# Topic Analysis of the Research Domain in Knowledge Organization: A Latent Dirichlet Allocation Approach

Soohyung Joo*, Inkyung Choi**, and Namjoo Choi***

*University of Kentucky, School of Information Science, Lexington, KY 40506, <soohyung.joo@uky.edu>

**University of Wisconsin-Milwaukee, School of Information Studies, Milwaukee, WI 53211, <ichoi@uwm.edu>

***University of Kentucky, School of Information Science, Lexington, KY 40506, <namjoo.choi@uky.edu>

Soohyung Joo, PhD, is an assistant professor in the School of Information Science at the University of Kentucky. He obtained a PhD in information studies from the University of Wisconsin-Milwaukee. His main research areas include online resource use, information retrieval, domain analysis, and data analytics. His research work has been published in information science journals, including the *Journal of the Association for Information Science and Technology, Information Processing & Management, and Knowledge Organization*.

Inkyung Choi is a PhD candidate at the School of Information Studies at the University of Wisconsin-Milwaukee. She holds a master's degree in library and information science from the iSchool at Syracuse University. Her research interests stem from intellectual curiosity about social and cultural pluralistic perspectives, which influence ways of organizing knowledge. Her dissertation research is a mixed-methods study investigating ways in which a globalized knowledge organization system can be adapted to a culturally different regional environment and the impacts of sociocultural factors on the adaptation of the system.

Namjoo Choi, PhD, is an associate professor at the School of Information Science, College of Communication and Information, University of Kentucky. He received his PhD in informatics from the University at Albany, State University of New York. His research interests include online communities, free/libre open source software, and technology adoption and post-adoption. His research work has been published in *IEEE Software, Journal of the Association for Information Science and Technology, Library & Information Science Research,* and *Journal of the Association for Information Systems*.

Joo, Soohyung, Inkyung Choi, and Namjoo Choi. 2018. "Topic Analysis of the Research Domain in Knowledge Organization: A Latent Dirichlet Allocation Approach." *Knowledge Organization* 45(2): 170-183. 43 references. DOI:10.5771/0943-7444-2018-2-170.

**Abstract:** Based on text mining, this study explored topics in the research domain of knowledge organization. A text corpus consisting of titles and abstracts was generated from 282 articles of the *Knowledge Organization* journal for the recent ten years from 2006 to 2015. Term frequency analysis and Latent Dirichlet allocation topic modeling were employed to analyze the collected corpus. Topic modeling uncovered twenty research topics prevailing in the knowledge organization field, including theories and epistemology, classification scheme, domain analysis and ontology, digital archiving, document indexing and retrieval, taxonomy and thesaurus system, metadata and controlled vocabulary, ethical issues, and others. In addition, topic trends over the ten years were examined to identify topics that attracted more discussion in the journal. The top two topics that received increased attention recently were "ethical issues in knowledge organization" and "domain analysis and ontologies." This study yields insight into a better understanding of the research domain of knowledge organization. Moreover, text mining approaches introduced in this study have methodological implications for domain analysis in knowledge organization.

Received: 23 May 2017; Revised: 19 August 2017; Accepted: 16 September 2017

Keywords: knowledge organization, KO, research, topic modeling, domain analysis, research trends

## 1.0 Introduction

According to Andersen and Skouvig (2006, 302), knowledge organization (KO) has been a field that investigates "the organization and representation of texts in various forms of information systems for the purpose of mediating, supporting, and producing social practices that constitute every kind of information system." Hjørland (2008) stated that KO encompasses various activities of document description such as indexing and classification, databases, archives and others. These activities of document description involve multiple stakeholders, such as librarians, archivists, subject specialists, and computer algorithms. Hjørland (2008) also observed that KO consists of multiple pillars related to library and information science, supporting learning and research activities, and concepts and theories of knowledge. The science of KO involves multi- and inter-disciplinary comprehension of knowledge and is concerned with the heuristics for conceptual ordering of that which is known or perceived (Smiraglia 2015a). In this way, prior discussions on definitions of KO imply the nature of KO is multifaceted, which includes multiple constituents, diverse objects, and activities. This complex nature of KO has resulted in various subordinate topics within the research field that are explored by researchers with diverse backgrounds and interests.

KO researchers have exerted concerted efforts to probe such diverse aspects of research topics in the KO domain. Multiple methods have been applied to understand the sphere of KO research including qualitative content analysis, bibliometric methods and natural language processing (e.g., Olson 2006; Roe et al. 2007; Smiraglia 2015a; Ibekwe-Sanjuan and Sanjuan 2010; Arboit et al. 2012). These efforts have successfully explained the nature of the KO research domain and guided the directions of KO research among the KO community. This study intends to contribute to this line of research by introducing a recent text mining approach, specifically Latent Dirichlet Allocation (LDA) topic modeling. To the best of our knowledge, LDA has not been used yet in research domain analysis in KO. LDA is an unsupervised machine-learning technique to uncover hidden topics from a large corpus of text documents by analyzing semantic relationships between observed terms (Munzert et al. 2014). Based on text mining and LDA topic modeling, topics were explored from text collected from the *Knowledge Organization* journal (*KO*), which is a principal scholarly venue in the KO field. More importantly, we analyzed topical trends over the recent decade to assess the changes of popular topics and identify the topics that recently received increased attention in the KO field.

## 2.0 Literature review

### 2.1 Research trends in KO

As in many other fields, there exist numerous studies analyzing publication data to identify research trends in KO. For example, by counting the references in the issues of 1991-3 of *Knowledge Organization Literature* according to its classification scheme, Dahlberg (1995) found that the bulk of references (1543/3402) fell into a few foundational classes (e.g., methodology of classing and indexing) and that there were some emerging topics (e.g., automatic classing and indexing techniques). Dahlberg (1995) further highlighted several trends in KO (e.g., the rising significance of KO automation research). In order to survey trends in one of the sub-fields in KO, subject analysis, McIlwaine and Williamson (1999) scanned and categorized relevant published works (e.g., journal articles) appearing in major venues in library and information science (LIS) over a ten-year period from 1988-98. Their findings revealed topics that were most popular during the period (e.g., universal classification systems) as well as an increase in topical diversity.

Olson (2006) performed a content analysis of the articles on organization of information (more specifically, bibliographic control as defined in Olson (2006)) that were published in *Library Quarterly* from volume 1-74 (1931-2004). The quantitative part of her study identified prominent themes throughout the volumes (e.g., cataloging codes) and also showed that the majority of the articles were published in the early years of the journal. In addition, Olson (2006) provided further discussions of the results from a thematic analysis focusing on the identified prominent themes. Roe et al. (2007) conducted two descriptive topical analyses by counting the subject terms assigned to the articles published in *Cataloging & Classification Quarterly* from volume 11 (1990) through volume 40 (2005). The results from the first analysis presented the topics that were most frequently assigned in each of the three article groups by decade (i.e., volumes 11-20, 21-30, and 31-40). The second analysis involved comparison of the first set of volumes (11-20) with the last (31-40) in terms of topical change, identified decreases (e.g., cataloging) and increases (e.g., authority control) in uses of topics. Building on Olson (2006) and McIlwaine (2003), Saumure and Shiri (2008) ran a qualitative content analysis using KO-related articles collected from the Library, Information Science, and Technology Abstracts (LISTA) database. Their findings underlined the growing role of metadata and the heightened prominence of issues pertinent to cataloging and classification since the advent of the Internet.

While the early efforts reviewed above mostly relied on manual or less automatic analysis and authors' knowledge of the field, researchers have begun to employ more advanced approaches in recent years (Ibekwe-Sanjuan and Sanjuan 2010). As one of the first attempts, Smiraglia (2009) used several bibliometric techniques including citation analysis, word and co-word analysis, and author co-citation analysis (ACA) to determine the characteristics and changes of a North American chapter of the International Society for Knowledge Organization (ISKO). The results of the analyses indicated that prolific North American authors had characteristics that were distinct from those of their non-North American peers, and also showed topics that were emergent in North America (e.g., knowledge organization online). In many other publications (some in series), Smiraglia has also been tracking the evolution of KO, by applying similar techniques to regional and international KO conferences (e.g., Smiraglia 2007; 2008; 2011a; 2011b; 2012; 2013a; 2014; 2015a). In order to map trends in KO research, Ibekwe-Sanjuan and Sanjuan (2010) analyzed KO-related journal articles published between 1988 and 2008 (838 out of 931 from the *Knowledge Organization* journal) with natural language processing (NLP), clustering, and information visualization techniques. They found that while the first decade (1988-97) was characterized more by mainstream topics (e.g., classification), the second decade (1998-2008) exhibited more technology-driven and specialized topics (e.g., terminology database). Arboit et al. (2012) employed both citation and social network analyses to identify the most productive authors in the five ISKO conferences from 2002-10 and then examined their relationships with the thematic categories occurring during that period.

