Keywords Redux—An Editorial*

Richard P. Smiraglia

Smiraglia, Richard P. **Keywords Redux—An Editorial.** *Knowledge Organization. 42(1)*, 3-7. 3 references.

* The data reported in this editorial were gathered by Hyoungjoo Park, to whom I am most grateful.

In *KO* volume 40 number 3 (2013) I included an editorial about keywords—both about the absence prior to that date of designated keywords in articles in *Knowledge Organization*, and about the misuse of the idea by some other journal publications (Smiraglia 2013). At the time I was chagrined to discover how little correlation there was across the formal indexing of a small set of papers from our journal, and especially to see how little correspondence there was between actual keywords appearing in the published texts, and any of the indexing supplied by either *Web of Science* or *LISTA* (Thomson Reuters' *Web of Science*™ (*WoS)* and EBSCOHost's *Library and Information Science and Technology Abstracts with Full Text* (*LISTA*). The idea of a keyword arose in the early days of automated indexing, when it was discovered that using terms that actually occurred in full texts (or, in the earliest days, in titles and abstracts) as search "keys," usually in Boolean combinations, provided fairly precise recall in small, contextually confined text corpora. A recent *Wikipedia* entry (Keywords 2015) embues keywords with properties of structural reasoning, but notes that they are "key" among the most frequently occurring terms in a text corpus. The jury is still out on whether keyword retrieval is better than indexing with subject headings, but in general, keyword searches in large, unstructured text corpora (which is what we have today) are imprecise and result in large recall sets with many irrelevant hits (see the recent analysis by Gross, Taylor and Joudrey (2014). Thus it seems inadvisable to me, as editor, especially of a journal on knowledge organization, to facilitate imprecise indexing of our journal's content.

Nevertheless, during 2014, in *Knowledge Organization* volume 41, I added lists of keywords below the abstracts of each article. I deliberately did not ask authors to supply keywords. Rather, I developed a set for each article by entering the complete text into the *Voyeur* tool from Hermeneutica.ca (http://hermeneuti.ca/voyeur/). I did not use the simple list of most frequently occurring words provided by *Voyeur*; rather I used that list to inform my compilation of terms (preferring "knowledge organization" as a term, to the keywords "knowledge" and "organization," for example). In the rare event that an important concept represented in the title of an article did not fall into this list, I added terms to represent that concept as well, after first checking to be certain the term actually occurred in the text of the article. In this manner, *Knowledge Organization* managed to complete a full year of publication with "keywords." I was amused on occasion, when authors asked to replace our keywords with terms of their own, none of which actually occurred anywhere in their texts. I decided to place a notice in the instructions for manuscript submission indicating that author keywords are not used, but that our keywords are supplied in conjunction with *Voyeur* text analysis.

2.0 Case study redux

As the end of the volume year approached I decided to revisit my earlier case study to see whether the presence of keywords in our publication had had any effect on indexing. To do so, I gathered comparative data using the first article in each number of volume 41—admittedly a convenience sample with no potential for generalizability. Table 1 shows the results of that comparison, giving the keywords from *KO,* alongside indexing discovered in either *WoS* or *LISTA*.

LISTA has not yet indexed much of the volume as of January 2015); the final issue has not yet been indexed by either *WoS* or *SCOPUS*. *SCOPUS* uses the supplied keywords, which, like the others, it identifies as "author keywords." In two cases *WoS* did not add to the keywords that were already present, but in two cases they did add "keywords plus." Gratifyingly, "knowledge organization" is one of those. It would seem we have convinced *WoS*

Knowledge Organization issue	Article	*KO* supplied keywords	*WoS* Keywords Plus	*LISTA* subject headings
41n1 2014	Olesun-Bagneux "The Memory Library"	library, memory, literature, Pinakes, Aristophanes, mechanics, scholars, Alexandria		library science; information science; libraries; classification of books; information retrieval; Alexandria (Egypt); Egypt
41n2 2014	Ménard and Dorey "TIIARRA"	images, participants, TIARRA, retrieval, taxonomy, categories, indexing, efficiency	American history; retrieval	taxonomy, digital images, websites, bilingualism, statistics, nglish language
41n3 2014	Satija, Madalli and Dutta "Modes of Growth of Subjects"	knowledge, subjects, growth, Ranganathan	knowledge; classification; systems	not in LISTA
41n4 2014	Szostak "Classifying the Humanities"	art, classification, classifying, works, scholarship, humanities, concepts, subjects	knowledge organization; information-seeking	not in LISTA
41n5 2014	Shiri "Making Sense of Big Data"	Big Data, facet analysis, research, metadata		not in LISTA
41n6 2014	Budd "Organizing Acts and Objects"	metaphysics, knowledge organization, knowledge, objects, acts, informing, information	not in WoS	not in LISTA

Table 1. Comparative Keywords from *KO*, *WoS*, and *LISTA*.

indexers that the term "knowledge organization" is something we take seriously. At least *LISTA* did not apply knowledge management terms, as was the case in the prior review, although the outdated early twentieth century term "library science" appears. (For an informative rant on that term see my blog post "Grocery Store Science" at https://lazykoblog.wordpress.com/2011/01/25/grocery-store-science/.)

