JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice
eISSN : 2287-4577 pISSN : 2287-9099

Research Paper
J Inf Sci Theory Pract 6(4): 39-53, 2018
https://doi.org/10.1633/JISTaP.2018.6.4.4

Topics and Trends in Metadata Research

Jung Sun Oh

School of Information and Library Science, University of North Carolina at Chapel Hill, NC, USA
E-mail: ohjungsun@gmail.com

Ok Nam Park*

Department of Library and Information Science, Sangmyung University, Seoul, Korea
E-mail: ponda@smu.ac.kr

ABSTRACT

While the body of research on metadata has grown substantially, there has been a lack of systematic analysis of the field of metadata. In this study, we attempt to fill this gap by examining metadata literature spanning the past 20 years. With the combination of a text mining technique, topic modeling, and network analysis, we analyzed 2,713 scholarly papers on metadata published between 1995 and 2014 and identified main topics and trends in metadata research. As the result of topic modeling, 20 topics were discovered and, among those, the most prominent topics were reviewed in detail. In addition, the changes over time in the topic composition, in terms of both the relative topic proportions and the structure of topic networks, were traced to find past and emerging trends in research. The results show that a number of core themes in metadata research have been established over the past decades and the field has advanced, embracing and responding to the dynamic changes in information environments as well as new developments in the professional field.

Keywords: topic modeling, metadata research, research trends, library and information science

Open Access

Accepted date: July 09, 2018
Received date: December 07, 2017

***Corresponding Author:** Ok Nam Park
Associate Professor
Department of Library and Information Science, Sangmyung University, 20 Hongjimun 2-gil, Jongno-gu, Seoul 03016, Korea
E-mail: ponda@smu.ac.kr

1. INTRODUCTION

Metadata lies at the intersection of multiple core areas of information science including knowledge organization and information retrieval. While having its root in traditional bibliographic control in libraries, the area has grown substantially along with the evolution and expansion of the Internet, encompassing principles and practices of resource description for both digital and non-digital materials.

The proliferation of distributed information repositories on the web brought about the need for standardized mechanisms for describing resources, which led to the developments of an array of metadata standards since the 1990s, including the Dublin Core Metadata Element Set (DCMES or DC), Metadata Object Description Schema (MODS), Encoded Archival Description (EAD), and Learning Object Metadata (LOM), to name just a few. Research addressing various issues related to the creation and use of metadata started to appear in scholarly publications in the mid-1990s, and since then the body of literature has grown substantially. Recently, there has been a new wave of advances in web technologies, including Linked Data, which extends the horizon for metadata research even further. Zeng and Qin (2016) noted, in one of the well-known textbooks on the subject of metadata, that "the last two decades of metadata development have witnessed a continual expansion and evolution of metadata research and practices at almost all levels and in almost all disciplines" (p. 18). Yet, to our best knowledge, there has been a lack of systematic analysis of the field of metadata.

In this study, we attempt to examine the body of literature on metadata and address questions regarding the development of the field. While metadata research and practice grow to span multiple disciplines and various areas of interest, as a starting point we set out to trace the growth of the field within library and information science (LIS). More specifically, we apply a text mining method, topic modeling, to the research papers on metadata in LIS literature for the past 20 years to discover main topics addressed in these papers, and the trends in those topics over time. What are the main themes/topics in the discussion of metadata? How has the field changed over time? What topics have gained increasing attention, and what topics have declined over the years? How are these topics interrelated and how have the relationships evolved as individual topics developed? These are questions that we intend to explore. In doing so, we present the topic modeling method combined with network analysis as a promising way for identifying major research topics and studying research trends in literature.

In the following, we will first review the topic modeling method in general and then discuss specifics of our methods. In the result section, the topics identified as a result of topic modeling and our interpretation of the modeling outputs will be presented first, and the analysis of research trends and topic networks will follow. Lastly, the main findings of the study will be discussed to conclude the paper.

2. TOPIC MODELING

Topic modeling has attracted much attention over the past ten years as a tool for computational analysis of large document collections. Based on a probabilistic model, topic modeling uncovers latent 'topics' in a collection of documents or a text corpus (Blei, Ng, & Jordan, 2003; Griffiths & Steyvers, 2004). It starts from an assumption that each document contains a mixture of topics, and the words in the document reflect those topics. The modeling algorithm then infers the hidden topics in a document collection using the observable data—documents and words therein—through unsupervised learning. A topic is represented as a semantically related cluster of words that are likely to appear together in text discussing the topic, with each word having different probabilities of occurrence with regard to the topic. A document, with the occurrences of words associated with different topics, can in turn be abstracted as a probabilistic mixture of those topics. In this way it is possible not only to discover topics addressed in a collection as a whole, but also to figure out what topics appear in which proportions in each document in the collection.

Topic modeling is in fact a label for a family of probabilistic learning algorithms for discovering topics from text corpora. The first and most common method, called Latent Dirichlet Allocation (LDA), was introduced in 2003 in the seminal paper of Blei, Ng, and Jordan (2003). LDA is a generative probabilistic model for a collection of documents, based on the joint probability distribution of the hidden variables (topics) and the observed variables (words in documents). Given a pre-specified number of topics K and a collection of documents containing a fixed set of words (a vocabulary) V, LDA computes the conditional distribution of the underlying topic structure and derives K topics, each as a multinomial distribution over the vocabulary. At the same time, LDA delivers topic assignments for documents, with each document being described as a multinomial distribution over topics (Blei, 2012).

There are several advantages of using topic modeling for the analysis of a large collection of documents. First, as an unsupervised learning technique, it requires no intervention or supervision during the modeling process once the input parameters, including the number of topics, are set. Therefore, using a tool for topic modeling is relatively simple and does not require sophisticated computational skills. The analyst may adjust the input parameters to, for instance, get a more or less find-grained result, but the rest is taken care of by the tool. Second, the model output is readily interpretable by human analysts. The discovered topics can be presented as a list of words with weights denoting the relative importance of each term in describing the topic. Moreover, an examination of the words associated with a topic together with the documents representative of the topic (i.e., documents of which a large proportion is allocated to the topic) helps clarify the meaning of the topics and verify the results. Third, the topics are known to be robust against the inherent ambiguity in language (e.g., synonymy, polysemy) as the model's reliance on word co-occurrence in effect sorts out different contexts in which a word appears. For instance, the term 'library' may appear in two different topics in a collection—in one topic it appears along with terms like 'catalog,' 'monograph,' or 'lending' while in another topic it comes next to 'java,' 'programing,' or 'functions.' The different meanings of the term in these two topics are apparent thanks to the other associated terms. Finally, one of the primary advantages of topic modeling is in its representation of documents as a mixture of multiple topics, which provides a 'soft' clustering or classification of documents. This sets this method apart from other clustering techniques such as K-Means where a document belongs to a single class. Instead of assigning a single topic to a document, topic modeling finds the proportions of multiple topics (the proportions of words associated with those topics) for each document. This more realistic and flexible representation of documents allows further exploration of relationships between topics and documents (Mimno, McCallum, & Mann, 2006).

