Digital Library Perspectives

Institutional repository research 2005-2015: a trend analysis using bibliometrics and text mining

Sujira Ammarukleart, Jeonghyun Kim,

Article information:

Received 31 July 2016
Revised 30 August 2016
Accepted 30 August 2016

Institutional repository research 2005-2015: a trend analysis using bibliometrics and text mining

Sujira Ammarukleart
Department of Information Sciences, University of North Texas, Denton, Texas, USA and Department of Information Science, Faculty of Humanities and Social Sciences, Chiang Mai Rajabhat University, Chiang Mai, Thailand, and

Jeonghyun Kim
Department of Information Sciences, University of North Texas, Denton, Texas, USA

Abstract

Purpose – This paper aims to investigate the longitudinal trends of research in the area of institutional repositories (IR) using bibliometric and text-mining methods.

Design/methodology/approach – The Library and Information Science Abstracts and the Web of Science citation databases were used as data sources. A total of 603 articles published in 109 peer-reviewed journals from 2005 to 2015 were collected and analyzed. The articles were analyzed in terms of publication trends, authorship patterns and keywords and phrases appearing in the article titles and abstracts.

Findings – The study shows that there has been a notable growth trend in research outputs, along with more participation and collaboration among institutes and countries. The study also found significant variability in the topics covered in the literature. In a comparison of the first period of 2005-2010 and the second period of 2011-2015, new research themes and foci, including research data, data management, linked open data, students and student research and an international audience, are observed in the later period.

Originality/value – This paper provides a comprehensive overview of publication, authorship and research themes in the IR research field. It describes the evolution of the intellectual structure of IR as a research field.

Keywords Institutional repositories, Text mining, Bibliometrics, Open access, Research trends, Research themes

Paper type Research paper

Introduction

Institutional repositories (IRs) debuted in the 1990s when college and university libraries began to collect electronic documents, digitize content in their special collections and make that content available via the internet. In 1999, the framework for developing interoperable archives was developed and became the *Open Archive Initiative*. In 2001 and 2002, the development of software tools, such as DSpace and EPrints, created open-source alternatives for IR development, and in 2002 and 2003, public statements on open access, including the *Budapest Open Access Initiative*, the *Bethesda Statement on Open Access Publishing* and the *Berlin Declaration on Open Access*, spurred development and implementation of IRs. In this historical context, IRs emerged as a new strategy and service to leverage open access to scholarship, disseminate the intellectual output of institutional

Digital Library Perspectives
Vol. 33 No. 3, 2017
pp. 264-278

2059-5816
DOI 10.1108/DLP-07-2016-0027

communities and enhance the visibility of the research outputs locally produced. Lynch (2003, p. 2) stated in the *ARL Bimonthly Report 226* that such a repository is "most essentially an organizational commitment to the stewardship of the digital materials, including long-term preservation where appropriate, as well as organization and access or distribution". To date, numerous initiatives have been launched to explore, research and assist in the development of IR solutions.

It is notable that a continuing and dramatic growth in IRs have been achieved over a relatively short period of time. This conclusion is supported by a large number of studies that report the current state and/or track the growth of IRs at regional, international and global levels (Lynch & Lippincott, 2005; van Westrienen & Lynch, 2005; Rieh *et al.*, 2007; Chen & Hsiang, 2009; Kennan & Kingsley, 2009; Fralinger & Bull, 2013; Dubinsky, 2014; Pinfield *et al.*, 2014). In the experimental stage of IR development, only some developed countries launched IR initiatives, including the USA, the UK, Canada and Australia, but since 2010, the number of IRs in other countries in East Asia, South America and Eastern Europe has steadily grown (Pinfield *et al.*, 2014). The growth of IRs also can be inferred from the increasing number of institutions listing their repositories in projects such as the *Directory of Open Access Repositories* (*OpenDOAR*). As of August 2016, the *OpenDOAR* lists a total of 2,702 IRs. The number of countries hosting IRs is 119, and the breakdown by geographic region is 128 in Africa, 252 in Latin America and the Caribbean, 448 in North America, 599 in Asia, 1,212 in Europe and 63 in Oceania[1]. These repositories vary quite significantly in their nature, type of content and software usage.

Accordingly, the academic and professional literature on IRs have increased considerably, and IRs as a field of research has expanded in recent years. Bhardwaj (2014) noted that the number of papers in this research field was quite low until 2005 but has rapidly increased in later years. Researchers and practitioners from the library and information science (LIS) community have become involved in this research field, as have individuals from other disciplines. They have shared their thoughts and experiences through best practice reports, project reports, case studies and research papers. Such significant efforts have contributed to advancing the development, implementation and assessment of IRs. Thus, the IR research field has now reached the maturity stage.

Given these developments, there is a need for a big-picture view of the state of IRs as a research field made possible by an understanding of what core issues have been addressed in the literature and how the literature has evolved in the past decade. This study aims to explore the research trends and development of the field and identify emerging research foci.

Literature review

Various issues related to IRs have become major topics for discussion in the scholarly community via both formal and informal forums. In this discussion, the need to develop a scheme to organize information and enrich taxonomic terminology has gained attention since the early 2010s. Bailey (2010) was the first to focus on classifying IR-related works. The *Institutional Repository Bibliography* was created to index and present selected English-language scholarly textual sources related to IRs. According to the *Institutional Repository Bibliography Version 2*, there are 11 sub-categories of IR-related works, including general issues related to IRs, country and regional IRs, multiple-institution repositories, specific institution repositories digital preservation, library-related issues, metadata, open access policy, research and development projects, research studies and software. This bibliography serves as a useful web-based indexing source for IRs and provides a broader picture of the field through sub-categories without, however, presenting a list of detailed research themes existing within the field.

