Afr. J. Lib. Arch. & Inf. Sc. Vol. 30, No. 1 (April 2020) 49-60

Big Data Research Outputs in the Library and Information Science: South African's Contribution using Bibliometric Study of Knowledge Production

Patrick Ajibade and Stephen M. Mutula
Information Studies Programme
University of KwaZulu-Natal
Pietermaritzburg, South Africa
ajibadep1@ukzn.ac.za
mutulas@uk.zn.ac

Abstract

The focus of this study was to evaluate research production in Library and Information Science (LIS) on big data and South African's contribution from 1992-2019. As advancement in technological innovation is changing the methods of digital collection development and dissemination of information in the fourth industrial revolution, big data technology will be reshaping library management systems through big data. Big data is defined as information overload due to the volume, varieties, velocity and veracity of the data which must be processed to get value. It is also useful information for efficient decision making or business intelligence. The data collection methods utilised bibliometric analysis as an intuitive approach to map research focus in big data and LIS contribution, by visualising the outputs using data harvesting capability of Web of Knowledge to export titles, authors, abstract, all keywords, citations, journal sources and bibliographies for further analysis. We performed bibliometric coupling, co-citation analysis, with a total dataset (n = 8,415), h-index =104, and an average citation per output (ACP=97). The findings showed that the LIS scholars contributions were very low (h-index = 29) and (ACI =15.47), and the USA (n=112,) China (n=45) and India (n=25) were the top leading countries in LIS and big data. The contribution of South Africa was very low (n=4). This research underscores that LIS big data contribution is very important for archiving and providing information services to manage petabytes data and information with automated controlled index terms and big data metadata management.

Keywords*:* Data Archiving, Bibliometrics, Big Data, Controlled Index Terms, Metadata Management, Faceted Classification, Web Archiving

Introduction

The proliferation of information technology (IT) has created another challenge of information explosion known as big data. The availability of the Internet and mobile technology infrastructure is set to change how future library services are rendered perpetually. This will require libraries to respond to an uncontrollable growing speed of data accumulation. Currently, libraries are facing inevitable shockwave owing to informatisation of knowledge which library services must respond to, using custom-made library applications (Weihong et al. 2012). Library and Information Science (LIS) has a huge role to play in this era of big data. Part of such a role involves how to manage big data, data processing and classification and big data archiving. It will require specialised technologies (hard and software) and training to deploy embryonic technologies to curate, manage and archive big data for research and other information services. As recently reported, the LIS challenges are also complex, not only in how to handle the high volume of big data (Gulgec, Shahidi and Matarazzo, 2017) but also in the ability to create agile metadata management. Therefore, most libraries will need to go beyond the traditional archiving practices into advanced practices which might require technological integration for big data archiving.

The significance of big data in improving and making libraries more agile has not been fully explored. To underscore the importance of a paradigm shift in the LIS field due to technological advancement and its alignment in the profession, the UK national archives and the Netherlands national library are now involved in web archiving (Di Pretoro and Geeraert, 2019). Web crawling is one of the most commonly adopted methods to harvest information on the websites for archiving. It is stated that most national libraries and archives are collaborating to work on web archiving (Di Pretoro and Geeraert, 2019). While web crawling cannot be equated with big data mining, it is a commendable effort to engage in web archiving, which is beyond the traditional archival practices. Owing to the complexity of technical know-how that is required to undertake a big data project, LIS should revisit the curriculum and training processes. The curriculum could be adjusted to teach basic big data metadata management and cloud services applications in the LIS profession. A previous study has expressed the importance of embedding big data training in the library and information science discipline (Munshi, 2016).

The library's collection development principle applies to managing big data, except that in big data, there is a problem of how to manage speed and a large volume of such data. The librarians' abstracting and indexing skills and reference management prowess are central to big data management. Librarians can be equipped with basic skill needed to use big data technologies and software for enhancing information services, faceted classification, collection development, and building controlled index terms and ontology. This paper conducted a bibliometric analysis of libraries' current alignment of big data technology with the Library and Information Sciences (LIS) based on the research outputs in big data.

Undoubtedly, the LIS profession is undergoing a "big data paradigm shift" because of the proliferation of information technology such as the Internet of Things (IoT), cloud computing and big data. The "big data" has also brought much confusion even within the LIS profession. For example, an article published in *Library Review* attempted to propose a thorough definition of big data as "information asset with high volume, velocity and variety which requires specific analysis methods and technology to transform the value" (De Mauro, Greco and Grimaldi (2016) of such large dataset. In contrast to this definition, big data only becomes useful information when the business intelligence has been extracted due to processing, mapping and reduction of either petabytes or zettabytes of data into manageable information asset that can help organisations or libraries make informed decisions. Besides the volume, variety and velocity, one of the characteristics of big data is the veracity and value. The veracity denotes the quality of a dataset (Jeble, Kumai and Patil, 2018) which refers to the factual and accuracy of the data, and the values, which is the business intelligence that will be derived from processing such huge amount of data for decision making.