Over the past decade, many studies show empirically that topic modeling discovers a semantically meaningful set of topics as well as inducing a sensible decomposition of individual documents in terms of those topics. Among others, those studies applying the method to discover topics in research publications reported its usefulness in finding subfields or topical divisions of a research field. Moreover, it was shown that, using the quantitative measures of topic proportions in research literature, it is possible to track changes in the relative prevalence of topics over time, and thereby trace the overall progression of a research field as well (Griffiths & Steyvers, 2004; Hall, Jurafsky, & Manning, 2008; Daud, 2012).

3. LITERATURE REVIEW

Studies related to research trends have already been conducted in several areas of LIS. This study reviewed previous studies regarding the methodology used in research efforts, research trends in knowledge organization, and research trends by means of topic modeling.

3.1. Methodology Used in Research Trends Analysis

The research methods used to analyze trends in research disciplines are distinguished largely by bibliometrics, content analysis, and social network analysis.

3.1.1. Bibliometrics

Bibliometrics is a methodology to apply quantitative methods for literature analysis, mainly quoted by utilizing the index method. It identifies the most highly cited journals in the areas of research, discipline, author, and author cooperation. Patra, Bhattacharya, and Verma (2006), who leveraged the Library and Information Science Abstracts (LISA) to investigate the tendency for bibliometrics literature, described the mid-range of the relevant literature, the language of literature, and authorship patterns. Blessinger and Hrycaj (2010) analyzed the 10 top Journal Citation Ranking (JCR) journals by impact factors to analyze trends in the LIS field of study. They examined 2,200 articles published from 1996 to 2004 and identified the most highly cited journals, articles, and subject areas.

3.1.2. Content Analysis

Content analysis is also a methodology to interpret documents by text analysis. It analyzes the structure, intention, and characteristics that appear in the text, and can be carried out by quantitative or qualitative methods. Shiri (2003) analyzed articles published from three conferences in 2012 to identify research trends in digital library areas, and the study found standards, architecture, usability, issues, and digital content used as focal issues in digital library areas. Julien, Pecoskie, and Reed (2011) conducted a content analysis to analyze trends in information behavior research. They analyzed 749 articles on information behavior published from 1999 to 2008 according to authorship, types of article, journal type, theoretical framework, user groups, degree of attention to users' cognitive processes,

and interdisciplinarity. They found there was an increase in interdisciplinarity in information behavior. Greifeneder (2014) studied 155 articles that were written in the field of information behavior between 2012 and 2014. They employed publication title, authors, publication years, methods, and topics, and main research topic, and identified information seeking as the main research topics, and qualitative methods as the primary methodology.

3.1.3. Social Network Analysis

Social network analysis (SNA) is a methodology to analyze and visualize the network characteristics of group, organization, and data objects. SNA evaluates the network focusing on frequency of keywords, network size, network centrality, and density. Cho (2013) studied articles published in the Republic of Korea and Japan between 2010 and 2012 that were focused on field knowledge organization. The frequency of keywords and network map of the main keywords were analyzed. Feicheng and Yating (2014) utilized SNA to investigate the co-occurrence of tags. They studied tags from the CiteULike and found centricity and groups of tags. They propose that SNA of online tags can be employed as a visualization tool and recommendation resources.

3.2. Research Trends in Knowledge Organization

Studies to investigate research trends regarding knowledge organization have not been carried out to any great extent. Pattuelli (2010) analyzed 34 courses related to knowledge organization in LIS schools in the United States, and outlined the topics and readings taught in the courses. Cho (2013) examined the knowledge organization literature in Japan and the Republic of Korea by network analysis. In addition, Hunter (2003) studied metadata research trends by survey, and found XML semantic web, metadata harvesting, web services, and so on as the main research areas. Parlmer, Zavalina, and Mustafoff (2007) performed an analysis of Institute of Museum and Library Services digital collections projects conducted in 2003 and 2008. They surveyed the project managers and gained an understanding of the metadata audience, metadata application, decision factors, and problems for metadata schema development.

3.3. Topic Modeling Utilization in Research Trends Analysis

Topic modeling has not been much employed in research trend analysis. Most studies have utilized the literature as a dataset to investigate the applicability of topic modeling algorithms.

Griffiths and Steyvers (2004) analyzed abstracts from *Proceedings of the National Academy of Sciences of the United States of America* (*PNAS*) published from 1991 to 2001. They employed LDA and a Markov chain Monte Carlo algorithm to infer about topic modeling. Topics that continue to decrease and increase in the dataset were presented along with topic related terms. Mimno and McCallum (2008) investigated 300,000 articles related to artificial intelligence. They argued that topic modeling can be usefully applied to discover main authors, topics, and predict lead authors for topics. Hall et al. (2008) analyzed 12,500 papers from a journal of the Association for Computational Linguistics Anthology. He employed the LDA technique and discovered hot topics that have been emphasized in anthology, cold topics, and the decline of the leading conference in the field. Daud (2012) carried out topic modeling analysis based on DataBase systems and Logic Programming dataset as well as temporal analysis by year for the main topic, and determined the key authors, key topics, and changes in key topics in the literature.

Previous studies have contained a number of flaws. Studies to understand research trends have been conducted in various areas but there is little specific literature on the research trends of metadata, and studies have relied on limited methodology such as surveys and descriptive content analysis. To unlock key areas, changes in key research areas, and interlinking among research areas, more intensive studies of metadata research need to be done. Looking at the research on topic modeling performed to date, topic modeling has employed literature as data set. Through this, it can be seen that topic modeling can be used to analyze research tendencies.

4. METHODS

4.1. Dataset

In order to investigate topics addressed related to metadata in LIS literature, we constructed our dataset by searching three databases commonly used in the LIS field—LISA, Library and Information Science Source (LISS), and Library, Information Science & Technology Abstracts (LISTA). The three databases have comparable search features, allowing us to construct the same or equitable queries. The searches were done in February 2015. From each database, we retrieved all the records for peer-reviewed scholarly papers written in English that have the term 'metadata' in the title, keyword, or subject fields. The initial dataset had a total of 5,473 records including 1,059 records from LISA, 2,065 records from LISTA, and 2,349 records from LISS. From

the initial dataset, we removed those records that did not contain an abstract, either having an empty abstract field or having text other than an actual abstract (e.g., copyright notes, accession number, etc.) in the abstract field. Since the three databases overlap in their coverage of LIS literature, there were numerous duplicate records in the initial set. We detected duplicates by comparing both title and abstract and removing them. We also removed five papers published before 1995 and 19 papers published in January/February, 2015. The resulting dataset included 2,713 records published in 370 journals between 1995 and 2014. Fig. 1 shows the number of papers in the final dataset by year.

As can be seen in the table, the research literature on metadata grew steeply in the first ten years, from only a handful of papers per year in the mid-1990s to almost 200 papers in 2005. The number of yearly publications has stabilized around 200 since then. The growth can also be illustrated by the number of journals represented in our dataset. Until the mid-1990s the papers on metadata appeared in only a few journals, but drastic increases occur in the last few years of the 1990s and early 2000s. For instance, between the year 2000 and 2001, the number of journals in our dataset rises from 24 to 41, coinciding with a notable increase of papers in 2001.