A number of studies have been carried out to make possible a better understanding of IRs as research field. Using a traditional bibliometric analysis, Bhardwaj (2014) examined a total of 436 articles published in 118 journals from 2001 through 2012. Those papers originated from 68 countries and had authors affiliated with 159 institutions. The study concluded that the growing number of publications, growing citation rates and increasing numbers of contributing institutions and authors detailed in the study are evidence of the field's achieving maturity within the past decade. However, the results could not provide a deeper understanding of the intellectual structure of the field fueling its growth.

In the same year, Cho (2014) mapped out the intellectual structure of the IR research field using a co-word analysis and multidimensional scaling techniques. A total of 204 articles indexed in the SCOPUS database between 1997 and 2012 were included in the study. Out of 564 author keywords in these articles, 32 keywords, which were selected on the basis of their occurring more than four times, were extracted and used to generate a co-occurrence matrix. The study found the main research subjects in relation to IRs included the IR as a significant means of open access, issues about scholarly communication and metadata as an information-retrieval tool. Further, the study identified eight subgroups: metadata, open access, IRs, digital libraries, DSpace, copyright, preservation and semantic webs; among those eight subgroups, open access, digital libraries and DSpace showed a strong correlation with IRs. The study concluded that the other five subgroups, which did not show a correlation with IRs, could be considered as independent research domains within the IR-related research field.

More recently, Stevenson and Zhang (2015) tracked the changing patterns of related topics within the field to understand the breadth and the depth of the development of IRs as a research field. Temporal data mining, multidimensional scaling and parallel coordinate analysis were used in understanding the temporal changes within the field and discovering subject themes which appeared in four time periods: Period I 1992-2001, Period II 2002-2005, Period III 2006-2009 and Period IV 2010-2013. The analysis demonstrated the maturity and development of the field; three meta-categories of IR research, theory and methods, application and management and implementation and technique, along with subject terms appearing in each category, were identified. A consistent pattern in terminology in the theory and methods category was observed, whereas a lack of consistency in regard to terminology and definitions relating to IRs was acknowledged in the application and management category. In addition, the most noticeable changes in subject term use were observed in the implementation and techniques category. Overall, Period III demonstrated the most dramatic changes in research among the three categories.

Methodology

Bibliometrics has been defined as "the study of the quantitative aspects of the production, dissemination, and use of recorded information" (Tague-Sutcliffe, 1992). Various bibliometric indicators have been used in previous studies to explore the features of publication activity, scientific collaboration, authorship characteristics and citation patterns.

Text mining is useful for extracting meaningful, non-trivial patterns or knowledge from a set of semi-structured or unstructured text data, such as full-text documents, e-mails, web posts and HTML text. According to Miner *et al.* (2012), the idea behind text mining is to turn text data into a numerical format for the purpose of performing subsequent analysis.

Data collection

The *Library and Information Science Abstracts* and the *Web of Science* databases were used for the collection of IR literature. These two major databases have served as international abstracting and indexing resources designed for identifying scholarly articles in the LIS

field. They have also been used as data sources in other bibliometric studies (Hung & Zhang, 2012; Han *et al.*, 2014). The terms "institutional repository" and "institutional repositories" were selected as keywords for the initial search in this study. Only these two terms were used as the authors wanted to capture the general theme of IRs expressed in articles' titles, abstracts, subjects and keywords. The focus of this study was articles published in journals written in English. The search was limited to peer-reviewed journal articles to enhance retrieval quality. Thus, other publication types, such as proceedings papers, dissertations, editorial materials and reviews published in journals, were excluded.

The sampling timeframe for this study was the period from 2005 to 2015. This selection was based on the observation that little IR research was carried out before the early 2000s although the number of IRs has steadily grown since the late 1990s. This conclusion is supported by Cho (2014), who found that although IR-related research started in 2000, progress in the field was slow until 2005, a year in which four papers were indexed in SCOPUS. The authors also could identify only eight articles related to IRs published in peer-reviewed journals from 2000-2004. Hence, the year 2005 was determined as a starting point for data collection. After the duplicate articles and non-relevant articles were eliminated, a total of 603 journal articles were retained for analysis.

Data analysis

The data analysis began with the examination of publication trends. Both the number of articles per year and the cumulative number of articles were determined. For the purpose of assessing the hypothesis of IR literature growth saturation, a linear regression model was performed. Further, the number of articles per journal title was counted and rank ordered so that the prolific journals in the field of IRs could be identified.

To identify an authorship pattern, the total number of authors who contributed to IR literature was counted. In addition, two types of collaboration – national collaboration and international collaboration – were reviewed. To reveal the characteristics of and trends in the geographic locations of authors, each author's nationality was observed. The author's nationality was determined based on the country in which his or her institution was located. The total author counting method was selected such that the data of all contributing authors could be captured, i.e. when a country's contribution was being counted, the absolute country counting approach used by Egghe *et al.* (2000) was used. Thus, the countries of all contributing authors were first recorded and then each participating country received one count.

To capture the themes and the conceptual evolution of the field, word analyses were conducted. The titles and abstracts of the articles were used as data sources for text mining. Research article titles represent the first contact a reader has with a potentially fruitful source of information in his or her field of interest; they are regarded as a "means of making visible the internal cognitive structure of a discipline" (Leydesdorff, 1989, p. 221). Article abstracts, which provide brief descriptions of articles' contents, can also be seen as doorways that persuade the reader to select an article. Both titles and abstracts were analyzed using text-mining techniques used in previous studies (Milojevic *et al.*, 2011; Zhang *et al.*, 2012; Assefa & Rorissa, 2013).