### 2.2 Latent Dirichlet allocation (LDA)

LDA is a relatively recent technique proposed by Blei at al. (2003) as a generative probabilistic model for topic modeling. It has been used to discover prevailing themes in collections of scholarly textual data (e.g., journal articles) in various research fields (Blei 2012). In the LIS context, for example, Sugimoto et al. (2011) examined topical trends in all dissertations completed at American Library Association (ALA)-accredited programs from 1930 to 2009. Their LDA analyses determined not only the core themes (e.g., information-seeking behavior) during the period, but also a number of substantial topical changes over time in LIS (e.g., decreasing use of the word library and its related terms). Lu and Wolfram (2012) proposed three methods (i.e., two word-based methods, and one topic-based method using LDA) for measuring author research relatedness. They tested the proposed methods against a more traditional ACA approach with the articles of the 50 most prolific LIS scholars, and they showed that the topic-based method yielded a more distinctive map than the others. Park and Song (2013) carried out a trend survey of LIS research in Korea by applying LDA to the articles published in four major LIS journals between 1970 and 2012 in Korea. Part of their results demonstrated that while several topics such as service and evaluation by libraries were in a growing trend, there were also topics that were in a decreasing trend. In a more recent study, Joo and Cahill (2017) employed LDA to identify topical trends in the field of school librarianship. The data was drawn from the articles published in the field's two leading journals during a ten-year period from 2006-15. They found that programing related concerns were most widely examined in the field, and that some discrepancies existed in terms of topic popularity between the two journals.

The applications of LDA to date have also included trend surveys in other research disciplines. For example, Griffiths and Steyvers (2004) used LDA to determine the topics covered by articles in the *Proceedings of the National Academy of Sciences* (*PNAS*). With the topics determined, they further demonstrated how different scientific domains were related to each other, and also showed topics that gained ("hot" topics) or lost ("cold" topics) popularity over time. Zheng et al. (2006) extracted major recurring topics from a corpus of protein-related MEDLINE articles using LDA and considered the potential of the extracted topics for better indexing and retrieval in biomedical research. Wang, Joo, and Lu (2014) collected a corpus of 550 Wikipedia documents retrieved from a range of search terms relevant to data science, and then, they used the LDA topic modeling to identify twenty-five key topics in the field of data science. Choi et al. (2017) investigated topical trends in personal information privacy research by analyzing relevant articles from Scopus with LDA. Based on the trends that emerged from the analysis, they identified some gaps in the research and made recommendations for future directions. Sun and Yin (2017) applied LDA to articles from twenty-two top-tier journals in transportation research and showed topics that were becoming more popular over time. Also, they ran additional temporal analyses and found different patterns by journal and country or region.

With the ever-growing body of work in KO, it becomes difficult to assess research trends without automatic techniques. Accordingly, as reviewed above, a number of studies have already taken more advanced approaches (e.g., Arboit et al. 2012; Ibekwe-Sanjuan and Sanjuan 2010; Smiraglia 2009; 2014) and offered insights into the dynamic changes in KO research. Along with these efforts, this study takes another step forward by employing LDA to examine recent trends and developments of KO research.

### 3.0 Research questions

In this study, we intend to explore topics that were studied in the *Knowledge Organization* journal over the decade from 2006 to 2015, using text mining to do so. The following three research questions guide the investigation of the present study:

1) What are the terms that *most* frequently occurred in the *Knowledge Organization* journal for the period from 2006 to 2015?
2) What are the research topics that emerged from the *Knowledge Organization* journal for the period from 2006 to 2015?
3) How have research topics changed in the *Knowledge Organization* journal over the past ten years from 2006 to 2015?

### 4.0 Methods

We collected ten years (2006-2015) of published research articles from the journal *Knowledge Organization*. It is the official journal of International Society for Knowledge Organization, which was founded in 1973 by Dr. Ingetraut Dahlberg. The journal has been a critical venue for international KO researchers in representing a variety of KO related subjects. Specifically, according to its main website (http://www.isko.org/ko.html), the major topics the journal covers, but does not limit itself to, are theoretical foundations and practical applications associated with all types of KO such as indexing, classification and thesauri, historical reviews of KO as a discipline, education and training, and terminological issues from general to specific fields. Non-research publications such as editorials, book reviews, journal updates, news, reports of events, and interviews were excluded. In total, 282 published research articles were included for the analysis.

The collected text was analyzed using text mining, to be more specific, term frequency analysis and LDA topic modeling. The titles and abstracts were collected from the selected 282 articles of *KO*. In this study, we delimited the analysis to the titles and abstracts of the collected articles as full-text of research articles are likely to contain some noise (information or unnecessary text that is not directly relevant to the content). In the analysis of research article text, titles and abstracts are considered well organized portions of information and contain key topics of article content (Joo and Cahill 2017). For frequency analysis, we applied three steps of text preprocessing, which are widely applied in textual analysis, including tokenization, stopwords elimination, and stemming. We first extracted all tokens from the collected documents (tokenization), and then stopwords were removed from the corpus. Stemming was applied, which is the process of extracting base or root forms of the observed words. As to stemming, we employed Porter's word stemming algorithm (http://snowball.tartarus.org/algorithms/porter/stemmer.html) to transform terms into stemmed format. We calculated frequencies for stemmed terms and created a term-frequency table with the most frequently observed terms.

Then, LDA topic modeling analysis was carried out to identify prevailing topics underlying the collected corpus of 282 articles. LDA is based on the assumption that a document exhibits multiple topics and each topic is represented as a distribution of observed terms (Blei, 2012). Details of the algorithm behind this method are too complex to present in this paper and is out of the scope of the study. Instead, we attempt to briefly summarize it using the graphical annotation as shown in Figure 1, which is a graphical representation of the LDA topic model. Each document exhibits

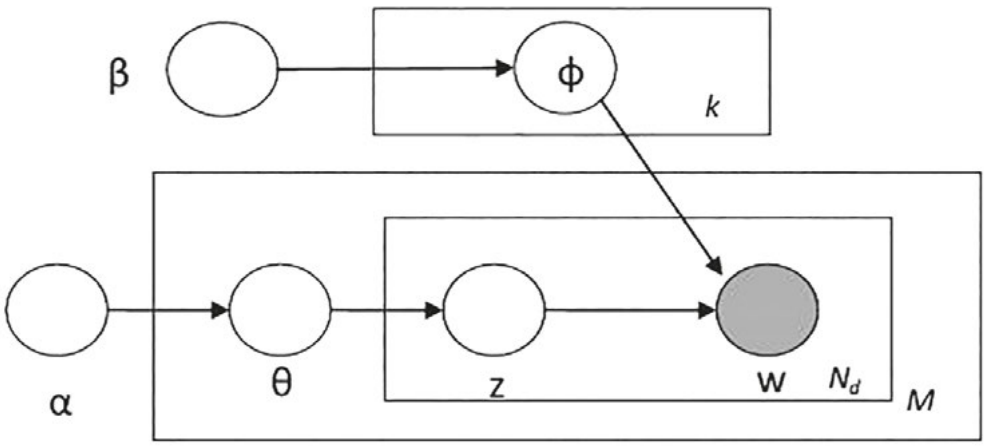


*Figure 1*. Graphic model representation of LDA (Lu and Wolfram 2012, 1975).

a distribution of topic "θ," and each topic consisting of terms (φ) can be generated from a latent Dirichlet distribution with a prior of "β," which is the same parameter of the per-topic word distribution. The LDA begins with a document "$\theta_d$" from "Dir(α)," and "α" is the Dirichlet-prior concentration parameter of the per-document topic distribution. A word ("w") in a document is supposed to be allotted to "z," selected from "$\theta_d$." Thus, "z" indicates the topic assignment for "w," and the word is selected according to "z" and "$\varphi_k$" (where "k" = number of topics, "$N_d$" = number of words in a document). After the iterations of this procedure until it converges, hidden topics underlying the corpus can be extracted (Lu and Wolfram, 2012; Blei, 2012).

In this study, we set "k" as twenty in the LDA model, which is to extract twenty hidden topics from the corpus. In addition, to investigate the trends over ten years, we conducted a term frequency analysis for each year separately and computed the proportions for frequently observed terms between years. We also analyzed the proportions of topics that occurred for each year to assess the changes of topic popularity over the ten years. In this way, we not only depicted prevailing topics in *KO* holistically but also examined the popularity of such topics over time in the KO research domain.