4.2. Topic Modeling

4.2.1. Preprocessing

The combined title and abstract from each record in our dataset constituted a single document for topic modeling. The documents were preprocessed prior to the topic modeling. After removing punctuation, numbers, and special characters, we deleted words belonging to a stop-word list or with a length less than 3 letters. Since most, if not all, documents contain the word 'metadata,' we also decided to remove the term as it has little value in differentiating topics. For the remaining words (tokens), we performed stemming. All data processing and analysis were done in the *R* computing environment (R Core Team, 2014). The *tm* package (Feinerer & Hornik, 2014) was used to preprocess the documents.

4.2.2. LDA Topic Modeling

For topic modeling, we applied LDA to the preprocessed corpus using the *mallet* package (Mimno & McCallum, 2008), which is an *R* wrapper for the Java-based software MALLET (McCallum, 2002), in *conjunction* with the *topicmodels* package (Grün & Hornik, 2011) in *R*.

As mentioned before, LDA requires a specification of the number of topics *K* as an input for running the algorithm. Considering that the actual number of topics or organizing themes in a corpus is usually unknown, finding the optimal number of topics to be discovered is an important and challenging part of the analysis. Often a trial-and-error approach is taken in practice, trying different *K* values (different number of topics) and settling with a number that leads to the most meaningful results. In our analysis, we

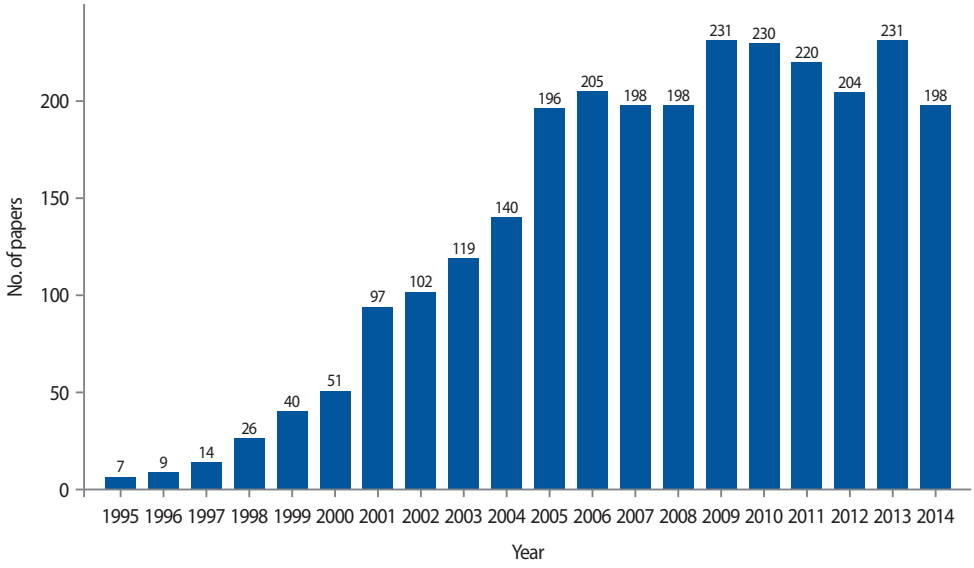

Fig. 1. The final number of papers by years removing duplicates.

tried *K* values of 20, 30, and 40 and chose 20 topics. In the subsequent analysis, we used the model output taken after 5,000 iterations of Gibbs sampling, with *K*=20.

4.2.3. Interpretation

In using topic modeling, assessing the quality of latent topics and interpreting the results is the key analytic task. In interpreting each topic derived from topic modeling, we examined 1) the most probable words (i.e., the words with a high probability of being associated with the topic), 2) the words more distinctive to the topic, and 3) the most representative papers addressing the topic (i.e., the papers where the vast majority or a significantly large proportion of its content words belongs to the topic). A label was then assigned to each topic.

The difference between the probable words and the distinctive words (1 and 2 above, respectively) related to a topic is explained in the following. As mentioned before, the outcome of topic modeling includes the distributions of words over topics. Therefore, a ranked list of words with a high probability of appearing in a topic can be easily created, and in fact, this list is most commonly used for topic interpretation. However, since the probability is affected by the overall frequencies of words, it is often the case that some generic words or common jargon in the domain take place in the top ranked list for multiple topics, making it difficult to differentiate the meanings of these topics. In order to solve this problem, different measures for identifying and ranking topic words were suggested. Among others, we adopt Sievert and Shirley (2014)'s measure of 'relevance' defined as relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w). This measure considers how uniquely or distinctively a word is associated with a topic, and thereby reduces the weight of those generic words occurring frequently in many topics while increasing the weight of most 'relevant' terms for a given topic. We examined the top thirty probable terms and top thirty relevant terms (at λ=0.6) for our interpretation of topics. In addition, as mentioned above, a number of documents where a given topic appears dominantly are also reviewed to verify the topic contents.

4.3. Research Trends and Topic Networks

In addition to finding out metadata-related topics addressed in research literature, we are interested in how these topic areas have evolved over time. As in Griffiths and Steyvers (2004), we conducted a post hoc analysis on the topic modeling outcomes to track changes in topic distributions over time. More specifically, the proportion of topics assigned to documents were aggregated by the publication year of the documents, and each topic's share in yearly publications was plotted to uncover any trends.

In order to look at the relationships between topics and the dynamics of the relationships over time, we employed network analyses of topics, using the topic distribution over documents. In topic modeling, each document is represented as a mixture of topics. In some cases, a document has a single prevalent topic with other topics having marginal proportions. In some other cases, a document comprises multiple topics each with a non-trivial proportion. We suppose that, if two topics are frequently addressed together in substantial proportions in the same documents, it indicates possible relationships between those topics. That is, relationships between topics are established based on the documents in which the given topics appear together.

5. RESULTS

As a result of performing the topic modeling of datasets, twenty latent topics were discovered. Table 1 shows the twenty topics with assigned labels, in the order of their prevalence in the entire dataset, along with the most probable words next to each topic.

As explained in the method section, in order to interpret the topic content and assign the label for each topic, we examined 1) the most probable words, 2) the most relevant words (words that are most distinctive to the given topic), and 3) the representative papers of the topic. Due to the space limitation, Table 1 includes only seven most probable terms for each topic. We will mention any notable differences between probable words and relevant words in our discussion of derived topics in the following.

Note that the topic modeling outcome included words in their stemmed form since we did stemming of words in the preprocessing phase, but for better readability, stems were changed back to complete words in Table 1. Note also that we removed the term 'metadata' before topic modeling because most, if not all, of the documents in our collection include the term, and therefore it does not have a value for identifying distinct topics in the collection. However, when we interpret the results presented in Table 1, it would be reasonable to presume that the top terms may be used in conjunction with the term 'metadata' or in a broad context of metadata related issues. Therefore, we used the term 'metadata' in topic labels where it seems appropriate.