In this study, Provalis Research's WordStat, a content-analysis and text-mining tool, was used. The authors first cleaned and preprocessed texts; this included checking spelling, fixing abbreviations/acronyms, removing numbers/punctuation marks/hyphens/stop words, lemmatization, etc. Then the phrase finder function within WordStat was used to classify phrases with a minimum of two words and a maximum of five words into appropriate categories within dictionaries. The authors combined frequently occurring words and phrases extracted from both title and abstract corpuses. It should be emphasized that some phrases and non-specific words appearing in the titles and abstracts that had

limited analysis value were removed. Word frequency analysis was performed across the extracted text corpus so that the most frequently occurring words and phrases could be identified. The authors adopted this technique of word count analysis to identify the predominant themes in texts. Krippendorff (2004, p. 59) argued that the frequency with which a word appears is considered an indicator of "the importance of, attention to, or emphasis on" a particular word or the idea or concept to which it is related.

To track changes in the publication trends, in authorship patterns and in the popularity of words and phrases used in the titles and abstracts, data were examined for the following two consecutive periods: 2005-2010, for which 345 articles were found, and 2011-2015, for which there were 258 articles.

Results

Publication trends

The total number of IR peer-reviewed research articles indexed in the *Library and Information Science Abstracts* and the *Web of Science* from 2005 to 2015 is 603. Figure 1 presents the distribution of the articles. The number of publications on IRs reached its peak in 2006 with a total number of 88 articles published in that year. However, in the following three years, the number of articles per year dropped by 46.59 per cent (from 88 in 2006 to 47 in 2010). Then the number of articles increased to 59 in 2011 and reached 63 in 2014. In 2015, the cumulative number of articles published in the IR field was 603, as presented in Figure 1. Although the annual number of publications has remained relatively static, the cumulative number of publications has increased significantly over time. The cumulative progression is represented by a linear model. The plot of the data revealed a high coefficient of determination in the period 2005-2010 ($r = 0.985$) and 2011-2015 ($r = 0.993$). The publication growth in the later period was even higher compared to the period from 2005 to 2010. It can be predicted that the number of peer-reviewed research articles on the topic of IRs will continue to grow at a high rate in the future. However, the number given for 2015 should be interpreted with caution because it is likely that the databases were not fully updated at the time of data collection.

The 603 articles analyzed in this study were published in 109 journals. Out of the 109 journals, the top 15 journals, which published about 50 per cent of the research output in the field of IRs, are presented in Table I. *Digital Library Perspectives* is at the top with 69 articles, representing about 11 per cent of the total articles, followed by the *Journal of Digital Information*, which published 26 articles. In fact, it is not surprising that those journals ranked as the first and second contributors to the IR literature as the scope of those journals covers a broad range of topics relating to digital

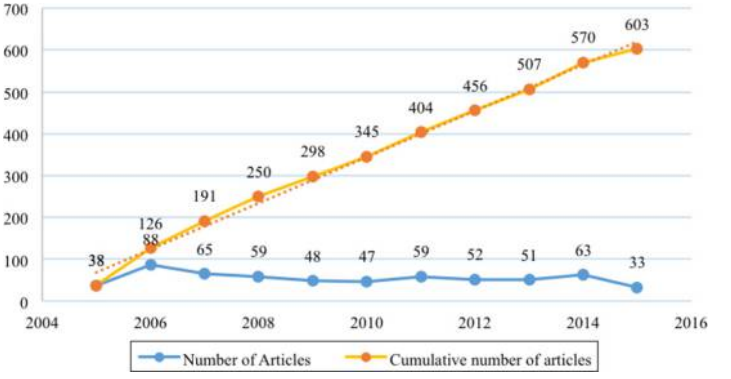

Figure 1.
IR publications from 2005 to 2015

No.	Journal	No. of articles	(%)
1	*Digital Library Perspectives*[a]	69	11.44
2	*Journal of Digital Information*[b]	26	4.31
3	*Library Hi Tech*	26	4.31
4	*Program: Electronic Library and Information Systems*	26	4.31
5	*The Journal of Academic Librarianship*	26	4.31
6	*Serials Review*	21	3.48
7	*The Electronic Library*	20	3.32
8	*The Serials Librarian*	19	3.15
9	*New Review of Information Networking*	15	2.49
10	*Cataloging & Classification Quarterly*	13	2.16
11	*Journal of the Association for Information Science and Technology*[c]	13	2.16
12	*Learned Publishing*	13	2.16
13	*Library Trends*	13	2.16
14	*First Monday*	11	1.82
15	*Online Information Review*	11	1.82

Notes: [a]The journal was formerly entitled as *OCLC Systems and Services*: *International Digital Library Perspectives* from 1993 to 2015; [b]The journal was discontinued in 2013; [c] The journal was formerly entitled as *Journal of the American Society for Information Science and Technology* from 2001 to 2013

Table I.
Top 15 journals

libraries, digital repositories and issues related to digital content and digital information. However, other journals listed in the top 15, such as the *Journal of the Association for Information Science and Technology* and *Library Trends*, cover a broad area of theory and practice in the LIS field; the appearance of IR publications in those journals reveals that IRs are not a specialized topic but rather present a basic challenge for the profession itself. Stevenson and Zhang (2015) state that the more mature the IR research field becomes, the more practitioners and researchers from the LIS community will get involved. From this standpoint, it may be concluded that the maturity of the IR research field has been achieved through the involvement of the LIS community in past decades as shown by the growing body of empirical research and practical experience studies in LIS journals.

