++, metadata researchers were confident in the validity of the small sample results.

It is important to point out that HIVE 1 and HIVE 2 work with different versions of LCSH (HIVE 1, 2014; HIVE 2, 2018). Given practical research constraints of time, we still proceeded with this work, recognizing that this limitation could have an impact on the results. Because we were only looking at the top-ten results, the research still gives enough insight into the algorithm's performance, although we recognize that further testing is needed.

An additional complication of the research was the presence of the historical Long S < ſ > in the 3rd edition of the Encyclopedia Britannica. Correction of the Long S in the OCR output is in the process of optimization, and further research is needed to determine whether or not the Long S affects the relevance scores when indexing with HIVE.

Conclusion and Next Steps

The objective of this research was to determine which of three keyword extraction algorithms integrated into the HIVE tool produces a greater proportion of relevant results when indexing entries across four 19th-century Encyclopedia Britannica editions. This was accomplished through a comparative topic relevance evaluation. The results presented above show that the RAKE keyword extraction algorithm provides significantly increased proportions of relevant results when automatically indexing historical Encyclopedia Britannica entries with the HIVE application. These preliminary results are essential for moving forward with automatic metadata generation for the 19th-Century Knowledge Project.

These results are specific to this particular collection of documents, though we believe that this research may have wider applicability to other historical multidisciplinary knowledge corpora. Further research is needed to determine wider applicability. Next steps include performing comparative topic relevance testing to refine RAKE's minimum word frequency parameter to accommodate for entries of varying lengths, as well as performing comparative topical relevance testing for before and after correction of the historical Long S in the 3rd edition. Additional research is also necessary to determine whether natural language keyword extraction should supplement these controlled vocabulary results, and whether the inclusion of named entity controlled vocabularies would improve results for automatically indexing this corpus.

Acknowledgements

We acknowledge the support of the NEH grant #HAA-261228-18 and IMSL grant #RE- 70-17-0094-17.

References

Bair, S. and S. Carlson. 2008. Where Keywords Fail : Using Metadata to Facilitate Digital Humanities Scholarship Where Keywords Fail : Using Metadata to Facilitate Digital Humanities Scholarship. *Journal of Library Metadata* 8(3): 249-262. https://doi.org/10.1080/19386380802398503

Bueno-de-la-Fuente, G., D.R. Mateos, and J. Greenberg. 2016. Chapter 10
- Automatic Text Indexing with SKOS Vocabularies in HIVE. In , 231-245: Elsevier Ltd.

Cosijn, E. 2010. Relevance Judgments and Measurements. *Encyclopedia of Library and Information Sciences*, Third Edition, 4512–19. https://doi.org/10.1081/e-elis3-120044537.

"Dryad Digital Repository." Dryad. Accessed May 2, 2019. http://www.datadryad.org/.

Golub, K., D. Soergel, G. Buchanan, D. Tudhope, M. Lykke, and D. Hiom. 2016. A Framework for Evaluating Automatic Indexing or Classification in the Context of Retrieval. *Journal of the Association for Information Science and Technology* 67(1): 3-16. https://doi.org/10.1002/asi

Hjørland, Birger. 2001. Towards a Theory of Aboutness, Subject, Topicality, Theme, Domain, Field, Content . . . and Relevance. *Journal of the American Society for Information Science and Technology* 52(9): 774-778.

Hjørland, B. and F.S. Christensen. 2002. Work Tasks and socio-cognitive Relevance: A Specific Example. *Journal of the American Society for Information Science and Technology* 53(11): 960-965.

Lancaster, F.W. 2003. *Indexing and Abstracting in Theory and Practice*. Third ed. London: Facet Pub.

Logan, P. Nineteenth-Century Knowledge Project. *Nineteenth-Century Knowledge Project*. Accessed April 3, 2019. https://tu-plogan.github.io/.

Medelyan, O., V. Perrone, and I. Witten. 2010. Subject Metadata Support Powered by Maui. *ACM*. doi:10.1145/1816123.1816204.

Medelyan, O. and I.H. Witten. 2008. Domain-independent Automatic Keyphrase Indexing with Small Training Sets. *Journal of the American Society for Information Science and Technology* 59(7): 1026-1040.

Moens, M. 2000. *Automatic Indexing and Abstracting of Document Texts*. 1. Aufl. ed. Vol. 6.;6;. Boston: Kluwer Academic Publishers. doi:10.1007/b116177.

Olson, H.A. and D. Wolfram. 2008. Syntagmatic Relationships and Indexing Consistency on a Larger Scale. *Journal of Documentation* 64(4): 602-615.

Rose, S., D. Engel, N. Cramer, and W. Cowley. 2010. Automatic Keyword Extraction from Individual Documents. In , 1-20. Chichester, UK: John Wiley & Sons, Ltd.

Saracevic, T. 2007. Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part III: Behavior and Effects of Relevance. *Journal of the American Society for Information Science and Technology* 58(13): 2126-2144.

Svenonius, Elaine. 2000. *The Intellectual Foundation of Information Organization*.
Cambridge, Mass: MIT Press.

Walsh, John. 2011. The use of Library of Congress Subject Headings in Digital Collections. *Library Review* 60(4): 328-343.