Table 1. Discovered topics

No	Topic name	Probable terms
1	Digital library projects	digital, collection, library, project, preservation, archive, access
2	Development or management of information systems/services	system, manage, service, base, develop, implement, integrate
3	Role of metadata or metadata librarians	library, resource, librarian, service, develop, technology, digital
4	Evaluation of metadata quality	study, paper, quality, analysis, research, result, data
5	Semantic web and ontology	semantic, web, ontology, model, knowledge, base, paper
6	Metadata standard development	project, standard, develop, resource, access, nation, work
7	Record management	manage, record, paper, system, research, knowledge, context
8	Cataloging	catalog, library, record, bibliographic, catalogue, author, marc
9	Social tag and folksonomy	tag, user, social, subject, term, index, folksonomy
10	Automatic extraction methods	document, base, method, extract, automatic, system, index
11	Search and retrieval	search, user, retrieve, image, music, query, result
12	Linked Data	data, link, research, map, science, scientific, geographic
13	Metadata harvesting	repository, institute, open, oai, harvest, research, protocol
14	Publishing and access	publish, article, journal, book, access, public, scholar
15	XML and encoding standards	standard, description, xml, archive, encode, article, document
16	Dublin Core	core, element, dublin, resource, standard, set, develop
17	Search engine and web sites	web, site, page, engine, meta, search, description
18	Domain metadata: education and heath	learn, education, object, student, health, resource, medical
19	Metadata for multimedia and social media	content, network, media, user, video, social, multimedia
20	Conference and meeting reports	library, conference, present, report, association, meet, discuss

5.1. Prominent Topics

In this section, we review topics 1 to 8 in detail, as they turned out to be the most prominent topics in metadata literature. Collectively, these topics account for 57.6 % of entire words (tokens) in the corpus, with each topic assuming more than 5%.

Topic 1 is about digital collection or digital library projects, as can be seen in the words falling in this topic with high probability. Included in the top third probable or relevant words are a set of words referring to the organizations in charge of such projects, including *library*, *archive*, *cultural*, *heritage*, and *institution*, as well as a group of words for information objects, such as *object*, *material*, and *image*. In addition, the words related to some key functions of metadata, such as *describe*, *access*, or *preservation*, also appeared high in the lists. The words associated with this topic occupy more than 10% of the corpus, making it the most prominent topic in the metadata literature in our dataset.

As expected, many of the representative papers of this topic report and share the experiences of building a digital collection or setting up a digital library, often discussing metadata related issues or challenges encountered in the process (e.g., Boyd & King, 2006; Woodley, 2002). There are also papers addressing specific aspects of a project, such as the implications of technical choices for digitization or the need for documenting metadata decisions (e.g., Lalitha, 2009; Symonds & May, 2009).

Topic 2 concerns the development or management of information systems/services. The main keywords for this topic include nouns such as *system*, *service*, *software*, and *application* as well as verbs like *develop*, *manage*, *implement*, and *integrate*. Also ranked high on the list of probable/relevant words are *design*, *architecture*, *model*, and *framework*. Combinations of these terms give a fairly good idea as to the topic area.

The papers in this topic discuss metadata based approaches or solutions for specific problems at hand in relation to information systems or services: for instance, access control (Yagüe, Maña, & Lopez, 2005), content management (Yeh, Chen, Sie, & Liu, 2014), or a federation of distributed resources (Aktas, Fox, & Pierce, 2010). Issues concerning the design and implementation of such solutions, or the proposed models or architectures, are commonly found in those papers.

Topic 3 deals with role of metadata or metadata librarians. The most probable words for this topic are *library*, *resource*,

librarian, *service*, *develop*, *technology*, and *digital*, but this topic can be better understood when the relevant words distinctive to this topic, such as *role*, *skill*, *future*, and *profession* are considered.

The papers in this topic appear to divide broadly into two groups, while commonly noting the challenges that the proliferation of electronic resources have brought to library services. One group tackles the problem of organizing electronic resources and discusses the increasingly important role of metadata (e.g., Medeiros, 2003; Emery, 2007). The other group centers on the discussion of professional roles of librarians in the digital era, reflecting on the influence of technologies on library and information services. Many state the need for reconfiguring technical services or cataloging practices to better meet current and future challenges, and call for attention to the changing roles and competencies of librarians (e.g., Schottlaender, 2003; Han & Hswe, 2011).

Topic 4 covers evaluation of metadata quality. The top ten words with the highest probabilities of being associated with this topic are *study*, *paper*, *quality*, *analysis*, *research*, *result*, *data*, *find*, *survey*, and *evaluate*. The rank order of relevant words differs slightly, with *quality*, *analysis*, *survey*, and *evaluate* placed higher.

The most representative papers of this topic report the results of studies on metadata quality. Some analyze a number of metadata records and identify patterns of problems or errors therein. Often certain criteria such as accuracy, consistency, and completeness are used to evaluate the quality of metadata (e.g., Chuttur, 2012). Some further suggest and/or test mechanisms for quality assurance (e.g., Park & Tosaka, 2010; Chuttur, 2014). Many papers present empirical studies where a variety of methods including experiments, focus group interviews, and surveys are employed. Yet, some discuss quality issues based on a review and an analysis of research and practice in the field (e.g., Park, 2009).

Topic 5 is focused on semantic web and ontology. Top keywords for this topic are *semantic*, *web*, *ontology*, *model*, *knowledge*, *base*, *paper*, *relation*, *concept*, *structure*, *domain*, and so on. As can be seen in the above list of keywords, where all except for one rather general term *paper* represent a coherent theme, there is little to no ambiguity about this topic area.

The papers in this topic review various semantic web technologies (e.g., Kanellopoulos & Kotsiantis, 2007), or discuss ontology modeling or implementation, including a conversion of existing controlled vocabulary or metadata (e.g., Qin & Paling, 2001).

Topic 6 is about metadata standard development. The first glance at the list of probable words in this topic does not give a clear idea as to its topic content, as it includes rather generic terms that appear in multiple topics, such as *project*, *develop*, *resource*, and *access*. Only the term *standard* is relatively unique to this topic. However, when we examine further down the list of the terms relevant to this topic, along with its representative papers, it becomes evident that this topic centers on metadata standards, especially the development of various national or international standards. The names of standards organizations, such as *NISO* and *ISO*, are often mentioned in the papers and related terms like *initiative*, *committee*, (*working*) *group*, or *programme* appear high on the list of relevant terms.

Some papers in this group discuss the importance of developing and adopting metadata standards (Lagace, Breeding, Romano Reynolds, & Han, 2013), some introduce a newly developed standard or provide updates on a standard under development (e.g., Feick, Henderson, & England, 2011), and some provide a review of an existing standard, often with a discussion of emerging issues (e.g., Mullen, 2001).

Topic 7 mainly concerns record management. The top keywords in this topic include *manage*, *record*, *paper*, and *system*. Additional words particularly relevant to this topic are *recordkeeping*, *context*, and *scheme*.

A variety of questions regarding metadata for record management or record keeping have been addressed, including the role, purpose, or capacities of metadata in the context of electronic record management, the specifications for records metadata, the methods for acquiring or capturing record keeping metadata, the need for standards and tools, and so on (e.g., Evans & Rouche, 2004; Evans, 2007; Cumming, 2007).

Topic 8 is related to cataloging or bibliographic description of resources, as clearly shown in the lists of probable and relevant topic words including *catalog*(*ue*), *record*, *library*, *bibliographic*, *MARC*, *RDA*, *access*, *control*, *description*, *resource*, and so on.

Although sharing the central theme, cataloging, research problems addressed in the papers range broadly from the pertinence of specific aspects of cataloging rules and/or principles in modern catalogs (e.g., Conners, 2008), to the applicability of cataloging tools to organization of Internet resources (e.g., Ferris, 2002), and to future directions for library catalog or cataloging rules (e.g., Wakimoto, 2009). Many of the recent papers report testing and implementation of the new cataloging standard, Resource Description and Access (RDA) (e.g., Danskin, 2014).