When comparing data from the two periods, it is important to note that while *Digital Library Perspectives* was the most active journal, contributing 60 articles in the first period (2005-2010), the *Journal of Academic Librarianship* became the most productive journal in the second period (2011-2015). It is worthy of note that *Digital Library Perspectives* released two special issues on IRs every year from 2007 through 2009, and the *Journal of Academic Librarianship* published a special issue on open access in 2013; these contributions may explain the high publication record in each period.

Authorship pattern

A total of 1,248 authors contributed to the 603 articles, 674 authors during the period of 2005-2010 and 574 authors during the period of 2011-2015, respectively. The number of authors contributing to each article ranged from 1 to 10. The largest percentage of contributions (43 per cent) were by single authors; this is followed by work by two authors with 31, and 26 per cent of articles were contributed by more than three authors. It can be confirmed that collaborative efforts are more common in the field of IRs. Among the co-authored articles, 94 per cent were collaborations by authors in the same country, whereas a very small proportion of the total articles (6 per cent) were collaborations by authors from different countries. This also reflects the long-standing tradition of institution-focused research in the IR field.

A comparison was made between the period of 2005-2010 and the period of 2011-2015, based on the proportion of articles written by more than one author. Between the two periods, the proportion of co-authorship grew from 51 per cent to 66 per cent. The proportion of international collaborations increased from 4 per cent in the first period to 8 per cent in the second period. Most international collaborations were collaborations between two institutions in two different countries.

Authors from a total of 58 countries contributed articles to the IR field from 2005 to 2015. Authors from the USA were responsible for 45 per cent, followed by the UK (13 per cent), India (6 per cent), Australia (5 per cent) and Canada (5 per cent). This study also found that the number of countries participating in research in the field increased from 35 in 2005-2010 to 49 in 2011-2015. This means that the distribution of authors among countries increased greatly. The authorship also spread from developed countries to developing countries. Table II shows the top ten contributing countries for each period. The list for the two periods presents some notable variations. A total of 15 countries were found on the list for the two periods. Only seven countries appear in the top ten lists for both periods.

Terms and topics

A list of words and phrases appearing in more than three articles was generated. It is unsurprising that the phrase "IR" showed the highest number of cases (58 per cent), followed by "repository", "research", "library", "development", "university", institution" and "open access". A wide range of terms identified in this study indicates that various topics have been covered in the IR literature in the past decade. Topics include, but are not limited to, IR policies/procedures/workflow, models/architecture/frameworks, skill sets and training needs of IR staff members, business models and plans, information technology infrastructure, tools and techniques, metadata creation, long-term preservation, copyright issues, IRs for data management, integration and interoperability issues, usage and impact, repository metrics and usability. In particular, the terms identified in this study, such as "project", "initiative", "case study", "strategy", "activity" and "experience", imply that a large number of published articles report primarily on the overview of institutional efforts and on activities, use cases, strategic contexts, best practices, practical considerations and lessons learned in particular institutions. Such articles often describe practical aspects of IR development, implementation, management and assessment. Table III shows the top 100 words/phrases frequently occurring in the titles and abstracts; the "number of cases" is the

Table II.
Top 10 countries

	2005-2010			2011-2015	
Country	No. of papers	(%)	Country	No. of papers	(%)
USA	164	47.54	USA	106	41.09
UK	58	16.81	India	21	8.14
India	17	4.93	UK	18	6.98
Australia	16	4.64	Spain	12	4.65
Canada	13	3.77	Malaysia	11	4.26
Germany	9	2.61	Nigeria	7	2.71
Netherlands	7	2.03	South Africa	7	2.71
Spain	7	2.03	Canada	6	2.33
Hong Kong	5	1.45	New Zealand	6	2.33
Malaysia	5	1.45	Finland	4	1.55
New Zealand	5	1.45	France	4	1.55

Words/phrases	No. of cases	% of cases	Words/phrases	No. of cases	% of cases
Institutional_repository	349	57.88	Case_study	53	8.79
Repository	258	42.79	Search	52	8.62
Research	199	33.00	Community	51	8.46
Library	178	29.52	Document	49	8.13
Development	167	27.69	Country	49	8.13
University	160	26.53	Requirement	48	7.96
Institution	154	25.54	Deposit	46	7.63
Open_access	151	25.04	DSpace	45	7.46
Content	125	20.73	Journal	45	7.46
Information	123	20.40	Preservation	45	7.46
Access	121	20.07	Strategy	43	7.13
Project	116	19.24	Framework	43	7.13
Author	113	18.74	Scholar	43	7.13
Data	112	18.57	Collaboration	42	6.97
Service	109	18.08	Method	41	6.80
Resource	102	16.92	Creation	40	6.63
Researcher	100	16.58	Factor	39	6.47
System	99	16.42	Academic_library	39	6.47
Archive	96	15.92	Database	39	6.47
Role	95	15.75	Thesis	39	6.47
Metadata	90	14.93	Success	38	6.30
Process	89	14.76	Publisher	38	6.30
Challenge	89	14.76	Infrastructure	38	6.30
Survey	87	14.43	Network	38	6.30
User	86	14.26	Solution	38	6.30
Model	83	13.76	Opportunity	38	6.30
Librarian	80	13.27	Student	37	6.14
Digital_repository	77	12.77	Publishing	37	6.14
Tool	77	12.77	Scholarship	36	5.97
Collection	76	12.60	Visibility	36	5.97
Faculty	75	12.44	Dissemination	36	5.97
Literature	74	12.27	Initiative	35	5.80
Digital_library	71	11.77	Self-archiving	34	5.64
Practice	71	11.77	Standard	34	5.64
Policy	70	11.61	Staff	34	5.64
Material	70	11.61	Implication	34	5.64
Benefit	70	11.61	Open_access_repository	33	5.47
Analysis	69	11.44	Organization	33	5.47
Management	66	10.95	UK	33	5.47
Implementation	66	10.95	Quality	32	5.31
Publication	65	10.78	Discipline	32	5.31
Review	63	10.45	Awareness	32	5.31
Knowledge	62	10.28	Record	32	5.31
Design	62	10.28	Science	31	5.14
Impact	62	10.28	Research_output	30	4.98
Technology	62	10.28	Cost	30	4.98
Scholarly_communication	61	10.12	Digital_collection	29	4.81
Software	60	9.95	Education	29	4.81
Activity	55	9.12	Plan	29	4.81
Web	53	8.79	Evaluation	28	4.64