Literature Review

The importance of big data in various research domain is growing, such as big data and cyber-physical and social systems (Wang et al., 2018), Internet of things (IoT) and bigdata (Sun et al., 2018) and in the LIS field, an assessment of data analytics from the prism of bigdata has been linked with LMS and policies (Chen et al., 2015). But the question remains how do you define big data architecture (Demchenko, De Laat, and Membrey, 2014), without exacerbating inherent confusion of what is the actual characteristics of big data? This study defined five major characteristics of big data with the usual "5 Vs" Volume, Velocity, Variety, Veracity and Value. A massive data (5 gigabytes) which is volume, in different format (variety), but, such data size which does not require hyper computational speed processing (velocity), to generate or perform data collection, irrespective of the fact that it is valuable cannot be referred to as a big data. Therefore, big data required deployment and uses of a supercomputer that is the capability of performing approximately zettabyte (1 trillion gigabytes) of calculations per nanosecond. The library function must be exposed to advances in technology such as big data (Wang, Xu, Chen, and Chen, 2016), because of the crucial role the libraries play in information organisation and management. Such role might include using big data for curating materials (Teets and Goldner, 2013), as it seems libraries are not catching up with opportunities big data advancement present, albeit its challenges in data processing, collection management and storing etc. (Golub and Hansson, 2017; Shan and Gang, 2013). Yet, the big data can significantly improve libraries' innovation in service delivery (Cuifeng, 2013; Simoviæ, 2018).

Big Data

One of the reasons the library and information profession must take significant interests in big data hinges on the fact that the profession is an information society whose currency of transactions is aggregated data. However, despite the enormous potential of big data to enrich the information society due to its value, this data must be mined, processed and reduced to usable quantity for decision making as business intelligence. By definition, big data presents a huge challenge to the LIS professions due to big data five Vs namely: size, velocity, volume, variety, veracity, and how to authenticate its veracity to derive the intended values. What makes big data are the five "Vs" characteristics, howbeit, big data is multidisciplinary (Hu and Zhang, 2017) and the LIS profession role is particularly vital to achieving knowledge organisation of big data. For example, Aydin, Akkineni and Angryk's (2016) study examined the method of modelling and indexing philosophy and physics of space and time trajectory (spatio-temporal) data, not just in a relational database, but in a non-relational repository. Their study concluded that using the indexing structure and data model has advantages. Therefore, libraries could combine their knowledge of building controlled index terms with the mapping of big data to develop an algorithm to handle automatic big data indexing.

Big Data Technologies and LIS Skills Requirements

The ability of the academic libraries to use their technical prowess in combination with software such as Python, JavaScript and R is vital for handling big data. A librarians' ability to use JavaScript functions to (map, filter, and reduce) big data will assist the library to have manageable data needed for knowledge organisation. The question remains whether South African LIS professionals are ready to explore how High-Performance Computing (HPC) and parallel programming or application could speed up the libraries' ability to collect and process zettabytes of data. For example, the Centre for High-Performance Computing in Cape Town usually allocates 24 nodes per users, who are registered under a principal research leader in a university. Findings indicated that parallel programming would substantially increase the ability to process data as the HPC take advantage of the distributed memory of the system (Yildirim, Ozdogan and Watson, 2016). De Mauro Greco and Grimaldi (2016) indicate some of the impact of big data in the LIS profession in terms of collection and organisation of information. Nevertheless, there are specific and complex aspects of big data that the LIS professional must focus attention on as the profession cannot afford to adopt a "catch me if you can" approach to big data while the librarians are expected to deploy software in processing information resources in this era of big data. This study was conducted using bibliometrics to analyse big data outputs within the domain of library and information science in the Web of Science and Scopus.

Problem Statement

The growth of databases has become exponential to the point of the term big data being used (Patel, Birla and Nair, 2012) because a vast amount of the world information is stored in the databases. However, the problem of the exponential growth of databases has not been solved even with the introduction of big data. For example, there are relational and non-relational, structured query language (SQL) and non-structured (NoSQL) databases that host data from different provenances/sources, formats and complex information architecture. Therefore, librarians must apply their skills to ensure knowledge organisation by processing petabytes volume of data with such velocity, veracity and variety that must be harvested from various provenances. Although technology such as Hadoop may help to solve some of the problems of storing big data compared to traditional or legacy data storing system, Hadoop in itself does not resolve big data metadata management issues such as controlled index terms, faceted classification regarding big data without the library and information professional's prowess.