5.2. Research Trends

The topics of prominence we looked at in the previous section are based on the proportion of words associated with a given topic within the entire corpus. That is, the result shows the distribution of the twenty topics across all the papers in the corpus that are published in the span of 20 years. In order to see the dynamics of these topics, we now turn our attention to the changes in the topic proportions in the corpus over time.

Since topics are assigned to each document with their respective proportions, when we aggregate the proportions of a given topic in the documents published in a year, we can obtain the share of the topic in that year's literature. This provides quantitative measures of rise and/or fall of topics in popularity over time, as well as their relative prevalence.

For the overall trend analysis, we plotted topic shares from 2000 to 2014. The publications from 1995 to 1999 were not included in the plots, since the number of publications during the period was too small for this analysis. The plots for all twenty topics were first drawn, and among them, we identified topics with an upward trend, a downward trend, and those reaching a peak at different points in time. In the following, we present those topics.

5.2.1. Emerging Topics

Topic 4 (evaluation of metadata quality), Topic 10 (document extraction methods), Topic 12 (Linked Data) have increased in volume since 2000 as shown in Fig. 2.

Topic 4 (evaluation of metadata quality) shows a steep increase from 2000 to 2010, with its topic share being more than tripled. This trend demonstrates that growing attention has been given to quality issues as metadata has become a pillar of information services.

Topic 10 (document extraction methods) is related to automatic extraction of metadata from documents. Noting both the need for and the expenses of creating metadata, the papers in this topic explore various strategies and methods for automatically generating metadata using different parts of documents per se. This topic shows a good deal of fluctuation, yet the overall topic share has increased considerably over time.

The papers associated with Topic 12 (Linked Data) show a rapid increase since 2006, the year when Tim Berners-Lee coined the term and outlined the key principles for publishing and connecting structured data on the web. Also included in this topic are papers on data mining, or on curation or retrieval of special data including scientific data and geographic data. This explains why the topic was present before 2006 and has a spike in 2002. The vast majority of the representative papers (24 out of the top 30 papers) in this topic, however, were published after 2006, many touching on the application or adoption of Linked Data concepts, technologies, and practices to the creation, transformation, and use of Libraries, Archives, Museums (LAM) metadata.

5.2.2. Declining Topics

Interestingly, topics showing a downward trend (Fig. 3) are related to metadata standards: topic 6 (standard development), topic 15 (XML and encoding standards), and topic 16 (Dublin Core). It appears that these topics attracted the most interest at the early stage of metadata research, before or around the beginning of the 2000s, but began to fade at least from the research front.

As described in the previous section, topic 6 has to do with the development of metadata standards, and many

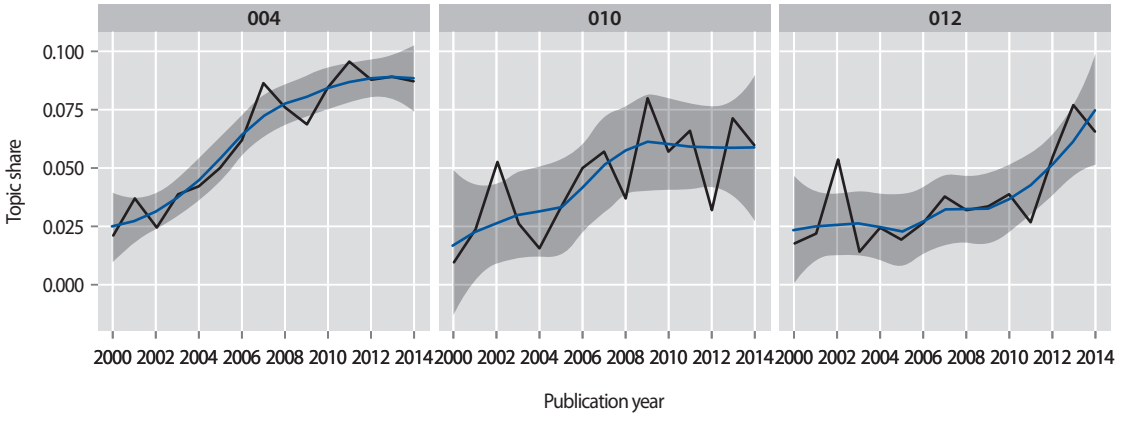

Fig. 2. Emerging topics showing upward trends.

papers in this topic introduced a then new standard or reported updates on the development or implementation of a standard.

Topic 15 is about XML and encoding standards. There is a surge of interest in this topic in the metadata community from 2000 to 2002, demonstrated by a special section on XML in *Library Hi Tech* (volume 19 in 2001). Many of the papers in this period provide an introduction to the suite of XML specifications and technologies, along with a discussion of their implications for metadata sharing. In addition, the release of the XML-compliant EAD version 1 in 1998 seems to kindle the interests of archivists in this topic, resulting in a series of papers on XML encoding of archival resources in following years.

Topic 16 about Dublin Core has declined steadily. As one of the first metadata standards that received international recognition and enjoyed wide adoption, Dublin Core took a central place in the early phase of metadata research, but has given its share to other emerging topics as the research horizon expands. In fact, this topic reached its peak in 1997 (not shown in the plot), where its topic share amounted to 16.8% of the entire corpus. The topic share continued to decrease to 8% in 2000, 3.8% in 2006, and finally to 2% in 2014.

5.2.3. Topics with Peaks and Valleys

The dynamic changes in the area are also observed in a set of topics that reached their peak (or valley) in different points in time at different pace (Fig. 4). Topic 9 (social tag and folksonomy) and topic 13 (metadata harvesting) gained sizable attention at one point then have gone down, and topic 8 (cataloging) showed an opposite pattern.

Topic 9 is about social tag and folksonomy. The steep increase of this topic in literature is closely related to the development of social bookmarking or tagging tools. Del.icio.us was founded in 2003, Flickr in 2004, and Ma.gnolia

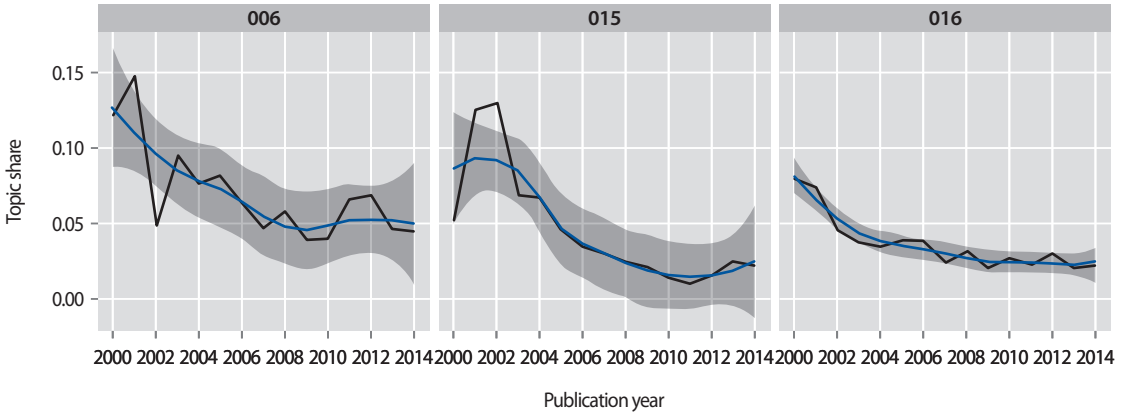

Fig. 3. Declining topics showing downward trends.