Table III.
Top 100 words/phrases

number of articles in which a word/phrase appears, and the "percentage of cases" is the percentage of articles in which the word/phrase appears.

To further analyze the evolution of the IR field over the past decade, the authors considered the differences in the percentage of cases in which a word/phrase appears between the periods of 2005-2010 and 2011-2015. As the two periods did not yield an equal number of publications, the percentage of articles where a term appears was chosen rather than the number of times that the term occurs in the corpus of texts. For comparative analysis, the words and phrases with two or more case occurrences were selected as a data sample and then the percentages of articles where they appear were compared. Table IV

Words/Phrases	2005-2010	2011-2015	Words/Phrases	2005-2010	2011-2015
Data curation and management			Technology/Tool		
Data	14.49	24.03	Technical infrastructure	2.03	1.16
Data set	2.90	5.43	Hardware	1.74	1.16
Scientific data	0.58	1.16	Software	12.17	6.98
Data curation	0.87	1.94	Open source software	4.06	3.88
Data management	0.87	3.49	DSpace	8.12	6.59
Data management plan	0	0.78	Eprints	5.51	1.16
Data management support	0	0.78	CONTENTdm	0.87	0.78
Data curation	0.87	1.94	Fedora	1.74	2.71
Digital curation	0	0.78	Cloud Computing	0.58	1.16
Digital preservation	3.48	3.88	Metadata		
Data sharing			Metadata	16.52	12.79
Open data	0	0.78	Metadata standard	0.87	0
Data reuse	0	1.16	Dublin Core	2.03	2.71
Data sharing	0.29	1.55	Metadata creation	0.87	0.78
Data sharing behavior	0	0.78	Metadata analysis	0.87	0.78
Data sharing policy	0	0.78	Linked data	0	1.55
Data repository			Content		
Data repository	0	5.04	Audio	0.58	1.55
Scientific data repository	0	0.78	Video	1.45	1.94
Disciplinary repository	0	1.55	Born digital	0.87	0.78
Subject repository	1.16	2.71	Journal article	2.03	1.55
Research			Conference paper	1.16	1.55
Researcher	12.46	22.09	Teaching material	0.87	0.39
Research activity	0.87	1.55	Dissertation	2.32	5.43
Research data	2.61	3.88	Thesis	4.35	9.30
Research service	0	0.78	Gray literature	0.87	1.55
Research support	0	0.78	Newspaper	0.29	0.78
Research information system	0.29	1.16	Stakeholder		
Country			Faculty	12.75	12.02
Australia	2.03	1.16	Student	5.22	7.36
Canada	1.16	1.55	Graduate student	0.29	1.94
UK	6.67	3.38	Postgraduate student	0	0.78
China	0	1.94	Reference librarian	1.74	0
India	2.90	6.59	Subject librarian	0	0.78
Indonesia	0	1.16	Library administrator	1.16	0
Japan	0	0.78	Repository administrator	0.58	1.16
Malaysia	0.87	1.94	Repository manager	2.03	2.33
Nigeria	0	2.71	Funding agency	0.29	2.71
South Africa	0	2.71	Academic institution	3.77	1.16

Table IV.
Growth of selected words and phrases

lists selective words/phrases that have a large relative increase and decrease in the percentage of cases between the two periods.

In the IR literature, terms related to data curation and management, such as "data", "research data", "scientific data", "data collection", "data curation", "data management" and "research data management", appeared more often in the 2011-2015 period. The terms "data management plan", "data management support" and "digital curation" appeared only in the second period. The emergence of those words and phrases may be due to government agencies' commitment to a long-term strategy for data resource provision and development of data policies[2] although the activities of managing and promoting the use of data and open access to research data have become a major concern in all scientific fields across the globe since the early 2000s. In light of this development, the institutional management of research data has become a major research agenda of digital curation, and there has been a call for a new role for IRs in supporting faculty researchers' data curation and management. Thus, scalable efforts within IRs have been made to further encourage participation in data curation.

Like data management mandated by funding agencies and journals, academic data sharing has received a considerable increase in attention in recent years. Accordingly, the data sharing behavior of particular groups of scientists or researchers has become a topic of interest in the IR field. This is because academic institutions have been challenged to provide well-designed IRs for their researchers to us in sharing data, and they find professional value in sharing data in IRs. This is confirmed by the finding of this study that the terms "data sharing behavior", "data sharing policy" and "data reuse" also appeared only in the second period. Also relevant to this point, the integration of IRs into the wider data repository ecology that includes research information systems and disciplinary data repositories was observed; this is evidenced by the growth of "data repository", "scientific data repository", "subject repository", "disciplinary repository", "Dryad data repository" and "geospatial data repository".