## 5.0 Results

First, we investigated which stemmed terms occurred frequently in the ten-year analysis. In total, 26,596 tokens were observed for 3,132 unique words after removing stopwords. The top 174 terms made up about a half of the entire tokens (49.98%) while 1,269 words were observed only once in the entire corpus. As shown in Figure 2, the observed terms exhibit a typical Zipf law pattern, a reverse J shape. Not surprisingly, the top four most frequent terms are "knowledg," "infor," "classif," and "organ," which showed more than 1% of the entire observed tokens. The stemming results generated "classif" and "classifi" separately even though they have the shared root of meaning. The stemming compiler transformed "classification" and "classifications" into "classif" while it does "classify," "classified," "classifying," and "classifies" into "classifi." If we combine the frequencies of "classif" and "classifi," it totals 429, which makes up 1.613% of the corpus tokens. The analysis of term frequency reveals that "classification" or "organization" of "knowledge" and "information" were the key themes in the *Knowledge Organization* journal. Then, "system," "research," "librari," "studi," and "concept" were among the top ten frequent terms. Table 1 shows the top 100 stemmed terms occurred more than fifty times across the corpus.

Table 2 presents the result of LDA topic modeling, which extracted twenty topics from the text corpus. The most prevalent topic turned out to be "theories" concerning knowledge organization (T18). Approximately 10.6% of the articles are involved with the topics of theories or epistemology. The topics related to classification scheme or facet structure were also popular (Topic 8; 9.57%). Then, domain analysis and ontologies (T19) and library book and collection related studies (T6) were observed to be present in more than 6% of the articles. Topics showed more than 5% among the corpus include: digital archiving (T1), document indexing and retrieval (T14), taxonomy and thesaurus system (T15), web data and topic map (T16),

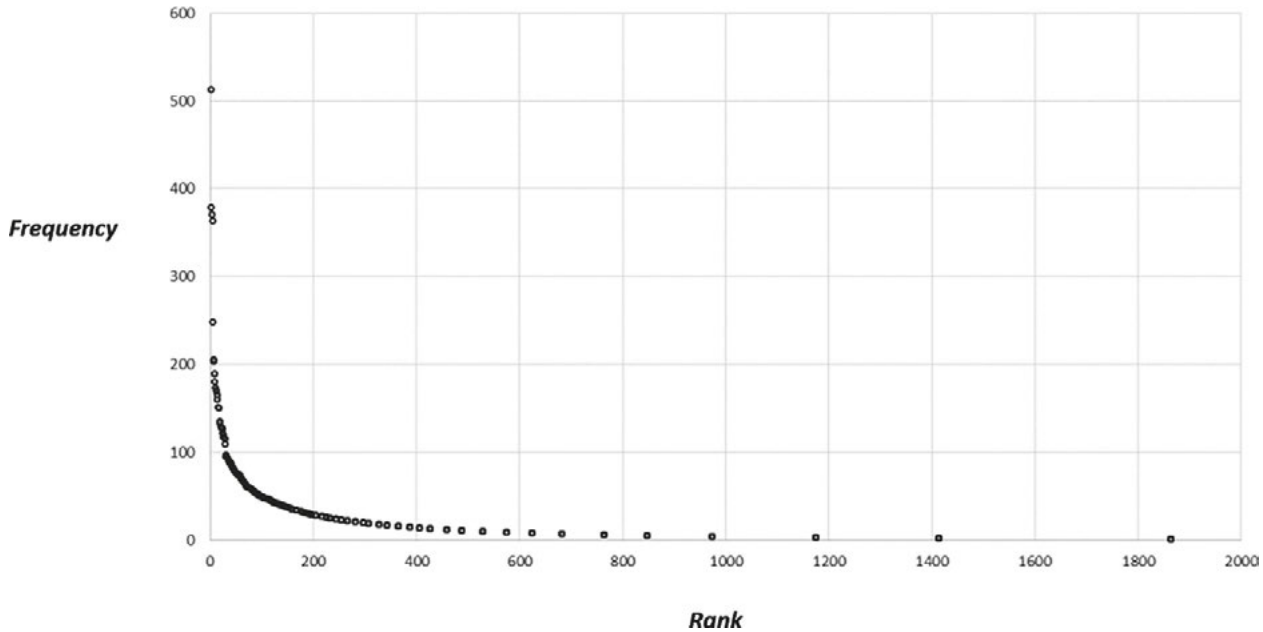


*Figure 2.* Term frequency pattern by rank.

| Term | Rank | Frequency | Percent | Term | Rank | Frequency | Percent |
|---|---|---|---|---|---|---|---|
| knowledg | 1 | 513 | 1.929% | scheme | 50 | 76 | 0.286% |
| inform | 2 | 379 | 1.425% | conceptu | 52 | 75 | 0.282% |
| organ | 3 | 371 | 1.395% | metadata | 52 | 75 | 0.282% |
| classif | 4 | 364 | 1.369% | object | 52 | 75 | 0.282% |
| system | 5 | 249 | 0.936% | propos | 52 | 75 | 0.282% |
| librari | 6 | 206 | 0.775% | languag | 56 | 74 | 0.278% |
| research | 7 | 204 | 0.767% | specif | 57 | 72 | 0.271% |
| paper | 8 | 190 | 0.714% | represent | 58 | 71 | 0.267% |
| studi | 9 | 181 | 0.681% | tool | 59 | 70 | 0.263% |
| subject | 10 | 174 | 0.654% | analyz | 60 | 69 | 0.259% |
| analysi | 11 | 171 | 0.643% | field | 60 | 69 | 0.259% |
| domain | 12 | 170 | 0.639% | vocabulari | 62 | 68 | 0.256% |
| concept | 13 | 166 | 0.624% | order | 63 | 67 | 0.252% |
| base | 14 | 161 | 0.605% | applic | 64 | 66 | 0.248% |
| relat | 15 | 152 | 0.572% | classifi | 65 | 65 | 0.244% |
| develop | 16 | 151 | 0.568% | facet | 65 | 65 | 0.244% |
| user | 16 | 151 | 0.568% | access | 67 | 64 | 0.241% |
| differ | 18 | 136 | 0.511% | repres | 67 | 64 | 0.241% |
| approach | 19 | 134 | 0.504% | record | 69 | 61 | 0.229% |
| document | 20 | 130 | 0.489% | resourc | 69 | 61 | 0.229% |
| index | 21 | 129 | 0.485% | search | 69 | 61 | 0.229% |
| term | 21 | 129 | 0.485% | archiv | 72 | 60 | 0.226% |
| ontolog | 23 | 128 | 0.481% | consid | 72 | 60 | 0.226% |
| semant | 24 | 123 | 0.462% | method | 72 | 60 | 0.226% |
| tag | 25 | 120 | 0.451% | practic | 72 | 60 | 0.226% |
| structur | 26 | 117 | 0.440% | design | 76 | 59 | 0.222% |
| web | 26 | 117 | 0.440% | examin | 76 | 59 | 0.222% |
| scienc | 28 | 115 | 0.432% | general | 76 | 59 | 0.222% |
| model | 29 | 109 | 0.410% | control | 79 | 58 | 0.218% |
| work | 30 | 97 | 0.365% | understand | 79 | 58 | 0.218% |
| present | 31 | 95 | 0.357% | book | 81 | 57 | 0.214% |
| relationship | 31 | 95 | 0.357% | cultur | 82 | 56 | 0.211% |
| retriev | 31 | 95 | 0.357% | includ | 82 | 56 | 0.211% |
| social | 34 | 93 | 0.350% | describ | 84 | 55 | 0.207% |
| result | 35 | 91 | 0.342% | topic | 84 | 55 | 0.207% |
| map | 36 | 90 | 0.338% | issu | 86 | 54 | 0.203% |
| author | 37 | 89 | 0.335% | perspect | 86 | 54 | 0.203% |
| ethic | 38 | 88 | 0.331% | taxonomi | 86 | 54 | 0.203% |
| need | 38 | 88 | 0.331% | identifi | 89 | 53 | 0.199% |
| provid | 40 | 87 | 0.327% | mean | 89 | 53 | 0.199% |
| theori | 41 | 85 | 0.320% | support | 89 | 53 | 0.199% |
| collect | 42 | 83 | 0.312% | level | 92 | 52 | 0.196% |
| context | 42 | 83 | 0.312% | construct | 93 | 51 | 0.192% |
| data | 44 | 82 | 0.308% | import | 93 | 51 | 0.192% |
| discuss | 44 | 82 | 0.308% | thesaurus | 93 | 51 | 0.192% |
| evalu | 46 | 80 | 0.301% | aim | 96 | 50 | 0.188% |
| process | 46 | 80 | 0.301% | bibliograph | 96 | 50 | 0.188% |
| content | 48 | 78 | 0.293% | methodolog | 96 | 50 | 0.188% |
| digit | 49 | 77 | 0.290% | potenti | 96 | 50 | 0.188% |
| articl | 50 | 76 | 0.286% | visual | 96 | 50 | 0.188% |

*Table 1.* Term frequency analysis of KO (terms are stemmed based on Porter Stemmer).