3.0 The use of keywords in our closest "sibling" journals

Curious to discover whether the journals closest to us intellectually use keywords, and if so, how, I decided to compare keywords for two years in *Journal of Documentation* and *Journal of the Association for Information Science and Technology*. We collected data for the first article in each issue of each journal in 2013 and 2014. Again this is purely a convenience sample for the purpose of generating basic information, thus I will not report quantitative results.

The guidelines from *Journal of Documentation* (*JDoc*) ask authors to provide keywords, but make it clear editorial prerogatives will be enforced (http://www.emeraldgrouppublishing.com/products/journals/author_guidelines.htm?id=jd):

> Please provide up to 10 keywords on the Article Title Page, which encapsulate the principal topics of the paper …. Whilst we will endeavour to use submitted keywords in the published version, all keywords are subject to approval by Emerald's in house editorial team and may be replaced by a matching term to ensure consistency.

In *JDoc* keywords appear at the bottom of the abstract with the header "keywords." *Journal of the Association for Information Science and Technology* (*JASIST*) does not mention keywords in its instructions for authors and no keywords appear in the texts of the articles. However, a list of keywords accompanies the article citation and abstract in the ASIST Digital Library online. Those same key-

words are supplied in the *LISTA* indexing for the article, where they are identified as "author-supplied keywords."

To discover the degree to which the keywords match frequently-occurring terms in the articles, we analyzed each text using *Voyeur*. Table 2 contains comparative data showing the keywords adjacent to the truncated term frequency maps for each article drawn from *JDo*c.

I think results in Table 2 are mixed. In most cases keywords from the title recur among the author-supplied keywords as well as among the most frequently-occurring terms. But there are some cases in which there is less correspondence. For example, Pattuelli and Rubinow's paper has the terms "knowledge organization" and "case study" in its title, and neither term appears on either list. "Semantic web" occurs among the author keywords but is not among the most frequently-occurring terms (according to our analysis the word "semantic" does occur, but with very low frequency). Matthew's paper title features "fixation" and "accretion" but neither word occurs among the keywords or the frequently-occurring terms. "Open access" is an important term in the title and keywords of the paper by Spezi et al. but is not among the most frequently-occurring terms. Another observation is that a number of frequently-occurring terms do not appear in either the title or among the keywords. Examples of this are "users" and "analytic" in Goodale and Clough, "Deleuze" in Faucher, and "factors" and "needs" in Robson and Robinson.

Table 3 shows similar results from *JASIST*:

Results are similarly mixed. Klavans et al., Lindenthal and Losee list keywords "collaboration," "trademarks" and "probabilistic" respectively, none of which occurs in either the title or the term list. Eschenfelder and Johnson name a "data commons" and use the words "scholarly" and "sharing" in their title, but none of those is directly indicated among the keywords or in the term list. Similarly, White's "belief dynamics," O'Brien and Lebow's "news interactions," Losee's "document ordering," and Velden's "ethnographic observations"—all title words—are concepts that do not recur in either the keywords or the term lists.

In all of the cases from either journal where we see some specific divergence it could be asserted that the specific terms to which I've drawn attention are subsumed under terms representing hierarchically broader concepts. Traditionally, however, acknowledging hierarchical relations has been the province of controlled vocabularies, and not of keyword searches. In every case, 7 from *JASIST* and 6 from *JDoc*, we have examples where terms that could be said to denote "keyness" as one potential measure, or increased specificity as another, are used in titles but not in the indexing. The inconsistency, if we want to consider it that, or alternatively, the failure of precision to co-occur across the indexing occurs in roughly a third of the potential instances among these journal articles.

Collectively the results tell us that the correspondence between title keywords, terms frequently occurring in texts (which are actual keywords), and keywords supplied as indexing is incomplete although not dramatically so. The real hazard for retrieval is that quite precise terms that occur in papers are not being used in the indexing.

For the moment it seems clear that keywords, if used with journal articles, are good editorial practice only when they represent either keyness or specificity or both. Thus it seems the most appropriate editorial stance continues to be edited lists. This will remain our editorial policy through 2015.