Additionally, there has been a need to secure data supporting academic research, and more emphasis is now placed on offering IRs as a research support service. Research-related terms, including "research support", "research service" and "research information system", showed an overall upward trend. A number of studies have tried to determine whether IRs can support research by making research outputs more visible and accessible (Schopfel *et al.*, 2014; Lee *et al.*, 2015). In fact, the term "research" has been consistently increasing in popularity not only in the IR field but also in other scientific fields. In tracking the shifts in subject emphasis in the broader LIS literature during the first 100 years, Lariviere *et al.* (2012) found that the term "research" has also consistently occurred in the LIS literature since the 1980s due to the growing interest in empirical investigations in the field.

Development and management of technical infrastructure for IRs has been a significant hurdle for most institutions. Compared with the 2011-2015 period, the terms representing IR technical infrastructure were prominent in the period of 2005-2010; there are noticeable instances of the use of repository technology vocabulary, such as "open source software", "open source digital library software", "DSpace", "EPrints", "Digital Commons" and "Fedora (Commons)". Special issues on the topic appearing in two journals support this conclusion: *Library Hi Tech* published a special issue on open source software in 2005 and *New Review of Information Networking* released its special issue on repository architecture in 2009. This demonstrates that early IR works focused more on adopting and applying the repository platform, hardware and software as part of the effort for successful IR implementation. This is in line with the finding of Stevenson and Zhang (2015), who noted that research related to

the technological advancement of IRs was commonly found in 2006-2009. Such efforts have continued as new technologies have emerged. In the 2011-2015 period, libraries started to plan and perform long-term digital preservation activities. The staggering growth of "cloud computing" is reflected, and new services, such as "DuraCloud", which leverages existing cloud infrastructure to enable durability and access to digital content, have been mentioned in the IR literature.

Metadata is a core component in the creation of repositories. Most institutions have implemented policies for metadata schema and authorized metadata creators for their IRs. The application, harvesting and interoperability of IR metadata have been studied in the IR field. Terms under the category of metadata, including "metadata", "repository metadata", "metadata creation", "metadata standard/schema", "Dublin Core", "MARC", "metadata element", "metadata record", "authority control", "metadata harvest", "metadata quality" and "metadata analysis", were fairly popular in the first period. It should be noted that the *Cataloging and Classification Quarterly* published a special issue on metadata and open access repositories in 2009. In the second period, 2011-2015, the term "linked data/linked open data" made a leap forward; this implies that the value of making data (a repository's content) available as linked open data has of late been recognized in the IR field.

In its early days, the use of IRs emphasized the deposit of textual research output. But the scope of repository content has gradually extended to cover various formats. It is now apparent that more objects and resources in various file formats are being handled in IRs. In our study, the terms "audio", "video" and "data" were more common in the second period. "File format" itself became a critical research topic in the IR field as deciding on an appropriate file format is an important issue of digital preservation. A wide range of different materials hosted in IRs is detected in the comparative analysis. Accordingly, the terms "conference paper", "teaching material" and "learning resource" often appeared in the first period, whereas the terms "thesis", "dissertation", "undergraduate thesis", "electronic thesis and dissertation (ETD)" and "grey literature" appeared in the second period. It seems reasonable to suggest that the early IR works were predominantly devoted to faculty members' published and unpublished works but that student research has become a significant and rapidly growing segment of content available in IRs. In particular, many articles on the institutional experience of implementing IRs for ETDs have been published by authors in many developing countries (Sheeja, 2012; Ezema & Ugwu, 2013; Hakimjavadi & Masrek, 2013).

In the same vein, early research on IRs primarily focused on issues related to faculty scholarship; the terms "faculty contribution", "faculty participation" and "faculty scholarship" were more frequently used in the period of 2005-2010. More generic terms used to describe the personnel who undertake research, including "researcher", "academic researcher" and "scientific researcher" were used in the period of 2011-2015. Terms representing other author/user groups, such as "student", "undergraduate student", "graduate student" and "postgraduate student", commonly appeared in the period of 2011-2015 as well. This is likely related to the field's growing interest in students as content creators and contributors. In addition, this also confirms that non-research-intensive institutions that emphasize student work have been late adopters of IRs and open access initiatives (Kocken & Wical, 2013) and that IRs in those small colleges and universities are growing.

The geographic focus of IRs has been extended throughout the past decade. While early IRs began appearing in the USA and UK in the 1990s, academic libraries in China began implementing them in the early 2000s (Hu *et al.*, 2013) and Middle Eastern countries began

exploring the possibilities in the 2010s (Ahmed & Al-Baridi, 2012). Along with this trend, IR research has become more pervasive at a global level. This is confirmed by a remarkable growth in the number of terms found by the current study referring to country names or continent names. In particular, the authors observed continent names, including "Africa", "Asia" and "Latin America", as well as country names, including "China", "India", "Japan", "Indonesia", "Malaysia", "Nigeria" and "Spain", which exhibited a remarkable increase in use in the period of 2011-2015. This finding is consistent with our observation regarding the geographical distribution of authors, which was reported in the *Authorship Pattern* section. It also demonstrates that the number of institutions worldwide actively engaged in the implementation of their own IRs has increased. A global shift was also noted in the presence of terms such as "global visibility", "international visibility" and "global accessibilities"; these terms were frequently used by institutions that had recently pioneered IR initiatives although the terms "visibility", "accessibility", "discoverability" and "web presence" appeared consistently in the IR literature. It should be noted that the term "USA" was not observed in our data set because many authors in the USA who conducted studies at the organizational level rarely mentioned their country name in either the titles or abstracts of the articles.