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Topic 1 (5.32%) | digit | archiv | record | access | represent | content | analyz |
| Topic 2 (4.61%) | research | activ | scientif | project | human | influenc | scienc |
| Topic 3 (3.55%) | need | develop | process | order | within | aim | main |
| Topic 4 (2.84%) | social | paper | cultur | critic | within | media | chang |
| Topic 5 (1.77%) | inform | object | design | level | repress | area | paper |
| Topic 6 (6.74%) | librari | book | name | number | collect | examin | given |
| Topic 7 (3.55%) | subject | scienc | articl | journal | search | result | origin |
| Topic 8 (9.57%) | classif | differ | scheme | class | facet | structur | tradit |
| Topic 9 (2.84%) | classifi | work | system | univers | studi | propos | consid |
| Topic 10 (2.84%) | concept | theori | paper | mean | characterist | framework | general |
| Topic 11 (3.90%) | metadata | collect | control | vocabulari | communiti | tool | element |
| Topic 12 (4.61%) | ethic | practic | question | profession | catalog | valu | code |
| Topic 13 (4.96%) | relat | semant | structur | term | present | network | base |
| Topic 14 (5.32%) | document | index | retriev | languag | context | term | approach |
| Topic 15 (5.32%) | system | evalu | develop | taxonomi | thesaurus | construct | support |
| Topic 16 (5.67%) | map | data | web | topic | applic | visual | servic |
| Topic 17 (3.55%) | relationship | model | author | current | express | conceptu | standard |
| Topic 18 (10.64%) | knowledg | organ | epistemolog | theoret | represent | role | foundat |
| Topic 19 (6.74%) | domain | analysi | ontolog | base | research | studi | approach |
| Topic 20 (5.67%) | user | tag | web | resourc | folksonomi | system | cognit |

*Table 2*. Topic modeling results (k=20).

and user tagging and folksonomies (T20). Additionally, the LDA topic modeling discovered the topics of "metadata and controlled vocabulary (T11)," "ethics in KO (T12)," "sematic relationships and structure (T13)," and so forth.

To examine the changes of research topics, the corpus was analyzed over time by year. First, term frequency was computed by year, and Table 3 presents those terms that were observed more than 0.6% in each year. The results indicate that frequent term patterns differed by year. In 2006, "subject," "classif," and "scheme" turned out to be the top three stemmed terms. Also, terms reflecting web environments were highly ranked, such as "web," "metadata," "system" and "semant." In 2007, "inform," "tag," and "classif" were listed as the top three stemmed terms. Particularly, we found Munk and Mork (2007a; 2007b) contributed two research articles regarding folksonomies and user tagging to the journal. In 2008, highly ranked terms were "knowleg," "organ," "relationship," "research," "classif," and "ontology" that accounted for over 1.26% respectively. Interestingly, there were several

| 2006 | | 2007 | | 2008 | | 2009 | | 2010 | |
|---|---|---|---|---|---|---|---|---|---|
| subject | 1.96% | inform | 1.22% | knowledg | 3.78% | knowledg | 2.21% | classif | 2.09% |
| classif | 1.56% | tag | 1.14% | organ | 1.93% | classif | 2.03% | inform | 1.41% |
| scheme | 1.49% | classif | 1.07% | relationship | 1.85% | librari | 1.61% | relat | 1.27% |
| web | 1.25% | develop | 0.99% | research | 1.77% | collect | 1.37% | system | 1.18% |
| differ | 1.17% | term | 0.99% | classif | 1.60% | system | 1.01% | knowledg | 1.00% |
| metadata | 0.94% | metadata | 0.91% | ontolog | 1.26% | index | 0.95% | subject | 0.86% |
| origin | 0.94% | differ | 0.84% | inform | 1.09% | classifi | 0.89% | udc | 0.82% |
| system | 0.94% | document | 0.84% | structur | 1.09% | philosophi | 0.83% | map | 0.77% |
| inform | 0.86% | design | 0.76% | base | 1.01% | inform | 0.72% | studi | 0.77% |
| semant | 0.86% | folksonomi | 0.76% | design | 0.93% | paper | 0.72% | base | 0.73% |
| index | 0.78% | librari | 0.76% | librari | 0.93% | retriev | 0.72% | research | 0.73% |
| organ | 0.78% | relat | 0.76% | relat | 0.93% | context | 0.66% | concept | 0.68% |
| record | 0.70% | user | 0.76% | approach | 0.84% | differ | 0.66% | evalu | 0.68% |
| term | 0.70% | creat | 0.69% | control | 0.84% | metadata | 0.66% | organ | 0.68% |
| applic | 0.63% | knowledg | 0.69% | studi | 0.76% | order | 0.66% | thesaurus | 0.68% |
| librari | 0.63% | number | 0.69% | system | 0.76% | organ | 0.66% | develop | 0.64% |
| resourc | 0.63% | system | 0.69% | vocabulari | 0.76% | web | 0.66% | domain | 0.64% |
| | | tool | 0.69% | develop | 0.67% | approach | 0.60% | paper | 0.64% |
| | | analysi | 0.61% | semant | 0.67% | model | 0.60% | resourc | 0.64% |
| | | descript | 0.61% | | | philosoph | 0.60% | servic | 0.64% |
| | | express | 0.61% | | | studi | 0.60% | term | 0.64% |
| | | organ | 0.61% | | | | | | |
| | | paper | 0.61% | | | | | | |
| | | present | 0.61% | | | | | | |
| | | research | 0.61% | | | | | | |
| | | subject | 0.61% | | | | | | |
| **2011** | | **2012** | | **2013** | | **2014** | | **2015** | |
| inform | 1.75% | inform | 1.86% | knowledg | 2.51% | knowledg | 1.86% | knowledg | 1.49% |
| classif | 1.56% | organ | 1.32% | organ | 1.79% | inform | 1.64% | domain | 1.28% |
| system | 1.18% | knowledg | 1.26% | inform | 1.31% | organ | 1.09% | inform | 1.22% |
| knowledg | 1.14% | classif | 0.99% | classif | 1.13% | research | 1.03% | analysi | 1.16% |
| user | 1.03% | system | 0.99% | ontolog | 1.13% | studi | 0.96% | organ | 1.03% |
| paper | 0.95% | concept | 0.93% | relat | 1.10% | tag | 0.90% | classif | 1.00% |
| organ | 0.91% | paper | 0.90% | index | 1.00% | develop | 0.88% | paper | 0.88% |
| term | 0.91% | ethic | 0.84% | concept | 0.89% | analysi | 0.85% | studi | 0.85% |
| model | 0.84% | librari | 0.81% | analysi | 0.86% | paper | 0.80% | ethic | 0.82% |
| base | 0.76% | document | 0.78% | domain | 0.86% | classif | 0.72% | user | 0.73% |
| web | 0.72% | develop | 0.75% | paper | 0.82% | domain | 0.72% | concept | 0.70% |
| differ | 0.68% | relationship | 0.72% | research | 0.82% | base | 0.69% | research | 0.70% |
| semant | 0.65% | studi | 0.66% | system | 0.82% | scienc | 0.69% | scienc | 0.70% |
| document | 0.61% | research | 0.60% | librari | 0.72% | system | 0.69% | content | 0.61% |
| propos | 0.61% | | | differ | 0.69% | approach | 0.66% | | |
| result | 0.61% | | | model | 0.69% | librari | 0.61% | | |
| | | | | subject | 0.69% | social | 0.61% | | |
| | | | | base | 0.65% | | | | |
| | | | | data | 0.62% | | | | |

*Table 3*. Most frequent terms for individual years (observed more than 0.6%).

articles that directly addressed the issues of "knowledge organization" in that year. For example, Hjørland (2008) defined the domain of KO in an article titled, "What is Knowledge Organization (KO)?," and also, Zeng (2008) discussed different aspects of knowledge organization system (KOS) with relevant examples. In 2009, "knowledge," "classif," "library," "collect," and "system" were among most frequent terms. Interestingly, there were several articles concerning classification and philosophical issues in the context of library collections. That explains why terms

"classif," "library," "classifi," and "philosophi," were observed with high frequency in 2009. In 2010, we found that most popular terms were "classif," "inform," "relat," "system," and "knowledge," which showed proportions of 1% or more respectively. In particular, several articles were contributed to the subject of classification in 2010 (e.g., Osinska 2010; Gnoli 2010; Jacob 2010).