References

Gross, Tina, Arlene G. Taylor and Daniel N. Joudrey. 2015. "Still a Lot to Lose: The Role of Controlled Vocabulary in Keyword Searching." *Cataloging & Classification Quarterly* 53 no. 1: 1-39.

"Keywords." 2015. *Wikipedia. http://en.wikipedia.org/wiki/Keywords* Accessed 23 February 2015.

Smiraglia, Richard P. 2014. "Keywords, Indexing, Text Analysis—An Editorial." Knowledge Organization 40: 155-59.

Issues	Authors	Titles	Keywords	Most frequently used terms
70n6, 2014	Goodale and Clough	Cognitive styles within an exploratory search system for digital libraries	Attitudes, Digital libraries, Individual differences, Cognitive style, Information behaviour, Educational informatics	users, analytic, wholist, paths, cognitive, style, task, search, path, information
70n5, 2014	Matthews	Knowledge fixation and accretion: longitudinal analysis of a social question-answering site	Communities, Individual behaviour, Knowledge processes, CQA, Q&A	Answer(s), question(s), quality, time, community, new, information, knowledge
70n4, 2014	Faucher	An information meta-state approach to documentation	Information theory, Becoming, Metastability, Transduction, Sense	information, document, process, systems, Deleuze, way, new, documentation, terms, view
70n3, 2014	White, Willis and Greenberg	HIVEing: the effect of a semantic web technology on inter-indexer consistency	Inter-indexer consistency, Indexing, Helping Interdisciplinary Vocabulary Engineering (HIVE)	indexing, terms, hive, relevant, consistency, inter-indexer, marked, results, keywords, vocabularies
70n2, 2014	Finnemann	Research libraries and the internet : On the transformative dynamic between institutions and digital media	Digitization, Research libraries, Digital media, Knowledge production Paper type Research paper	media, digital, new, internet, knowledge, research, libraries, institutions, computer, materials
70n1, 2014	Alexander	Devising a framework for assessing the subjectivity and objectivity of information taxonomy projects	Subjectivity, Classification, Information science, Taxonomies, Knowledge organisation, Sociology of science, Epistemology, Organisational politics, Organisational culture	projects, taxonomy, project, criticism, public, e.g, organisation, process, framework, criteria
69n1, 2013	Bawden	Knowledge, documentation and a London location	United Kingdom, Knowledge management	London, account, Bloomsbury, library, Ashton, book, intellectual, British, history, information
69n1, 2013	Robson and Robinson	Building on models of information behaviour: linking information seeking and communication	Information behaviour, Communications, Theory, Models, Information	information, model, behaviour, seeking, models, communication, factors, figure, needs, sources
69n3, 2013	Spezi, Fry, Creaser, Probets and White	Researchers' green open access practice: a cross - disciplinary analysis	Open access, Repositories, Self - archiving, Behaviour, Attitudes, PEER project, Disciplinary differences, Research work	sciences, researchers, version, article, likely, research, journal, repository, respondents, articles
69n4, 2013	Lingard	Information, truth and meaning: a response to Budd's prolegomena	Information, Sense - making, Critical realism, Ontology, Information - seeking behaviour, Abduction, Meaning - making, Case studies, Philosophical concepts	information, truth, meaning, properties, sense-making, Budd('s), definition, action, human
69n5, 2013	Gobinda	Sustainability of digital information services	Sustainable development, Digital information services, Sustainability, Social sustainability, Economic sustainability, Environmental sustainability	information, digital, sustainability, services, research, social, access, economic, sustainable, use
69n6, 2013	Pattuelli and Rubinow	The knowledge organization of DBpedia: a case study	DBpedia, Linked open data, Semantic web	Dbpedia, data, Wikipedia, properties, infobox, knowledge, ontology, jazz, templates

Table 2. Journal of Documentation 2013-2014 keywords and most frequently-occurring terms.

