Keywords, Indexing, Text Analysis: An Editorial[†]

Richard P. Smiraglia

† The editor would like to gratefully acknowledge Andrea Scharnhorst, colleague at the e-Humanities Group, Royal Netherlands Academy of the Arts and Sciences, Amsterdam, for referring him to Hermeneutica.ca's *Voyeur* text analysis tool.

Recently I was asked in earnest why KO doesn't have keywords. To which my reply was to LOL. Really—I laughed, out loud, and then I said "but it does, in every line!"

1.0 A rant

It took awhile for the real issue to settle into my brain so I could give a cogent response. Of course, what my inquisitioner was asking was why we don't have a list of author-contrived "keywords" underneath the abstract like so many other journals do. I was chagrined to realize the practice has become so ubiquitous that people entering the discipline think it is normal. Indeed, if one googles "keywords" one finds sets of instructions for phonying up the proper list of keywords on manuscripts such that one can somehow effect future retrieval of the article online. But of course, this is all based on assumptions about a) keywords; b) use of those lists of keywords; c) the role of indexing; and d) the proper functioning of information retrieval (IR).

So first let me say that one reason there are no lists of author-contrived keywords in *Knowledge Organization* is that when I became editor they were not being used. Although I have added review and processing dates ("received, revised, accepted") to encourage submission of manuscripts—potential authors can see that most papers submitted to this journal that get published, do so within six months, and that is pretty fast in the world of information journals (some of which take years from submission to publication). But as you can tell from my choice of words, I do not think lists of author-contrived keywords are useful. I do not decide whether to read an article in a journal based on those lists. I make my decision based on actual keywords—the ones in the title—and then I read the abstract to see whether I think the article is either of interest to me or of use to my research. And I thought I knew that indexing services did not use those lists either. The entire use of them seems to stem from a misperception that somehow adding "weighted" terms to the printed page in a journal would improve retrieval using indexing databases. The fact is, the only thing that improves retrieval is formal indexing. We have managed to get both Thompson Reuters and EBSCOHost to index our journal, and in the case of EBSCOHost to make the full-text available online through library subscription portals. That indexing is what will affect the rate at which articles published in our journal are discovered, read, ingested, and cited.

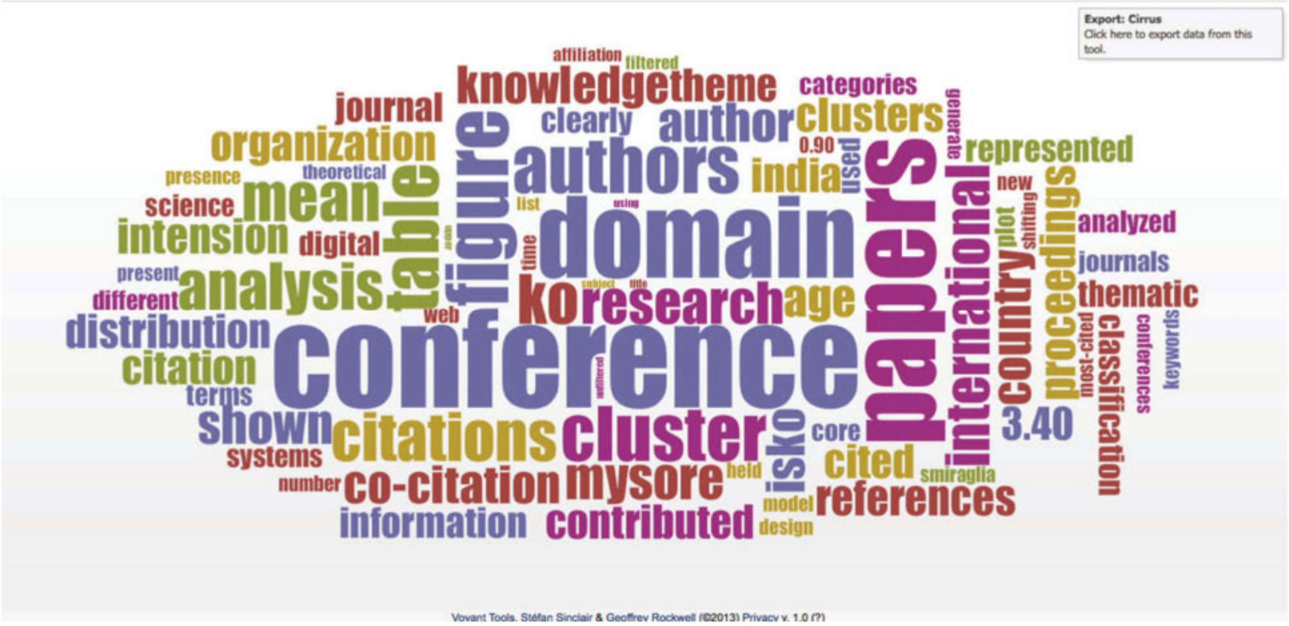
But there is more, of course, to my objection to keywords, and most of it stems from what I perceive to be a naïve understanding of information retrieval. Of course, information retrieval relies on keywords. But it relies on their presence in actual text, and in proximity to other terms (or, keywords). The reality is that actual keywords are everywhere in any journal, ours included.