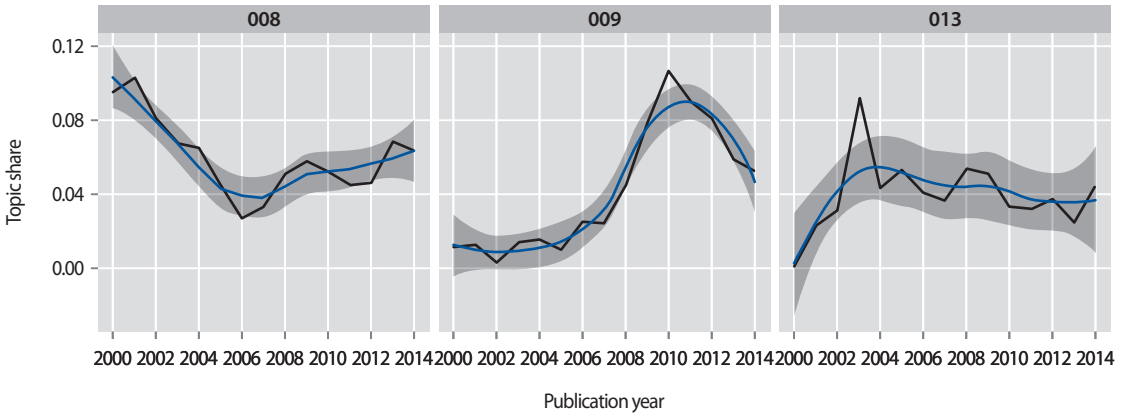

Fig. 4. Topics with peaks and valleys showing dynamic changes.

in 2006. The idea of building up a bottom-up taxonomy, dubbed as *folksonomy*, using user-generated tags was booming as those tools gained enormous popularity. Many libraries have also started incorporating tagging features into their catalogs. The published work on this topic peaked in 2010 but has rapidly decreased since then, following the downturn of tagging services in general.

Topic 13 represents a coherent set of discussions on metadata harvesting and issues related to constructing and maintaining metadata repositories. The research on metadata harvesting took off shortly after the introduction of Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) version 1.0 in early 2001, and the sudden burst of this topic between 2002 and 2003 coincides with the release of OAI-PMH version 2.0 specification. While the high level of attention to this topic per se in literature did not sustain itself after its peak in 2003, the breath of discussions appear to have expanded, as shown in its increased relationships with other topics (presented in the next section).

Topic 8 is an interesting case. It showed a clear downward trend until 2006, but bucked up the trend since then. This change appears to be a response to the substantial developments undergoing in the cataloging field, including the work on the new cataloging rule, RDA, and the continuing discussion on improving or substituting Machine Readable Cataloging (MARC).

5.3. Topic Networks

Having identified the main themes of metadata discourse and having looked at the changes in prominence of such topics over time, we now are interested in how these topics are interrelated and how the strengths of the connections between topics have changed during the fifteen year period.

As described in the method section, in order to look into the relationships between topics, we adopted a network analytic method. Using topic proportions in each paper, we derived connections between topics based on the papers addressing two or more topics together. In order to determine those papers spanning multiple topics, we set the threshold at 30%. That is, if a paper's topic composition consists of two or more topics each in a proportion of 30% or more, the paper constitutes a potential link between the topics. Between 2000 and 2014, about 32% of the papers (838 out of 2,617) fell under this category.

Fig. 5 shows the topic network considering all papers published between 2000 and 2014. The size of a node is proportional to its degree, and the thickness of a link between two nodes denotes the strength of their connections

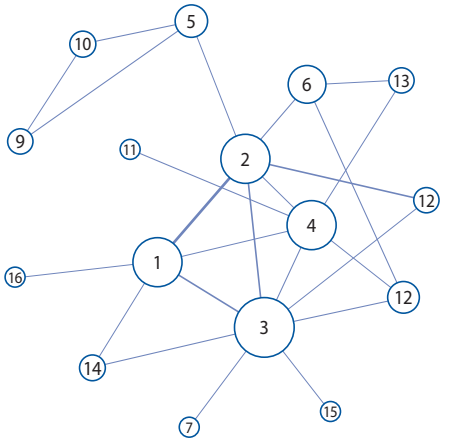

Fig. 5. Topic network considering all papers published between 2000 and 2014.

determined by the number of papers addressing the two together. Each link in this network has a minimum value of 10, which means ten or more papers addressed the topics on each side of the link together. The isolated nodes, topics with no connections, are removed from the figure.

Given the condition of ten or more shared documents to establish connections, 16 out of 20 topics have at least one connection with other topic(s). The node with the highest degree, the most well-connected topic, is topic 3, having links to eight other topics. This is not surprising since the role of metadata or metadata librarians is a topic that can be discussed in various contexts. Most of the prominent topics assuming a large proportion in the corpus (topics 1 through 8) have a relatively high degree ranging from three to eight, except for topic 7 (record management). Topics 1 through 4 clearly constitute the tightly-knit core of the network, to which smaller topics are connected with varying strengths. It is notable that the more technically oriented topics, topics 5, 9, and 10, form a clique somewhat separated from the core.

In order to look at how the relationships between topics have evolved over time, as well as to get more insights into their composition, we divide the 15-year period into three sub-periods (period 1: 2000–2004, period 2: 2005–2009, period 3: 2010–2014) and draw a topic network for each sub-period. Considering the relative sizes of the document sets in these periods, we adjusted the number of shared documents to create links between topics—a minimum of four shared documents was used for period 1, and a minimum of five for period 2 and period 3. Fig. 6 presents the three sub-period networks.

As can be seen in Fig. 6, the networks show considerable differences not only in the volume of connections but also

in their structure. Overall, it is clear that increasingly more connections among topics have arisen as time passes, but at the same time there are notable changes in topics assuming central positions.

In the first five year period (2000–2004), the most noticeable difference compared to subsequent periods is the prominence of topic 6 (metadata standard development) and topic 15 (XML and encoding standards) both in terms of their degree and the strengths of connections that they have with other topics. Topic 16 (Dublin Core) is present only in period 1, being connected to topic 6 and topic 15. Note that these topics all showed a downtrend from around 2000, as described in the previous section. The fact that they remained salient in the network suggests that, while declining in volume, these topics relating to base standards still had an important place in discussion of other topics. These topics, however, become peripheral in subsequent periods.

In period 2, topics 1 (digital library projects) and topic 2 (development or management of information systems/services) moved to the center of the network. Topic 1's central position, with its links to a variety of topics, indicates that various metadata research topics were often introduced and discussed in a context of digital projects. Topic 2, on the other hand, showed a tendency of having connections with more tech-oriented topics, including topic 5 (semantic web and ontology) and topic 12 (Linked Data). It is also notable that topic 4 (evaluation of metadata quality) started to appear on the network in this period.