Conclusion

The primary purpose of this study has been to identify longitudinal research themes and trends in the area of IRs. The authors have presented the findings of a bibliometric approach to describe the field's publication landscape and growth, as well as the authorship types and distribution. The authors have also presented the findings of a text-mining approach to identify the research themes and foci. In particular, our study focuses on the extraction of domain knowledge relating to the field as it appears in taxonomies used in articles published in peer-reviewed journals because such taxonomies better reflect the advancement of any professional and scientific field, including research on IRs.

While IRs emerged in the 1990s, our study confirms that IR research did not flourish until 2005. The number of publications has dramatically increased due to the growing needs and perceived benefits of open access publishing, increased emphasis on faculty self-archiving in IRs and funding agencies' data sharing/management mandates. We may assert that the maturity of the IR research field has been advanced by the involvement of the LIS community in the past decade, a conclusion supported by the growing body of empirical research studies published in a large number of LIS journals. In particular, the emerging research trends and themes identified in our study may provide useful information for both researchers and practitioners in the field in setting guidelines and a direction for future research.

Although the study provides important insights on the evolution of the research themes, a few limitations deserve consideration. This study examined peer-reviewed journal articles published from 2005 to 2015 indexed in the *Library and Information Science Abstracts* and *Web of Science*. Although these two databases are considered major index databases in the LIS field, some articles related to IRs published in other peer-reviewed journals might have been excluded. As this study included only articles written in English, our data set might be biased toward articles published in English-speaking countries. Additionally, the study did not cover other types of publications such as books, conference papers, theses and dissertations. Those publications can also be used as useful indicators for tracking the literature growth and identifying research trends and themes within the IR research field.

Notes

1. Countries were grouped according to regions outlined in the *United Nations Geoscheme* (The United Nations Statistics Division, 2013).
2. For instance, the National Science Foundation's Data Management Plan requirement took effect in January 2011.

References

Ahmed, S. and Al-Baridi, S (2012), "An overview of institutional repository developments in the Arabian Gulf region", *OCLC Systems and Services: International Digital Library Perspectives*, Vol. 28 No. 2, pp. 79-89, doi: http://dx.doi.org/10.1108/10650751211236613.

Assefa, S.G. and Rorissa, A. (2013), "A bibliometric mapping of the structure of STEM education using co-word analysis", *Journal of the Association for Information Science and Technology*, Vol. 64 No. 12, pp. 2513-2536, doi: 10.1002/asi.22917.

Bailey, C.W. (2010), "Institutional repository bibliography, Version 2", available at: http://hdl.handle.net/10760/14515 (accessed 12 July 2016).

Bhardwaj, R.K. (2014), "Institutional repository literature: a bibliometric analysis", *Science & Technology Libraries*, Vol. 33 No. 2, pp. 285-302, doi: 10.1080/0194262X.2014.906018.

Chen, K. and Hsiang, J. (2009), "The unique approach to institutional repository: practice of national Taiwan university", *The Electronic Library*, Vol. 27 No. 2, pp. 204 -221, doi: http://dx.doi.org/10.1108/02640470910947566.

Cho, J. (2014), "Intellectual structure of the institutional repository field: a co-word analysis", *Journal of Information Science*, Vol. 40 No. 3, pp. 386-397, doi: 10.1177/0165551514524686.

Dubinsky, E. (2014), "A current snapshot of institutional repositories: growth rate, disciplinary content and faculty contributions", *Journal of Librarianship and Scholarly Communication*, Vol. 2 No. 3, doi: http://doi.org/10.7710/2162-3309.1167.

Egghe, L., Rousseau, R. and Van Hooydonk, G. (2000), "Methods for accrediting publications to authors or countries: consequences for evaluation studies", *Journal of the American Society for Information Science*, Vol. 51 No. 2, pp. 145-157, doi: 10.1002/(SICI)1097-4571(2000)51:2<145::AID-ASI6>3.0.CO;2-9.

Ezema, I.J. and Ugwu, C.I. (2013), "Electronic theses and dissertations in Nigeria university libraries: status, challenges and strategies", *The Electronic Library*, Vol. 31 No. 4, pp. 493 -507, doi: http://dx.doi.org/10.1108/EL-08-2011-0118.

Fralinger, L. and Bull, J. (2013), "Measuring the international usage of US institutional repositories", *OCLC Systems & Services: International Digital Library Perspectives*, Vol. 29 No. 3, pp. 134-150, doi: 10.1108/OCLC-10-2012-0039.

Han, P., Shi, J., Li, X., Wang, D., Shen, S. and Su, X. (2014), "International collaboration in LIS: global trends and networks at the country and institutional level", *Scientometrics*, Vol. 98 No. 1, pp. 53-72, doi: 10.1007/s11192-013-1146-x.

Hakimjavadi, H. and Masrek, M.N. (2013), "Evaluation of interoperability protocols in repositories of electronic theses and dissertations", *Program: Electronic Library and Information Systems*, Vol. 47 No. 1, pp. 34-59. No

Hu, D., Luo, A. and Liu, H. (2013), "Open access in China and its effect on academic libraries", *Journal of Academic Librarianship*, Vol. 39 No. 1, pp. 110-112, doi: 10.1016/j.acalib.2012.11.009.

Hung, J. and Zhang, K. (2012), "Examining mobile learning trends 2003-2008: a categorical meta-trend analysis using text mining techniques", *Journal of Computing in Higher Education*, Vol. 24 No. 1, pp. 1-17, doi: 10.1007/s12528-011-9044-9.

Kennan, M.A. and Kingsley, D.A. (2009), "The state of the nation: a snapshot of Australian institutional repositories", *First Monday*, Vol. 14 No. 2, doi: http://dx.doi.org/10.5210/fm.v14i2.2282.