2.0 A case study

I decided to undertake a little editorial experiment by using the contents of the last two issues of *Knowledge Organization*. Volume 40 (2013) number 1 contained an editorial, 4 peer-reviewed articles, a book review, a classification issues report, and two substantive letters to the editor. Volume 40 (2013) number 2 contained 5 peer-reviewed articles, some ISKO news, and a bibliographic essay book review; unfortunately at the time this was written number 2 had not been indexed by either service. I decided to compare keywords drawn from Thompson Reuters' Web of Science™ and EBSCOHost's *Library and Information Science and Technology Abstracts with Full Text* (*LISTA*) to the actual keywords pulled from the texts. Full texts were uploaded to *Voyeur* from Hermeneutica.ca—*The Rhetoric of Text Analysis* (http://hermeneuti.ca/voyeur/) to derive most frequently used terms (applying an English language stoplist). Table 1 contains those comparative results.

	Web Of Science keywords-plus	***LISTA* with full text (EBSCO)**	***Voyeur*** — Most frequent words
Smiraglia—ISKO 12's Bookshelf—Evolving Intension: An Editorial	none	knowledge management — congresses; information technology; conferences & conventions; international society for knowledge organization (organization) — congresses; mysore (india : state); india	**744** unique words conference (29), papers (23), domain (21), figure (17), authors (14)
Hjørland— User-based and cognitive approaches to knowledge organization: a theoretical analysis of the research literature	information-science; critique; behavior	knowledge management; library science; information science; information technology; subjectivity; iphone (smartphone)	**2679** unique words information (139), cognitive (105), knowledge (65), studies (60), science (59).
Corrochano et al.— Spanish Research in Knowledge Organization (2002-2010)	none	knowledge management; bibliometrics; information storage & retrieval systems; databases; globalization	**1580** unique words authors (71), knowledge (59), organization (46), table (34), research (32)
Tennis—Ethos and Ideology of Knowledge Organization: Toward Precepts for an Engaged Knowledge Organization	none	knowledge management; metadata; buddhism; critical theory; ideology; language & languages	**1280** unique words knowledge (73), organization (57), action (49), work (48), violence (46).
Almeida Campos et al.—Information Sciences Methodological Aspects Applied to Ontology Reuse Tools: A Study Based on Genomic Annotations in the Domain of Trypanosomatides	knowledge organization; systems	information science; bioinformatics; qualitative research; trypanosomatidae; ontology; biomedical materials	**1716** unique words ontology (113), ontologies (82), information (44), terms (43), reuse (42)
Channon— The Unification of Concept Representations: An Impetus for Scientific Epistemology	none		**3036** unique words science (82), time (81), phenomena (77), event (62), schematic (57).
Martinez-Avila and San Segundo— Reader-interest classification concept and terminology historical overview	none		**2132** unique words classification (138), library (123), reader-interest (114), libraries (69), public (66)
Marcondes— Knowledge Organization and Representation In Digital Environments: Relations Between Ontology and Knowledge Organization	none		**1516** unique words ontology (51), knowledge (47), information (41), classification (36), domain (36)
Oikarinen and Kortelainen—Challenges of Diversity, Consistency and Globalty in Indexing of Local Archeological Artifacts	none		**1978** unique words archeological (85), artifacts (79), subnumbers (56), knowledge (48), cataloguing (45)
Sienkiewicz and Kijenska-Dabrowski— Knowledge creation and commercialization activities in Polish public HEUs in the area of technical and engineering sciences	none		**1174** unique words research (61), 00 (55), number (42), activity (34), publications (34).
	Web Of Science keywords-plus	***LISTA* with full text (EBSCO)**	***Voyeur*** — Most frequent words
Smiraglia—ISKO 12's Bookshelf—Evolving Intension: An Editorial	none	knowledge management — congresses; information technology; conferences & conventions; international society for knowledge organization (organization) — congresses; mysore (india : state); india	conference (29), papers (23), domain (21), figure (17), authors (14)
Hjørland— User-based and cognitive approaches to knowledge organization: a theoretical analysis of the research literature	information-science; critique; behavior	knowledge management; library science; information science; information technology; subjectivity; iphone (smartphone)	information (139), cognitive (105), knowledge (65), studies (60), science (59).
Corrochano et al.— Spanish Research in Knowledge Organization (2002-2010)	none	knowledge management; bibliometrics; information storage & retrieval systems; databases; globalization	authors (71), knowledge (59), organization (46), table (34), research (32)
Tennis—Ethos and Ideology of Knowledge Organization: Toward Precepts for an Engaged Knowledge Organization	none	knowledge management; metadata; buddhism; critical theory; ideology; language & languages	knowledge (73), organization (57), action (49), work (48), violence (46).
Almeida Campos et al.—Information Sciences Methodological Aspects Applied to Ontology Reuse Tools: A Study Based on Genomic Annotations in the Domain of Trypanosomatides	knowledge organization; systems	information science; bioinformatics; qualitative research; trypanosomatidae; ontology; biomedical materials	ontology (113), ontologies (82), information (44), terms (43), reuse (42)