In period 3, topic 3 (role of metadata or metadata librarians) emerged as a center of the network, with a degree of nine. It indicates that a discussion of the function of metadata or the professional role of metadata librarians takes place in a variety of topics in more recent literature. In addition, topic 12 (Linked Data) and topic 14 (publishing and access) show a considerable growth in links, reflecting new trends in research.

The composition of networks and the changes therein portray how topics develop in relation to other topics. For instance, looking at topic 13 (metadata harvesting), the topic was first connected to topic 6 in period 1, topics 1 and 2 in period 2, and finally topics 3 and 4 in period 3. That is, papers introducing the ideas and mechanisms of metadata harvesting, including the OAI protocol standard itself, first formed a link between topic 13 and topic 6. Then papers describing a project creating regional or national repositories using OAI-PHM and papers concentrating on design and implementation of services building upon metadata repositories appeared, connecting this topic to topic 1 and topic 2, respectively. Finally, papers focusing on the evaluation of consistency of metadata elements and values across harvested records constitute the link between topic 13 and topic 4 in period 3. In addition, discussions on academic libraries' role in managing and promoting OAI compliant institutional repositories spanned both topic 13 and topic 3.

6. DISCUSSION AND CONCLUSION

In this study, we collected 20 years of metadata literature, and identified and labeled 20 topics present in metadata literature. We found a set of sensible and coherent research topics from the literature spanning 20 years, and also traced a number of research trends. Among those, more prominent topics, in terms of their proportion in literature,

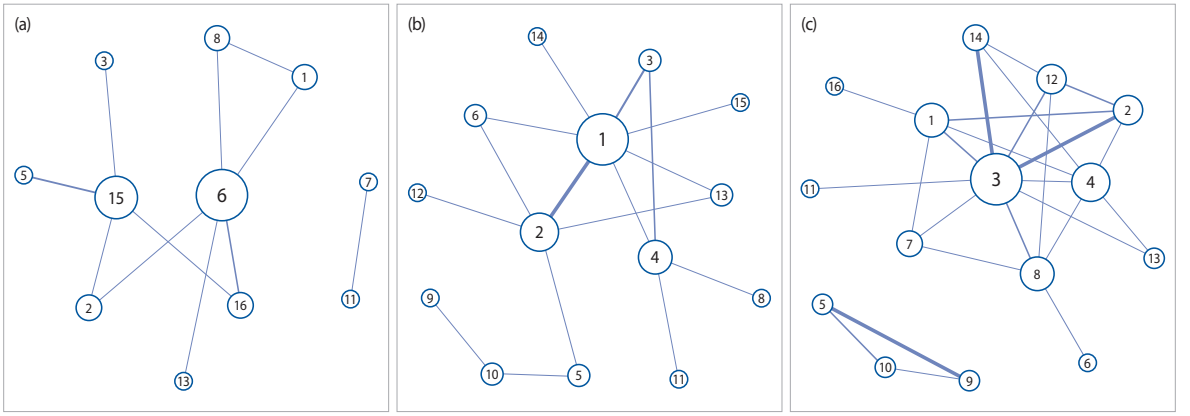

Fig. 6. Evolution of topic networks. (a) Period 1 (2000–2004), (b) period 2 (2005–2009), and (c) period 3 (2010–2014).

were examined in detail. Some overarching topics such as digital library projects, development of information systems/services, the role of metadata and metadata librarians, and evaluation of metadata quality turned out to be prevalent in metadata research, while more specifically focused topics such as semantic web and ontology, and record management are also found to have a significant share.

We also looked at research trends by comparing relative proportions of topics by year. In addition, as a means of gaining a better insight into how the topics have developed over time in connection with one another, we conducted a network analysis based on the distribution of topics over documents. Overall, the results show that, while some core topics more or less have retained their relatively large proportions in metadata literature, many topics exhibit considerable changes in popularity over time. At the same time, connections between topics continue to grow in volume and in diversity. A variety of factors may affect the rise and/or fall of a topic as well as the formation of relationships among topics, but some trends appear to be closely tied to the overall development of the field.

In the early days of metadata research, discussions related to building the infrastructure and some basic mechanisms for discovery and access in digital environments prevailed. Needless to say, development of international or national metadata standards as well as those proposed by communities in different domains constituted a large part of such groundwork, and metadata literature was once flooded with papers examining various aspects of standard development and deployment. Topic 6 (initiatives for standard development), topic 16 (Dublin Core), and topic 15 (XML and encoding standard) fall under this category. Although their importance in the field of metadata is hardly diminished, as the horizon of research has expanded, the proportions of these topics all show a decline later on.

Practitioners and researchers in metadata fields are keen on new technologies or developments in information environment, as shown in the topics peaking at different points in time. For instance, responding to the surge of social tagging tools, papers addressing various approaches to harnessing user generated tags for creating metadata or enhancing metadata-based services appeared soon after. Once a technology or an innovation is widely adopted in the field and its application is tested and reported in the context of metadata, words referring directly to the technology tend to dwindle in research papers. However, the decline in the relative share of a topic may reflect shifts in focus or integrations with other topics. Our network analysis of topic relations demonstrated this point, as explained with the case of topic 13 (metadata harvesting). The focus of discussion on this topic has moved from the harvesting standard itself to the development of repository services and to evaluation of the qualities of harvested metadata.

Metadata research has also been tightly connected to movements in the professional field of LIS. The turn of the trend of topic 8 (cataloging) is related to the development of the new model and standard for cataloging, which triggered the resurge of research interest in the topic in the context of broader metadata issues. Not only has the share of this topic increased, the topic appeared together with various other topics in recent literature, including topic 12 (Linked Data), indicating that the discourse surrounding this topic has extended beyond the traditional boundaries of library catalogs.

As the field has matured, empirical studies assessing the quality of metadata or examining the efficacy of current approaches to creation and use of metadata have emerged. The strong rise of topic 4 (evaluation of metadata quality) and the increased interest in automatic extraction methods (topic 10) showcase this trend. In addition, the recent stream of papers on Linked Data indicate that metadata research is continuously evolving in response to new developments in the global information infrastructure.

Finally, the increasingly dense and diversified connections among topics, as shown in the three sub-period topic networks, testify that topics emerged and evolved over time, not in isolation but in connection with one another. Note that the networks are constructed based on those papers addressing multiple topics together. The extended connections among topics therefore indicate that researchers become more attentive to related topics and attempt to incorporate relevant ideas and discourse into their work. Spawning connections to topic 3 and topic 4 in the recent topic network suggest that the discussions on the role of metadata for various purposes as well as the considerations of quality issues are becoming a cornerstone of metadata research.

Since the dataset in this study was constructed by searching well-known databases in the LIS field, the findings may reflect more the perspective of researchers and practitioners of the LIS field rather than a broader metadata research community. Metadata is an interdisciplinary field, and the current study has included only LIS journals. Therefore, it is necessary to analyze research trends by analyzing metadata research in various fields in future research. In addition, since the research data were analyzed by 2014, it is necessary to carry out the time series analysis continuously including the studies conducted thereafter.

Following up this initial effort to grasp the development of metadata research, we plan to expand the scope of the dataset in a future study to encompass research streams in other related fields.

REFERENCES

Aktas, M. S., Fox, G. C., & Pierce, M. (2010). A federated approach to information management in grids. *International Journal of Web Services Research*, 7(1), 65-98.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77-84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022.