In 2011, again, "classif (1.56%)" was one of the most common terms. Interestingly, the term "user (1.03%)" was ranked fifth, which reflects the occurrence of a relatively larger number of research papers related to users in 2011. For example, Petric et al. (2011) investigated user profiling on a digital library while Kipp (2011) compared user, author and professional indexing. In 2012, the term "ethics" was relatively more highly ranked than in other years. In particular, several articles covering the issues of ethics in KO were published in volume 39 issue 5. In 2013, "ontolog" was listed amongst top five most frequent terms. The 2013, volumes included several articles concerning ontologies and domain analysis. In 2014, the top five terms turned out to be "knowledge," "inform," "organ," "research," and "studi." In 2015, the terms "domain" and "analysi" were ranked highly, which revealed the prevalence of domain analysis research.

Next, we investigated the trends of topics extracted from the LDA method over the period of analysis. Here, we only interpreted the extracted topics showing more than 3% of the entire document set. As shown in Table 4, we did not observe strong, explicit linear patterns of increase or decrease of topic proportions, but certainly, there were certain popular topics in individual years. For example, the topic of digital archiving (T1) reached a temporary acme in 2013, but it was rarely observed between 2006 and 2007. Library collection related research (T6) exhibited three high humps in its pattern by showing intermittent popularities over the period. Similarly, T8 (classification scheme and facet structure) also presented irregular patterns. The topic relevant to metadata and controlled vocabularies (T11) was consistently popular across the ten years while relatively more popular in 2007 and 2015. The topic related to ethics (T12) was most popular in 2015. In particular, volume 42 issue 5 contains the "Proceedings of the 3rd Milwaukee Conference on Ethics in Knowledge Organization." T12 was also popular in 2012 and 2013. T3 (semantic structure and relationship) has steadily occurred across the ten years while it was most popular in 2014. Similarly, T14 (document indexing and retrieval) was also discussed steadily across the period of analysis, except for 2008 and 2010. The topic of taxonomy and thesaurus systems (T15) showed a peak in 2010. The topics relevant to epistemology and theories (T18) were observed unceasingly across the ten years. That topic (T18) was most prevalent in 2013, the year when a special issue was published focusing on theory driven research, "Special Issue: Paradigms of Knowledge and its Organization: The Tree, the Net and Beyond," edited by Fulvio Mazzocchi and Gian

| | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|---|---|---|
| **T1** | 0.0% | 0.0% | 6.7% | 6.7% | 6.7% | 6.7% | 13.3% | 33.3% | 13.3% | 13.3% |
| **T2** | 0.0% | 0.0% | 0.0% | 7.7% | 0.0% | 15.4% | 38.5% | 15.4% | 23.1% | 0.0% |
| **T3** | 0.0% | 0.0% | 0.0% | 20.0% | 10.0% | 10.0% | 20.0% | 20.0% | 20.0% | 0.0% |
| **T4** | 0.0% | 0.0% | 0.0% | 12.5% | 12.5% | 0.0% | 25.0% | 0.0% | 37.5% | 12.5% |
| **T5** | 0.0% | 0.0% | 20.0% | 0.0% | 0.0% | 20.0% | 0.0% | 0.0% | 40.0% | 20.0% |
| **T6** | 0.0% | 5.3% | 0.0% | 21.1% | 0.0% | 0.0% | 26.3% | 10.5% | 10.5% | 26.3% |
| **T7** | 10.0% | 0.0% | 10.0% | 0.0% | 10.0% | 20.0% | 0.0% | 0.0% | 20.0% | 30.0% |
| **T8** | 14.8% | 0.0% | 3.7% | 7.4% | 14.8% | 11.1% | 18.5% | 11.1% | 3.7% | 14.8% |
| **T9** | 12.5% | 0.0% | 0.0% | 25.0% | 12.5% | 12.5% | 25.0% | 0.0% | 12.5% | 0.0% |
| **T10** | 0.0% | 12.5% | 0.0% | 0.0% | 12.5% | 25.0% | 12.5% | 12.5% | 12.5% | 12.5% |
| **T11** | 9.1% | 18.2% | 9.1% | 9.1% | 9.1% | 9.1% | 9.1% | 0.0% | 9.1% | 18.2% |
| **T12** | 7.7% | 7.7% | 0.0% | 0.0% | 0.0% | 7.7% | 15.4% | 15.4% | 0.0% | 46.2% |
| **T13** | 7.1% | 7.1% | 14.3% | 7.1% | 14.3% | 7.1% | 7.1% | 7.1% | 21.4% | 7.1% |
| **T14** | 6.7% | 6.7% | 0.0% | 13.3% | 0.0% | 20.0% | 13.3% | 6.7% | 20.0% | 13.3% |
| **T15** | 13.3% | 0.0% | 6.7% | 6.7% | 33.3% | 0.0% | 6.7% | 0.0% | 13.3% | 20.0% |
| **T16** | 0.0% | 6.3% | 0.0% | 6.3% | 25.0% | 18.8% | 12.5% | 12.5% | 6.3% | 12.5% |
| **T17** | 0.0% | 10.0% | 10.0% | 10.0% | 10.0% | 0.0% | 30.0% | 10.0% | 10.0% | 10.0% |
| **T18** | 3.3% | 6.7% | 13.3% | 3.3% | 6.7% | 6.7% | 13.3% | 26.7% | 16.7% | 3.3% |
| **T19** | 0.0% | 5.3% | 10.5% | 0.0% | 5.3% | 10.5% | 0.0% | 21.1% | 15.8% | 31.6% |
| **T20** | 6.3% | 12.5% | 0.0% | 0.0% | 6.3% | 31.3% | 0.0% | 12.5% | 12.5% | 18.8% |

*Table 4*. Topic trends for the years between 2006 and 2015.

Carlo Fedeli (volume 40 issue 6). T19 (domain analysis and ontologies) was found to be most popular in recent years, especially from 2013 and 2015. To be more specific, domain analysis became one of the main topics in *KO* after 2013. Particularly, volume 42 issue 8 was published as a special issue, titled "Domain Analysis Revisited," specifically covering the subject of domain analysis. "User related studies" (T20) was most popular in 2011, which reaffirms the findings from the term frequency analysis. Figure 3 shows the five topics that received increased attention recently. It indicates that "ethics" and "domain analysis" were the topics with most increased popularity over the decade in the KO research domain.

Finally, we computed the proportions of top fifty most frequent terms by year (Table 5). The top two terms with *most* increasing patterns turned out to be "domain" and "analysi." This suggests that domain analysis was a recent hot topic in the journal. Also, the term "ethic" was another popular term that exhibits recent increased *popularity* in the journal. On the contrary, the terms "classif," "differ," and "subject" exhibited a decline over the past ten years in the *Knowledge Organization* journal.

## 6. 0 Discussion and conclusion

Based on text mining, this study explored research topics that appeared in the *Knowledge Organization* journal for the past decade from 2006 to 2015. We extracted all terms from the titles and abstracts of 282 research articles and made a corpus of stemmed terms after tokenization, stopwords elimination and stemming. Then, we tallied frequencies of those terms from the corpus and identified popular topics in the journal based on LDA topic modeling. Term frequency analysis identified popular terms that occurred in the journal over the ten years. The top four terms, which showed more than 1% of the entire corpus respectively, are "knowledge," "inform," "organ" and "classif." These four terms well represent the fundamental nature of KO research, which covers the key concepts of "organization or classification of knowledge and information." Also, most frequent terms implied key issues that were *popular* amongst KO researchers for the period of analysis. For example, terms related to libraries, domain, subject, user, index, ontology, semantic, web, and tag, among others were ranked within the top thirty most frequent terms. Some of these terms represent recent web environments. For instance, terms such as semantic, web, ontology, user, and tag are closely related to recent discussions of semantic web and collaborative indexing, which have been popular in the latest decade. Also, we can infer that domain analysis became one of the hot topics recently, as the terms "domain" and "analysi" were ranked eleventh and twelfth respectively. Terms concerning ethics and theories were ranked within the top fifty most frequent terms, revealing that ethical issues and theories were frequently discussed in the *Knowledge Organization* journal. This implies that the journal has been a scholarly venue, which focuses on theoretical contribution in KO, beyond the channels for practitioners.