Table 1. Indexing of contents of *Knowledge Organization* v. 40 nos. 1-2 (2013)



Export: Cirrus
Click here to export data from this tool.

Voyant Tools, Stéfan Sinclair & Geoffrey Rockwell (©2013) Privacy v. 1.0 (?)

Figure 1. Hermeneutic.ca's *Voyeur* word cloud for Smiraglia ISKO 12's Bookshelf

The results are a bit disturbing. "Keywords" were added only three times in the Web of Science indexing of the five papers from vol. 40 no. 1. More unsettling is the terminology used in the subject "terms" (really subject headings) assigned by *LISTA*—note that they have represented "knowledge organization" in every case as "knowledge management." In Hjørland's paper about knowledge organization theory, the outdated and inaccurate term "library science" has been applied (Hjørland does not use the term in his text—it appears in two citations to Ranganathan). This is inaccurate and misleading at best, and dangerous for our domain at worst, because it clearly misleads searchers and ultimately prevents ingestion and citation of our research.

Hermeneutica's text-analytical tool is impressive and potentially very powerful, providing not just a word count for each article but also a count of the number of unique words in each. Stoplists may be applied to full texts, and the analysis provides a colorful word cloud that illustrates the most-used terms in the text. Clicking on any term in the cloud generates a frequency graph about the use of the term and a keyword in context (KWIC) map of the text allowing visualization of usage. Figure 1 is a screen capture of the word cloud for the Smiraglia editorial.

Underneath the word cloud *Voyeur* gives a summary of the number of unique words, and of the most frequently used words (with the occurrence totals). Below that a frequency distribution of terms is available. Clicking on any highlighted term in the summary or in the frequency distribution will generate the word trends analysis graph and KWIC display. Various bits of data may be exported as well. In Table 1 the most frequently used words from each text are given together with the occurrence totals. These are the real keywords from these papers. Just to take the experiment one step further, we compare these keywords to the WoS and *LISTA* terms in Table 2.

Only the papers from vol. 40 no. 1 are included in Table 2, of course. What is immediately obvious is how little correspondence there is across the three; yellow highlighting shows terms that occur in more than one source. Some of the frequently-used words are, in fact, title keywords in every paper but the editorial. But the frequently-used terms are the most accurate and descriptive in every case. An interesting question arises, which is whether authors fail to use important terms frequently enough in their texts to cause them to fall into an empirically extracted list of frequently-used terms. For example, the term "evolving intension" is used in the title of the editorial in this case study, but that term does not appear in the most frequently used terms list. In such cases, when authors name important concepts but then describe them in text with more specific terms, the important key terms might fail to be extracted.