Issues	Authors	Titles	Keywords	Most frequently used terms
65n1, 2014	Klavans, LaPlante and Golbeck	Subject matter categorization of tags applied to digital images from art museums	collaboration; museums	tags, images, image, search, terms, tag, collection, subject, art, users
65n2, 2014	Björk, Laakso, Welling and Paetau	Anatomy of green open access	-	oa, green, articles, copies, repositories, al, article, et, authors, institutional
65n3, 2014	Waltman and Costas	F1000 Recommendations as a Potential New Data Source for Research Evaluation: A Comparison With Citations	citation analysis	publications, recommendations, citations, publication, recommendation, number, citation, cited, score
65n4, 2014	Larivière, Lozano and Gingras	Are elite journals declining?	bibliometrics	papers, journals, top, journal, proportion, elite, citations, most-cited, published, nature
65n5, 2014	Lindenthal	Valuable Words: The Price Dynamics of Internet Domain Names	trademarks; intellectual property; Internet	domain, domains, tld, prices, names, price, sales, market, com
65n6, 2014	Kinley, Tjondronegoro, Partridge and Edwards	Modeling users' web search behavior and their cognitive styles	end user searching; information seeking; human computer interaction	repositories, repository, opendoar, oa, number, data, growth, december, content, countries
65n7, 2014	Harvey and Harvey	Privacy and security issues for mobile health platforms	mobile communications, privacy, computer crime	security, data, health, mobile, information, services, devices, privacy, use, web
65n8, 2014	Mak	Archaeology of a digitization	History	Eebo, books, digital, digitizations, STC, history, database, English, power, particular
65n9, 2014	Eschenfelder and Johnson	Managing the Data Commons: Controlled Sharing of Scholarly Data	data, information access, digital rights management	data, use, access, repositories, repository, controls, control, required, depositors
65n10, 2014	Bergman, Whittaker and Falk	Shared files: The retrieval perspective	document management, document retrieval	files, retrieval, file, participants, sharing, retrievals, folders
65n11, 2014	White	Belief dynamics in web search	user studies, information seeking	search, yes, beliefs, answer, results, participants, belief, answers, questions, result
65n12, 2014	Pinfield, Salter, Bath, Hubbard, Millington, Anders and Hussain	Open-Access Repositories Worldwide, 2005–2012: Past Growth, Current Characteristics, and Future Possibilities	scholarly communication, diffusion of innovation, open access publications	repositories, repository, opendoar, number, growth, data, december, content, countries
64n1, 2013	Huang and Soergel	Relevance: An Improved Framework for Explicating the Notion	relevance, aboutness	relevance, information, user, topical, systems, topic, need, matching, framework, users
64n2, 2013	Chen, Hu, Milbank and Schultz	A visual analytic study of retracted articles in scientific literature	content analysis, visualization (electronic), information filtering	retracted, articles, article, retraction, citation, cited, al, et, scientific, citations
64n3, 2013	Frandsen and Nicolaisen	The ripple effect: Citation chain reactions of a nobel prize	bibliometrics	Nobel, publications, citation, prize, scientific, references, citations, difference, r-squared, Aumann
64n4, 2013	Sugimoto and Thelwall	Scholars on soap boxes: Science communication and dissemination in TED videos	video communications, scholarly communication, webometrics	TED, videos, YouTube, science, citations, online, metrics, talks, video, comments,
64n5, 2013	Mirel, Tonks, Song, Meng, Xuan and Ameziane	Studying PubMed usages in the field for complex problem solving: Implications for tool design	information seeking, end user searching, qualitative research	users, information, results, pair, query, relevance, disease, research, PubMed, knowledge
64n6, 2013	Wainer and Valle	What happens to computer science research after it is published? Tracking CS research lines	bibliometrics, computer science, quantitative research	papers, conference, journal, conferences, originals, journals, research, published, authors, paper
64n7, 2013	Nicholas, Clark, Rowlands and Jamali	Information on the go: A case study of Europeana mobile users	usage studies, web usage studies, user behavior	mobile, users, Europeana, information, page, use, content, search, behavior, fixed
64n8, 2013	O'Brien and Lebow	Mixed-methods approach to measuring user experience in online news interactions	text mining, content filtering, automatic classification	news, time, information, participants, measures, physiological
64n9, 2013	Boyack, Small and Klavans	Improving the accuracy of co-citation clustering using full text	citation analysis, citation networks, full text databases	co-citation, text, references, reference, cluster, proximity, using, article, coherence, weighting
64n10, 2013	Rosemblat, Resnick, Auston, Shin, Sneiderman, Fizsman and Rindflesch	Extending SemRep to the public health domain	natural language processing, knowledge representation, semantic analysis	semantic, health, concepts, UMLS, new, domain, SemRep, public, relations, types
64n11, 2013	Losee	The effect of assigning a metadata or indexing term on document ordering	probabilistic indexing, performance, metadata	documents, term, document, relevant, metadata, terms, indexing, performance
64n12, 2013	Velden, Carl Lagoze	The extraction of community structures from publication networks to support ethnographic observations of field differences in scientific communication	network analysis, qualitative research, scholarly communication	field, research, network, topic, area, areas, chemistry, groups, collaboration, group

Table 3. Journal of the Association for Information Science and Technology keywords and most frequently-occurring terms