LDA topic modeling provided a semantic level analysis of terms in the KO research domain and uncovered hidden topics underlying the journal. Topic modeling results revealed that a number of articles were concerned with

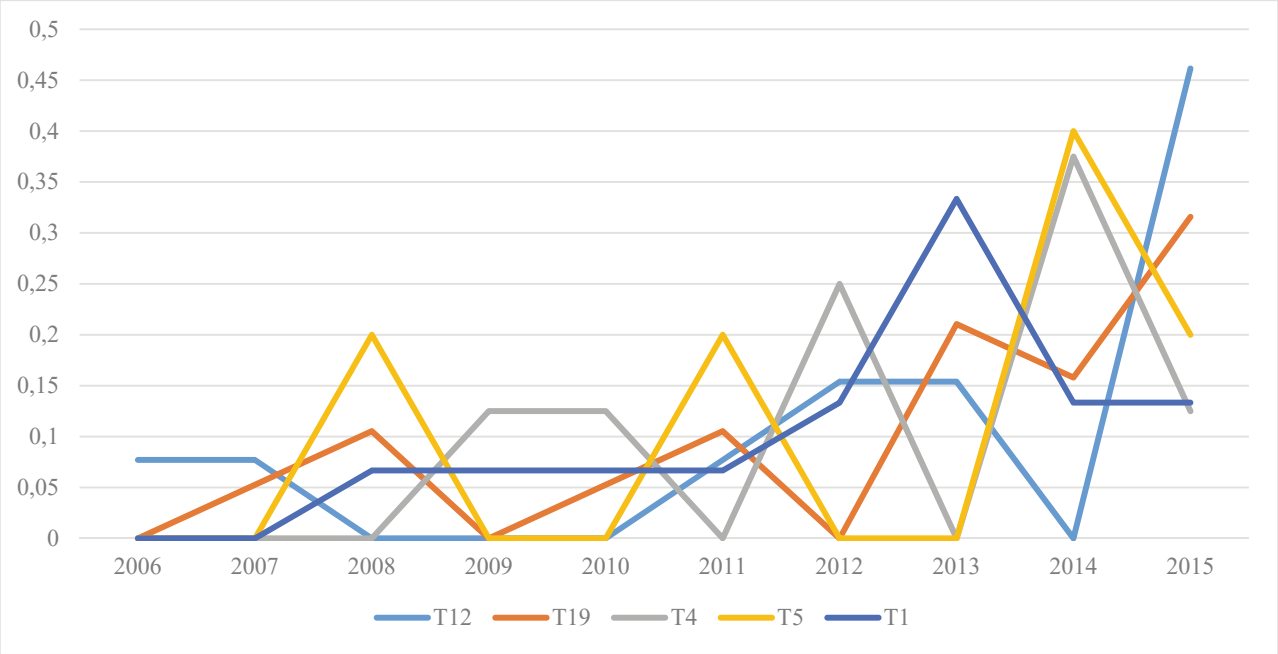


*Figure 3.* Top 5 increasing pattern topics (T12, T19, T4, T5, and T1).

| | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | Linear Slope |
|---|---|---|---|---|---|---|---|---|---|---|---|
| domain | 0.16% | 0.15% | 0.42% | 0.30% | 0.64% | 0.53% | 0.15% | 0.86% | 0.72% | 1.28% | 0.095 |
| analysi | 0.16% | 0.61% | 0.50% | 0.12% | 0.23% | 0.42% | 0.42% | 0.86% | 0.85% | 1.16% | 0.082 |
| inform | 0.86% | 1.22% | 1.09% | 0.72% | 1.41% | 1.75% | 1.86% | 1.31% | 1.64% | 1.22% | 0.067 |
| ethic | 0.00% | 0.00% | 0.08% | 0.06% | 0.05% | 0.00% | 0.84% | 0.24% | 0.05% | 0.82% | 0.066 |
| scienc | 0.00% | 0.08% | 0.34% | 0.24% | 0.27% | 0.19% | 0.42% | 0.24% | 0.69% | 0.70% | 0.064 |
| knowledg | 0.47% | 0.69% | 3.78% | 2.21% | 1.00% | 1.14% | 1.26% | 2.51% | 1.86% | 1.49% | 0.050 |
| paper | 0.47% | 0.61% | 0.50% | 0.72% | 0.64% | 0.95% | 0.90% | 0.82% | 0.80% | 0.88% | 0.045 |
| studi | 0.31% | 0.30% | 0.76% | 0.60% | 0.77% | 0.53% | 0.66% | 0.38% | 0.96% | 0.85% | 0.045 |
| organ | 0.78% | 0.61% | 1.93% | 0.66% | 0.68% | 0.91% | 1.32% | 1.79% | 1.09% | 1.03% | 0.043 |
| data | 0.00% | 0.15% | 0.08% | 0.36% | 0.27% | 0.38% | 0.12% | 0.62% | 0.56% | 0.24% | 0.043 |
| concept | 0.47% | 0.38% | 0.34% | 0.42% | 0.68% | 0.53% | 0.93% | 0.89% | 0.45% | 0.70% | 0.041 |
| social | 0.16% | 0.38% | 0.34% | 0.06% | 0.36% | 0.19% | 0.48% | 0.38% | 0.61% | 0.30% | 0.026 |
| content | 0.39% | 0.08% | 0.34% | 0.00% | 0.18% | 0.11% | 0.45% | 0.24% | 0.24% | 0.61% | 0.024 |
| base | 0.39% | 0.08% | 1.01% | 0.24% | 0.73% | 0.76% | 0.57% | 0.65% | 0.69% | 0.43% | 0.023 |
| theori | 0.08% | 0.53% | 0.25% | 0.06% | 0.09% | 0.53% | 0.21% | 0.34% | 0.16% | 0.58% | 0.020 |
| model | 0.39% | 0.08% | 0.25% | 0.60% | 0.32% | 0.84% | 0.15% | 0.69% | 0.24% | 0.46% | 0.019 |
| evalu | 0.08% | 0.30% | 0.00% | 0.30% | 0.68% | 0.30% | 0.12% | 0.21% | 0.32% | 0.36% | 0.017 |
| provid | 0.23% | 0.15% | 0.42% | 0.24% | 0.41% | 0.30% | 0.51% | 0.38% | 0.29% | 0.36% | 0.016 |
| research | 0.23% | 0.61% | 1.77% | 0.48% | 0.73% | 0.49% | 0.60% | 0.82% | 1.03% | 0.70% | 0.016 |
| map | 0.31% | 0.23% | 0.25% | 0.00% | 0.77% | 0.46% | 0.36% | 0.31% | 0.16% | 0.46% | 0.011 |
| author | 0.16% | 0.46% | 0.42% | 0.06% | 0.23% | 0.34% | 0.54% | 0.24% | 0.48% | 0.24% | 0.010 |
| ontolog | 0.08% | 0.30% | 1.26% | 0.42% | 0.05% | 0.57% | 0.24% | 1.13% | 0.56% | 0.09% | 0.007 |
| articl | 0.16% | 0.23% | 0.42% | 0.30% | 0.45% | 0.38% | 0.27% | 0.17% | 0.11% | 0.52% | 0.006 |
| relat | 0.31% | 0.76% | 0.93% | 0.12% | 1.27% | 0.46% | 0.39% | 1.10% | 0.40% | 0.52% | 0.001 |
| discuss | 0.47% | 0.30% | 0.59% | 0.18% | 0.14% | 0.30% | 0.27% | 0.34% | 0.37% | 0.49% | -0.001 |
| tag | 0.47% | 1.14% | 0.00% | 0.00% | 0.55% | 0.42% | 0.30% | 0.17% | 0.90% | 0.46% | -0.001 |
| present | 0.47% | 0.61% | 0.34% | 0.24% | 0.18% | 0.38% | 0.30% | 0.55% | 0.50% | 0.33% | -0.003 |
| document | 0.23% | 0.84% | 0.50% | 0.42% | 0.32% | 0.61% | 0.78% | 0.45% | 0.50% | 0.30% | -0.004 |
| user | 0.55% | 0.76% | 0.42% | 0.36% | 0.55% | 1.03% | 0.48% | 0.55% | 0.19% | 0.73% | -0.005 |
| process | 0.47% | 0.15% | 0.17% | 0.48% | 0.36% | 0.34% | 0.33% | 0.58% | 0.19% | 0.12% | -0.008 |
| need | 0.39% | 0.38% | 0.50% | 0.54% | 0.18% | 0.27% | 0.27% | 0.41% | 0.27% | 0.43% | -0.010 |
| result | 0.55% | 0.46% | 0.17% | 0.48% | 0.18% | 0.61% | 0.24% | 0.41% | 0.50% | 0.21% | -0.011 |
| work | 0.55% | 0.15% | 0.59% | 0.42% | 0.45% | 0.34% | 0.36% | 0.31% | 0.40% | 0.33% | -0.011 |
| retriev | 0.16% | 0.53% | 0.25% | 0.72% | 0.36% | 0.46% | 0.36% | 0.10% | 0.42% | 0.21% | -0.012 |
| develop | 0.55% | 0.99% | 0.67% | 0.48% | 0.64% | 0.30% | 0.75% | 0.41% | 0.88% | 0.40% | -0.018 |
| approach | 0.47% | 0.53% | 0.84% | 0.60% | 0.36% | 0.57% | 0.42% | 0.41% | 0.66% | 0.30% | -0.018 |
| index | 0.78% | 0.53% | 0.17% | 0.95% | 0.36% | 0.30% | 0.30% | 1.00% | 0.32% | 0.21% | -0.027 |
| digit | 0.55% | 0.38% | 0.08% | 0.30% | 0.32% | 0.23% | 0.42% | 0.27% | 0.19% | 0.06% | -0.027 |
| structur | 0.47% | 0.23% | 1.09% | 0.48% | 0.55% | 0.53% | 0.42% | 0.41% | 0.42% | 0.15% | -0.031 |
| context | 0.31% | 0.46% | 0.59% | 0.66% | 0.23% | 0.27% | 0.33% | 0.24% | 0.24% | 0.21% | -0.031 |
| system | 0.94% | 0.69% | 0.76% | 1.01% | 1.18% | 1.18% | 0.99% | 0.82% | 0.69% | 0.33% | -0.031 |
| web | 1.25% | 0.08% | 0.08% | 0.66% | 0.41% | 0.72% | 0.30% | 0.38% | 0.42% | 0.27% | -0.034 |
| relationship | 0.00% | 0.23% | 1.85% | 0.18% | 0.32% | 0.42% | 0.72% | 0.14% | 0.19% | 0.12% | -0.037 |
| librari | 0.63% | 0.76% | 0.93% | 1.61% | 0.55% | 0.53% | 0.81% | 0.72% | 0.61% | 0.43% | -0.038 |
| term | 0.70% | 0.99% | 0.34% | 0.30% | 0.64% | 0.91% | 0.48% | 0.21% | 0.42% | 0.40% | -0.040 |
| collect | 0.39% | 0.46% | 0.08% | 1.37% | 0.18% | 0.49% | 0.27% | 0.03% | 0.19% | 0.18% | -0.043 |
| semant | 0.86% | 0.53% | 0.67% | 0.36% | 0.32% | 0.65% | 0.24% | 0.48% | 0.53% | 0.15% | -0.045 |
| subject | 1.96% | 0.61% | 0.42% | 0.18% | 0.86% | 0.57% | 0.54% | 0.69% | 0.56% | 0.58% | -0.065 |
| differ | 1.17% | 0.84% | 0.59% | 0.66% | 0.36% | 0.68% | 0.27% | 0.69% | 0.45% | 0.24% | -0.069 |
| classif | 1.56% | 1.07% | 1.60% | 2.03% | 2.09% | 1.56% | 0.99% | 1.13% | 0.72% | 1.00% | -0.082 |