As a final step I entered the titles of the ten papers into WordStat™ and generated a co-word analysis as a visualization of keywords in this small group of papers. Figures 2 and 3 show the dendrogram and three-dimensional MDS (Multi-dimensional Scaling) plot that result.

This plot is a fair visualization of the small corner of the domain represented by these ten papers. (No goodness of fit statistics are given here because there really are too few cases involved in this "case" study.) The central role of the "concept" is clear, as is the leading position of "ontology" and the importance of "science" and "research." The group is anchored by "knowledge organization."

	Web Of Science keywords-plus	***LISTA* with full text (EBSCO)**	***Voyeur*—** Most frequent words
Smiraglia—ISKO 12's Bookshelf—Evolving Intension: An Editorial	none	knowledge management — congresses; information technology; conferences & conventions; international society for knowledge organization (organization) — congresses; mysore (india : state); india	conference (29), papers (23), domain (21), figure (17), authors (14)
Hjørland— User-based and cognitive approaches to knowledge organization: a theoretical analysis of the research literature	information-science; critique; behavior	knowledge management; library science; information science; information technology; subjectivity; iphone (smartphone)	information (139), cognitive (105), knowledge (65), studies (60), science (59).
Corrochano et al.— Spanish Research in Knowledge Organization (2002-2010)	none	knowledge management; bibliometrics; information storage & retrieval systems; databases; globalization	authors (71), knowledge (59), organization (46), table (34), research (32)
Tennis—Ethos and Ideology of Knowledge Organization: Toward Precepts for an Engaged Knowledge Organization	none	knowledge management; metadata; buddhism; critical theory; ideology; language & languages	knowledge (73), organization (57), action (49), work (48), violence (46).
Almeida Campos et al.—Information Sciences Methodological Aspects Applied to Ontology Reuse Tools: A Study Based on Genomic Annotations in the Domain of Trypanosomatides	knowledge organization; systems	information science; bioinformatics; qualitative research; trypanosomatidae; ontology; biomedical materials	ontology (113), ontologies (82), information (44), terms (43), reuse (42)

Table 2. Comparison of *LISTA* key terms and *Voyeur*-derived keywords

3.0 Some concluding thoughts

The role of what we call keywords in scholarly discourse has increased to the point that authors add them to manuscripts submitted to *Knowledge Organization* even though we do not ask for them (and delete them in editing). The actual use of keywords is unclear; I doubt readers use them much but it is possible that indexers rely on them. Perhaps that is why the formal indexing in this case study is so problematic. The potential use of keywords for retrieval and indexing seems clear. That is, the presence of keywords, whether in a separate list or in their usual place in the text, has the potential to influence the formal indexing of research, and also to influence resource-location or selection by researchers.

What is less clear is how those keywords should be generated. Empirical extraction of the terms is most accurate and therefore most reliable for indexing, retrieval or just for text analysis. Should editorial policy change to incorporate the use of formal keywords in *Knowledge Organization* it would make the best sense to generate the terms empirically, using text analysis tools designed for statistical term extraction.