Kocken, G.J. and Wical, S.H. (2013), "I've never heard of it before: awareness of open access at a small liberal arts university", *Behavioral & Social Sciences Librarian*, Vol. 32 No. 3, pp. 140-154, doi: 10.1080/01639269.2013.817876.

Krippendorff, K. (2004), *Content Analysis: An Introduction to Its Methodology*, 2nd ed., Sage, Thousand Oaks, CA.

Lariviere, V., Sugimoto, C.R. and Cronin, B. (2012), "A bibliometric chronicling of library and information science's first hundred years", *Journal of the American Society for Information Science and Technology*, Vol. 63 No. 5, pp. 997-2012, doi: 10.1002/asi.22645.

Lee, J., Burnett, G., Vandegrift, M., Baeg, J.H. and Morris, R. (2015), "Availability and accessibility in an open access institutional repository: a case study", *Information Research*, Vol. 20 No. 1, available at: http://InformationR.net/ir/20-1/paper661.html (accessed 8 July 2016).

Leydesdorff, L. (1989), "Words and co-words as indicators of intellectual organization", *Research Policy*, Vol. 18, pp. 209-223, doi: 10.1016/0048-7333(89)90016-4.

Lynch, C.A. (2003), "Institutional repository: essential infrastructure for scholarship in digital age", *ARL Bimonthly Report*, Vol. 226, pp. 1-7.available at: www.arl.org/storage/documents/publications/arl-br-226.pdf (accessed 12 March 2016).

Lynch, C.A. and Lippincott, J. (2005), "Institutional repository deployment in the united states as of early 2005", *D-Lib Magazine*, Vol. 11 No. 9, available at: www.dlib.org/dlib/september05/lynch/09lynch.html (accessed 8 July 2016).

Milojevic, S., Sugimoto, C.R., Yan, E. and Ding, Y. (2011), "The cognitive structure of library and informationscience: analysis of article title words", *Journal of the Association for Information Science and Technology*, Vol. 62 No. 10, pp. 1933-1953, doi: 10.1002/asi.21602.

Miner, G., Elder, J. and Hill, T. (2012), "History of text mining", in Miner, G. (Ed.), *Practical text Mining and Statistical Analysis of Non-Structured Text Data Applications*, Academic Press, Saint Louis, MO, pp. 3-28.

Pinfield, S., Salter, J., Bath, P.A., Hubbard, B., Millington, P., Anders, J.H.S. and Hussain, A. (2014), "Open-access repositories worldwide, 2005-2012: past growth, current characteristics, and future possibility", *Journal of the Association for Information Science and Technology*, Vol. 65 No. 12, pp. 2404-2421, doi: 10.1002/asi.23131.

Rieh, S.Y., Markey, K., Jean, B., Yakel, E. and Kim, J. (2007), "Census of institutional repositories in the US: a comparison across institutions at different stages of IR development", *D-Lib Magazine*, Vol. 13 Nos 11/12, available at: www.dlib.org/dlib/november07/rieh/11rieh.html (accessed 8 July 2016).

Schopfel, J., Chaudiron, S., Jacquemin, B., Prost, H., Severo, M. and Thiault, F. (2014), "Open access to research data in electronic theses and dissertations: an overview", *Library Hi Tech*, Vol. 32 No. 4, pp. 612 -627, doi: 10.1108/LHT-06-2014-0058.

Sheeja, N.K. (2012), "Knowledge management and open access e-theses: Indian initiatives", *Library Review*, Vol. 61 No. 6, pp. 418-427, doi: http://dx.doi.org/10.1108/00242531211284339.

Stevenson, J.A. and Zhang, J. (2015), "A temporal analysis of institutional repository research", *Scientometrics*, Vol. 105 No. 3, pp. 1491-1525, doi: 10.1007/s11192-015-1728-x.

Tague-Sutcliffe, J. (1992), "An introduction to informetrics", *Information Processing & Management*, Vol. 28 No. 1, pp. 1-3, doi: 10.1016/0306-4573(92)90087-G.

The United. Nations Statistics Division (2013), "Composition of macro geographical (continental) regions, geographical sub-regions, and selected economic and other groupings", available at http://unstats.un.org/unsd/methods/m49/m49regin.htm (accessed 20 July 2016).

van Westrienen, G. and Lynch, C.A. (2005), "Academic institutional repositories: deployment status in 13 nations as the mid 2005", *D-Lib Magazine*, Vol. 11 No. 9, available at: www.dlib.org/dlib/september05/westrienen/09westrienen.html (accessed 8 July 2016).

Zhang, J., Xie, J., Hou, W., Tu, X., Xu, J., Song, F., Wang, Z. and Lu, Z. (2012), "Mapping the knowledge structure of research on patient adherence: knowledge domain visualization based co-word analysis and social network analysis", *PLoS ONE*, Vol. 7 No. 4, pp. 1-7, doi: 10.1371/journal.pone.0034497.

Further reading

Milojevic, S. (2012), "Multidisciplinary cognitive content of nanoscience and nanotechnology", *Journal of Nanoparticle Research*, Vol. 14 No. 1, pp. 1-28, doi: 10.1007/s11051-011-0685-4.

Roberts, C.W. (2000), "A conceptual framework for quantitative text analysis", *Quality and Quantity*, Vol. 34, pp. 259-274, doi: 10.1023/A:1004780007748.

Corresponding author

Sujira Ammarukleart can be contacted at: SujiraAmmarukleart@my.unt.edu and mesujira@gmail.com