*Table 5*. Proportions of term occurrence for individual years (top 50 terms) and their linear trends over 10 years (ranked by linear slope based on the least squared method).

theories in KO. This indicates that the *Knowledge Organization* journal is a scholarly venue that emphasizes theories related to KO. We observed that many of the articles covered relevant theories in KO. For example, Hjørland, Olson, Fox and Tennis are representative scholars who contributed their theory-driven work to the *Knowledge Organization* journal. One of the strengths of the *Knowledge Organization* journal lies in its coverage of discussion of theoretical and philosophical foundations in KO, in addition to practical matters or empirical findings. Also, the LDA analysis results affirmed that traditional classification and facet structure related research were considered important steadily in the *Knowledge Organization* journal for the decade. According to the LDA results, domain analysis and ontologies became another popular area of interest among the KO community. In this area, Smiraglia (e.g., 2013b; 2015b) has been a leading scholar, and other researchers, such as Castanha et al. (2014) and López-Huertas (2015) also contributed to the studies of ontology and domain analysis. The LDA topic modeling results also uncovered diverse topics that were discussed in the KO research domain over the past ten years. For example, the *Knowledge Organization* journal covered the topics of digital archives, document indexing and retrieval, taxonomy and thesaurus, topical mapping, user tagging and folksonomies, and so forth, between 2006 and 2015.

To examine topic trends we analyzed the proportions of topics by year from 2006 to 2015. The results did not present explicit linear patterns over time for most topic cases. Rather, irregular patterns were observed for most topics. Certainly, we observed particularly popular topics in each year. For example, the topics of user tagging and folksonomies were more likely to be observed in 2011, when it was around the time tagging was popular in KO with the increased use of social media tools. Several researchers, such as Kipp (2011), Rafferty (2011), Mai (2011) and Park (2011), contributed to the area of user tagging, social tagging and folksonomies in 2011. In addition, we identified the topics for domain analysis and ontologies received increased attention recently. Domain analysis is an area that has exhibited increasing popularity since 2013. The patterns revealed by the terms "domain" and "analysi" had the greatest increase in probabilities in recent years. Smiraglia has led this area in the ISKO community by contributing both theories and methodologies relevant to domain analysis (e.g., Smiraglia 2013b; 2014; 2015a; 2015b).

The topic of "ethical issues" was discussed a lot in 2015, and it was partly due to a special issue that includes publications from the 3rd Milwaukee Conference on Ethics in Knowledge Organization. Both the 2012 special issue from the second ethics conference and the 2015 special issue from the third ethics conference have the term, "ethic," in the top ten. The 2015 special issue presented terms directly related to T2 (ethics in KO) such as "ethic," "practice," and "valu," while the 2012 special issue involved more terms of "social," "culture," and "critic" representing T4 (social cultural issues in KO), which discussed the social and conceptual background of ethical issues in KO. This implies that special issues of *Knowledge Organization* may have an impact on the direction of research trends within KO or, conversely, may reflect areas where a critical mass of interest has developed.

This study also yields a methodological contribution to KO. To the best of our knowledge, the present study is the first attempt to employ LDA topic modeling to explore research topics in the KO field. KO researchers have exerted efforts to understand the research domain of KO based on different methods, such as qualitative content analysis (e.g., Olson 2006; McIlwaine 2003; Saumure and Shiri 2008), bibliometric techniques (e.g., Smiraglia 2009; Arboit et al. 2012), and natural language processing and clustering (Ibekwe-Sanjuan and Sanjuan 2010). The LDA method has not been widely introduced yet to the KO field. LDA topic modeling is an unsupervised machine learning technique that can be used to objectively discover hidden themes or topics underlying a large set of textual documents in a certain domain (Blei 2012; Munzert et al. 2014). As shown in this study, text mining based on the LDA approach can be useful to explore research topics in the KO domain. This method is advantageous when a discipline consists of multiple facets of research agenda like KO. As the body of KO research products is getting bigger, automatic data collection and text mining analysis have become imperative to assess the domain of KO. The benefit of using text mining is that it directly examines the content of documents by analyzing relationships among observed terms objectively. Moreover, text mining can be used for other types of domain analysis in KO research, not limited to research topic analysis. Domain analysis basically classifies and identifies different hierarchical or categorical structures of a certain domain consisting of multiple facets. Based on text analysis, we can create a term-document matrix, which enables us to calculate correlation coefficients between documents. Using correlation information among documents, we can automatically cluster or classify documents, which will be useful for domain analysis. In addition, supervised machine learning techniques, such as the "support vector machine" algorithm, can be applied to classify information objects in a particular domain into different categories automatically based on unique characteristics and features of the objects of interest.

This study has several limitations. First, the discipline of KO is an international, interdisciplinary research field, and the *Knowledge Organization* journal is one of the channels that the KO community uses. Even though the *Knowledge Organization* journal is one of the most representative scholarly venues in KO, 282 articles from the journal do not represent

the entirety of the KO research domain. ISKO and regional ISKO individual chapters provide other scholarly communication venues, such as conference proceedings, and KO researchers also publish in other LIS journals. The current study did not investigate those additional venues. Also, this study is limited to the analysis of topics, but it did not examine the relationships among authors in the field. Second, not all topics generated by LDA topic modeling were apparent or discernible. It was not easy to clearly interpret and label all topics extracted. This limited the interpretation and understanding of the findings. Third, the dataset included articles published in special issues. In trend analysis, those special issue articles might have caused somewhat biased results and interpretation. These limitations illustrate a need for future research to investigate topics by researcher to identify the relationships between topics and key researchers. In addition, it is necessary to expand the documents of the KO research domain to include major conference proceedings in ISKO and regional ISKO individual chapters as well as other LIS journal articles to which KO researchers have contributed. Because now all issues of the *Knowledge Organization* journal are available online, we also can extend our analysis of KO to all the journal issues. In carrying out an LDA analysis for the whole corpus of KO journals, we expect to discover dynamic topical trends over time more holistically. In addition, as discussed, the analysis of special issues along with topical trends will discover the impact of social factors on the development of KO research. We plan on a next study to examine the relationships among key researchers based on their publication content using an extended LDA model, which incorporates authors into LDA. All these planned efforts will contribute to the drawing of a better portrait of the KO research domain.

**References**

Andersen, Jack and Laura Skouvig. 2006. "Knowledge Organization: A Sociohistorical Analysis and Critique." *The Library Quarterly* 76: 300-22.