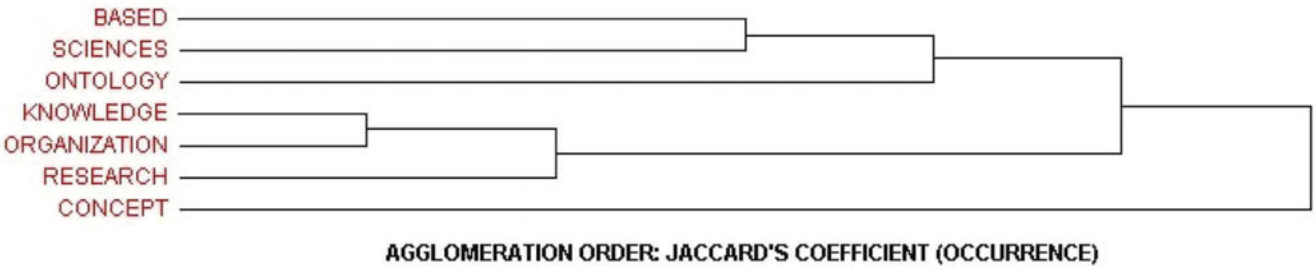


Figure 2. Title keyword co-occurrence dendrogram

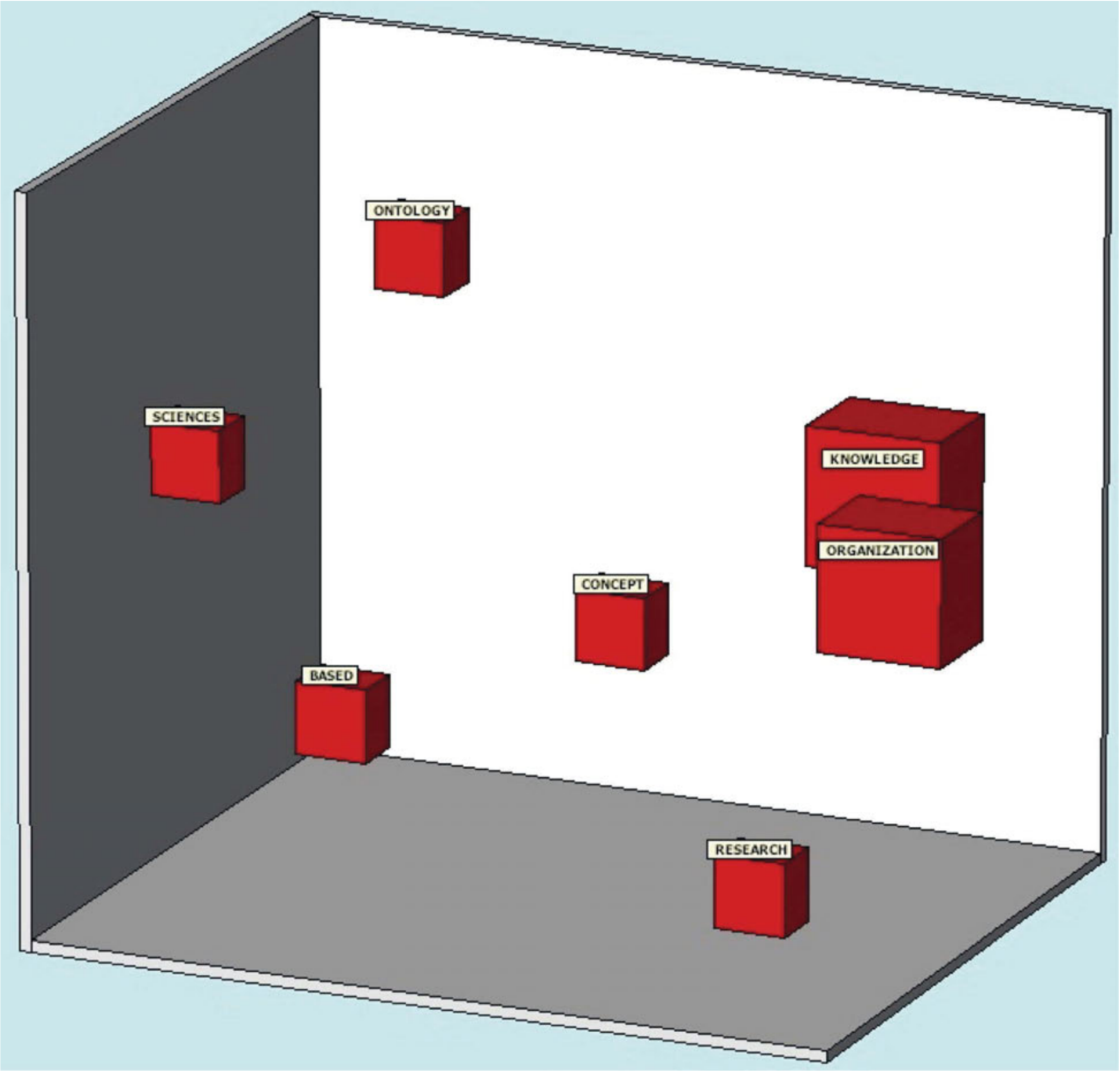


Figure 3. Title keyword co-occurrence three-dimensional MDS plot