Blessinger, K., & Hrycaj, P. (2010). Highly cited articles in library and information science: An analysis of content and authorship trends. *Library & Information Science Research*, 32(2), 156-162.

Boyd, K., & King, D. (2006). South Carolina goes digital: The creation and development of the University of South Carolina's Digital Activities Department. *OCLC Systems & Services*, 22(3), 179-191.

Cho, J. (2013). The recent trends of information organization research in Japan and Korea. *Library Collections, Acquisitions, and Technical Services*, 37(3-4), 107-117.

Chuttur, M. Y. (2012). An experimental study of metadata training effectiveness on errors in metadata records. *Journal of Library Metadata*, 12(4), 372-395.

Chuttur, M. Y. (2014). Investigating the effect of definitions and best practice guidelines on errors in Dublin Core metadata records. *Journal of Information Science*, 40(1), 28-37.

Conners, D. (2008). A ghost in the catalog: The gradual obsolescence of the main entry. *The Serials Librarian*, 55(1-2), 85-97.

Cumming, K. (2007). Purposeful data: The roles and purposes of recordkeeping metadata. *Records Management Journal*, 17(3), 186-200.

Danskin, A. (2014). Implementing RDA at the British Library. *CILIP Update*, 40-41.

Daud, A. (2012). Using time topic modeling for semantics-based dynamic research interest finding. *Knowledge-Based Systems*, 26, 154-163.

Emery, J. (2007). Ghosts in the machine: The promise of electronic resource management tools. *The Serials Librarian*, 51(3-4), 201-208.

Evans, J. (2007). Evaluating the recordkeeping capabilities of metadata schemas. *Archives and Manuscripts*, 35(2), 56-84.

Evans, J., & Rouche, N. (2004). Utilizing systems development methods in archival systems research: Building a metadata schema registry. *Archival Science*, 4(3-4), 315-334.

Feicheng, M., & Yating, L. (2014). Utilising social network analysis to study the characteristics and functions of the co-occurrence network of online tags. *Online Information Review*, 38(2), 232-247.

Feick, T., Henderson, H., & England, D. (2011). One identifier: Find your oasis with NISO's I2 (institutional identifiers) standard. *The Serials Librarian*, 60(1-4), 213-222.

Feinerer, I., & Hornik, K. (2014). tm: Text Mining Package: A framework for text mining applications within R. Retrieved September 22, 2018 from http://CRAN.R-project.org/package=tm.

Ferris, A. M. (2002). Cataloging internet resources using MARC21 and AACR2: Online training for working catalogers. *Cataloging & Classification Quarterly*, 34(3), 339-353.

Greifeneder, E. (2014, Semptember). *Trends in information behaviour research*. Paper presented at ISIC: the information behaviour conference (part 1), Leeds, United Kingdom.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl 1), 5228-5235.

Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, 40(13), 1-30.

Hall, D., Jurafsky, D., & Manning, C. D. (2008). Studying the history of ideas using topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 363-371). Hawaii: Association for Computational Linguistics.

Han, M. J., & Hswe, P. (2011). The evolving role of the metadata librarian. *Library Resources & Technical Services*, 54(3), 129-141.

Hunter, J. (2003). Working towards MetaUtopia: A survey of current metadata research. *Library Trends*, 52(2), 318-344.

Julien, H., Pecoskie, J. L., & Reed, K. (2011). Trends in information behavior research, 1999-2008: A content analysis. *Library & Information Science Research*, 33(1), 19-24.

Kanellopoulos, D. N., & Kotsiantis, S. B. (2007). Semantic

web: A state of the art survey. *International Review on Computer and Software*, 2(5), 428-442.

Lagace, N., Breeding, M., Romano Reynolds, R., & Han, N. (2013). Everyone's a player: Creation of standards in a fast-paced shared world. *The Serials Librarian*, 64(1-4), 158-166.

Lalitha, P. (2009). Importance of digitization of cultural and heritage materials. *SRELS Journal of Information Management*, 46(3), 249-266.

McCallum, A. (2002). MALLET: A Machine Learning for Language Toolkit. Retrieved September 22, 2018 from http://mallet.cs.umass.edu.

Medeiros, N. (2003). A pioneering spirit: Using administrative metadata to manage electronic resources. *OCLC Systems and Services*, 19(3), 86-88.

Mimno, D., & McCallum, A. (2008). Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *Proceedings of 24th Conference on Uncertainty in Artificial 1 Intelligence* (pp. 411-418). Arlington: AUAI Press.

Mimno, D., McCallum, A., & Mann, G. S. (2006). Bibliometric impact measures leveraging topic analysis. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital libraries* (*JCDL '06*) (pp. 65-74). New York: ACM.

Mullen, A. (2001). GILS metadata initiatives at the state level. *Government Information Quarterly*, 18(3), 167-180.

Palmer, C. L., Zavalina, O. L., & Mustafoff, M. (2007, June). Trends in metadata practices: A longitudinal study of collection federation. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 386-395). New York: ACM.

Park, J. R. (2009). Metadata quality in digital repositories: A survey of the current state of the art. *Cataloging & Classification Quarterly*, 47(3-4), 213-228.

Park, J. R., & Tosaka, Y. (2010). Metadata quality control in digital repositories and collections: Criteria, semantics, and mechanisms. *Cataloging & Classification Quarterly*, 48(8), 696-715.

Patra, S. K., Bhattacharya, P., & Verma, N. (2006). Bibliometric study of literature on bibliometrics. *DESIDOC Journal of Library & Information Technology*, 26(1), 27-32.

Pattuelli, M. C. (2010). Knowledge organization landscape: A content analysis of introductory courses. *Journal of Information Science*, 36(6), 812-822.

Qin, J., & Paling, S. (2001). Converting a controlled vocabulary into an ontology: The case of GEM. *Information Research: An International Electronic Journal*, 6(2). Retrieved September 22, 2018 from http://www.informationr.net/ir/6-2/paper94.html.

R Core Team (2014). *R: A language and environment for statistical computing*. Vienna, Austria: The R Foundation for Statistical Computing.

Schottlaender, B. E. C. (2003). Why metadata? Why me? Why now? *Cataloging and Classification Quarterly*, 36(3-4), 19-29.

Shiri, A. (2003). Digital library research: Current developments and trends. *Library Review*, 52(5), 198-202.

Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* (pp. 63-70). Baltimore: Association for Computational Linguistics.

Symonds, E., & May, C. (2009). Documenting local procedures: The development of standard digitization processes through the Dear Comrade project. *Journal of Library Metadata*, 9(3-4), 305-323.

Wakimoto, J. C. (2009). Scope of the library catalog in times of transition. *Cataloging & Classification Quarterly*, 47(5), 409-426.

Woodley, M. S. (2002). A digital library project on a shoestring. *Library Collections, Acquisitions, and Technical Services*, 26(3), 199-206.

Yagüe, M. I., Maña, A., & Lopez, J. (2005). A metadata-based access control model for web services. *Internet Research*, 15(1), 99-116.

Yeh, J., Chen, C., Sie, S., & Liu, C. (2014). X-System: An extensible digital library system for flexible and multi-purpose contents management. *International Journal of Digital Library Systems*, 4(1), 25-40.

Zeng, M. L., & Qin, J. (2016). Metadata (2nd ed.). Chicago: American Library Association.