Arboit, Aline Elis, Maria Claudia Cabrini Gracio, Ely Francina Tannuri de Oliveira and Leilah Santiago Bufrem. 2012. "The Relationship between Authors and Main Thematic Categories in the Field of Knowledge Organization: A Bibliometric Approach." In *Categories, Contexts and Relations in Knowledge Organization: Proceedings of the Twelfth International ISKO Conference 6-9 August 2012 Mysore, India*, ed. A. Neelameghan and K. S. Raghavan. Advances in knowledge organization 13. Würzburg: Ergon-Verlag, 44-50.

Blei, David M., Andrew Y. Ng and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993-1022.

Blei, David M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55, no. 4, 77-84.

Castanha, Renata Cristina Gutierres and Maria Cláudia Cabrini Grácio. 2014. "Bibliometrics Contribution to the Metatheoretical and Domain Analysis Studies." *Knowledge Organization* 41: 171-74.

Choi, Hyo Shin, Won Sang Lee and So Young Sohn. 2017. "Analyzing Research Trends in Personal Information Privacy Using Topic Modeling." *Computers & Security* 67: 244-53.

Dahlberg, Ingetraut. 1995. "Current Trends in Knowledge Organization." In *Organización del conocimiento en sistemas de información y documentación: Actas del I Encuentro de ISKO-España, Madrid, 4 y 5 de noviembre de 1993*, ed. Francisco Javier García Marco. Zaragoza: Librería General, 7-25.

Gnoli, Claudio. 2010. "Classification Transcends Library Business." *Knowledge Organization* 37: 223-29.

Griffiths, Thomas and Mark Steyvers. 2004. "Finding Scientific Topics." *Proceedings of the National Academy of Sciences of the United States of America* 101: 5228-35. doi: 10.1073/pnas.0307752101

Hjørland, Birger. 2008. "What Is Knowledge Organization (KO)?" *Knowledge organization 35*: 86-101.

Ibekwe-SanJuan, Fidelia and Eric SanJuan. 2010. "Knowledge Organization Research in the Last Two Decades: 1988-2008." In *Paradigms and Conceptual Systems in Knowledge Organization: Proceedings of the Eleventh International ISKO Conference 23-26 February 2010 Rome, Italy*, ed. Claudio Gnoli and Fulvio Mazzocchi. Advances in Knowledge Organization 12: 115-21.

Jacob, Elin K. 2010. "Proposal for a Classification of Classifications Built on Beghtol's Distinction between 'Naive Classification' and 'Professional Classification.'" *Knowledge Organization* 37: 111-20.

Joo, Soohyung and Maria Cahill. 2017. "Exploring Research Topics in the Field of School Librarianship based on Text Mining." *School Libraries Worldwide* forthcoming.

Kipp, Margaret E.I. 2011. "Tagging of Biomedical Articles on CiteULike: A Comparison of User, Author and Professional Indexing." *Knowledge Organization* 38: 245-61.

Lee, Seongsin. 2016. "A Study on Research Trends in Public Library Research in Korea Using Keyword Networks." *Libri* 66, no. 4: 263-68. doi:10.1515/libri-2016-0052

Lopez-Huertas, María J. 2015. "Domain Analysis for Interdisciplinary Knowledge Domains." *Knowledge Organization* 42: 570-80.

Lu, Kun and Dietmar Wolfram. 2012. "Measuring Author Research Relatedness: A Comparison of Word-based, Topic-based, and Author Cocitation Approaches." *Journal of the American Society for Information Science and Technology* 63: 1973-86.

Mai, Jens-Erik. 2011. "Folksonomies and the New Order: Authority in the Digital Disorder." *Knowledge Organization* 38: 114-22.

Mcllwaine, Ia C. and Nancy J. Williamson. 1999. "International Trends in Subject Analysis Research." *Knowledge Organization* 26: 23-29.

Munk, Timme Bisgaard and Kristian Mørk. 2007a. "Folksonomy, the Power Law & the Significance of the Least Effort." *Knowledge Organization* 34: 16-33.

Munk, Timme Bisgaard and Kristian Mork. 2007b. "Folksonomies, Tagging Communities, and Tagging Strategies-An Empirical Study." *Knowledge Organization* 34: 115-27.

Munzert, Simon Christian Ruoba, Peter Meiboner and Dominic Nyhuis. 2014. *Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining*. Chichester: Wiley.

Olson, Hope A. 2006. "Codes, Costs, and Critiques: The Organization of Information in Library Quarterly, 1931-2004." *The Library Quarterly* 76: 19-35.

Osinska, Veslava. 2010. "Visual Analysis of Classification Scheme." *Knowledge Organization* 37: 299-306.

Park, Heejin. 2011. "A Conceptual Framework to Study Folksonomic Interaction." *Knowledge Organization* 38: 515-29.

Petrič, Karl, Teodor Petrič, Marjan Krisper and Vladislav Rajkovic. 2011. "User Profiling on a Pilot Digital Library with the Final Result of a New Adaptive Knowledge Management Solution." *Knowledge Organization* 38: 96-113.

Rafferty, Pauline. 2011. "Informative Tagging of Images: The Importance of Modality in Interpretation." *Knowledge Organization* 38: 283-98.

Roe, Sandra K., Rebecca Culbertson and Laurel Jizba. 2007. "Cataloging & Classification Quarterly, 1990-2006." *Cataloging & Classification Quarterly* 44: 39-52.

Sarkar, Dipanjan. 2016. *Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insights from Your Data.* [New York?]: Apress.

Saumure, Kristie and Ali Shiri. 2008. "Knowledge Organization Trends in Library and Information Studies: A Preliminary Comparison of the Pre- and Post-Web Eras." *Journal of Information Science* 34: 651-66.

Smiraglia, Richard P. 2009. "Modulation and Specialization in North American Knowledge Organization: Visualizing Pioneers." In *Proceedings from North American Symposium on Knowledge Organization. Vol. 2: Syracuse, NY.,* ed. Elin K. Jacob and Barbara Kwasnik*,* 35-46

Smiraglia, Richard P. 2011a. "ISKO 11's Diverse Bookshelf: An Editorial." *Knowledge Organization* 38: 179-86.

Smiraglia, Richard P. 2011b. "I Simposio Internacional sobre Organizacion del Conocimiento, Bibliotecologia y Terminologia: An Editorial." *Knowledge Organization* 38: 3-8.

Smiraglia, Richard P. 2012. "Universes, Dimensions, Domains, Intensions and Extensions: Knowledge Organization for the 21st Century." In *Categories, Contexts and Relations in Knowledge Organization: Proceedings of the Twelfth International ISKO Conference 6-9 August 2012 Mysore, India*, ed. A. Neelameghan and K. S. Raghavan. Advances in knowledge organization 13. Würzburg: Ergon-Verlag, 6-9.

Smiraglia, Richard P. 2013a. "The Epistemological Dimension of Knowledge Organization." *IRIS-Revista de Informação, Memória e Tecnologia* 2: 2-11.

Smiraglia, Richard P. 2013b. "Is FRBR a Domain? Domain Analysis Applied to the Literature of The FRBR Family of Conceptual Models." *Knowledge Organization* 40: 273-82.

Smiraglia, Richard P. 2014. "II Congresso Brasileiro em Representação e Organização do Conhecimento: Knowledge Organization in Rio 2013—An Editorial." *Knowledge Organization* 41: 105-12.

Smiraglia, Richard P. 2015a. *Domain Analysis for Knowledge Organization: Tools for Ontology Extraction*. Chandos Information Professional Series. Waltham, MA: Chandos.

Smiraglia, Richard P. 2015b. "Domain Analysis of Domain Analysis for Knowledge Organization: Observations on an Emergent Methodological Cluster." *Knowledge Organization* 42: 602-11.

Sugimoto, Cassidy, Daifeng Li, Terrell G. Russell, S. Craig Finlay and Ying Ding. 2011. "The Shifting Sands of Disciplinary Development: Analyzing North American Library and Information Science Dissertations Using Latent Dirichlet Allocation." *Journal of the American Society for Information Science and Technology* 62: 185-204.

Wang, Yanyan, Soohyung Joo and Kun Lu. 2014. "Exploring Topics in the Field of Data Science by Analyzing Wikipedia Documents: A Preliminary Result." *Proceedings of the Association for Information Science and Technology* 51: 1-4.

Zeng, Marcia Lei. 2008. "Knowledge Organization Systems (KOS)." *Knowledge Organization* 35: 160-82.

Zheng, Bin, David C. McLean and Xinghua Lu. 2006. "Identifying Biological Concepts from a Protein-Related Corpus with a Probabilistic Topic Model." *BMC Bioinformatics* 7: 58-68. doi:10.1186/1471-2105-7-58