Arguably, the librarians have to use their data management knowledge and apply such relevant skills in big data metadata management. This might require the ability to programme software to handle automatic indexing of curated data. Furthermore, this paper argues that bibliometric analysis could be useful in identifying indexed terms for knowledge organisation (Hjorland, 2013). As such, the ability of LIS professionals to manage big data metadata is vital for the project owners to derive maximum benefit, which is expected from processing big data. For example, one of the reasons Google search engines is so effective is because of extensive resources descriptions of libraries and metadata management. Search queries results are a result of already developed controlled vocabularies, faceted classification algorithms and controlled index terms, all of which are essential for search result accuracy. Controlled vocabularies are essential owing to different research fields (computer science, engineering, social science, etc.) collaborating and working on big data.

Resulting from the challenges mentioned above, librarians will require training/retraining to handle algorithms and agile software that can handle semi or automatic indexing based on predefined controlled index terms. Christensen (2017) suggests that the systemic indexing will use previous search terms to handle iteration of subsequent indexing processes. Furthermore, a study in Amsterdam indicated that metadata is now the regular currency to pay for communication (van Dijck, 2014), more importantly, due to unique problems presented by big data because of the volume, veracity and varieties of the information from different sources/provenances and the nature of the information architecture. This study carried out this bibliometric analysis to map LIS big data research outputs to facilitate access to, and efficient use of a large amount (yottabytes) of data which requires teraflops of computational processing power to analyse. The bibliometric analysis is a useful approach to measure knowledge production in a field of study or particular subjects. This approach has been used by other scholars to analyse indicators such as authorship, collaboration and publication trends and sources of the journal (Cobo Serrano, 2018). Similar studies on big data in medical science have been carried out.

The main aim of this study was to conduct a bibliometric analysis of big data outputs in the domain of library and information science by mapping trends, and growth of scholarly contributions of the LIS scholars, as well as ascertain the contribution of South Africa to big data.

Objectives of the Study

- To assess global big data research trends based on the countries' outputs.
- To find out the library and information science research productivity on Big Data.
- To find out the area of LIS focus on Big Data and its implication for future research trajectories.
- To find out South African scholars' contributions and outputs on LIS and Big Data.

Research Methodology

The searches were limited to four databases on the Web of Science (WoS) Knowledge repository Core Collection. The data was indexed in the Science Citation Index Expanded (SCI-EXPANDED), Social Sciences Citation Index (SSCI), Arts and Humanities Citation Index (A and HCI) from 1975-2019, and the Emerging Sources Citation Index (ESCI) from 2015-2019. The data presented was based on the global output across all fields of studies/research areas in the Web of Science. The composition of the extracted LIS outputs types dataset was; articles (*n* = 221), editorial material (*n* = 51), book review (*n* = 33), review (*n* =12), and conference proceedings (*n* =4). The title was used as the field tag for the search strategy and separated our search terms with a Boolean operator "OR", to enable us to achieve highest possible recall ratio and recall precision based on all the controlled index terms in the databases. The controlled index terms used were "big data" (TI="big data" OR TI="big-data" OR TI= "bigdata"), and also the Boolean operator "OR". The timespan was limited from 1956 to 2019 October. One of the reasons why "massive data" was not included in the search strings was based on this paper definition of big data, and as alluded to from literature about the five "Vs" that make up big data, and such relative term "massive data" does not equal to "big data". This technique was adopted by (Ajibade and

Mutula, 2019). Furthermore, a related study on green innovation used a single search string (Albort-Morant, Henseler, Leal-Millán, and Cepeda-Carrión, 2017), and "project management (Cobo Serrano, 2018). There were 8, 415 outputs based on the search criteria, with h-index=104, ACP=97 (average citation per item), STC=81,645 (sum of times cited). For the analysis, the extracted data was imported from the search criteria into the desktop, and data cleaning was performed and the output analysed. Some 322 big data outputs were published in 88 LIS journals across 396 institutions from the WoS databases. A sample of 396 organisations was used as the unit of analysis to analyse output, citations and collaboration pattern. Subsequently, for the type of analysis, bibliographic coupling was used based on document, organisations and sources of the outputs.

Findings and Discussions

Outputs by Year

The LIS research outputs by year which are indexed in the SSCI, A and HCI, and ESCI in the Web of Science core collection databases (see methodology) span ten years from 2011-2019. Year 2019 accounted for 41 outputs (12.733%), 2018, 68 outputs (21.429%), 2017, 73 outputs (22.671%), 2016, 59 outputs (18.323%), 2015, 39 outputs (12.112%), 2014, 23 outputs (7.143), 2013 and 2012, 8 outputs each which accounted for (2.484), and 2011 LIS research outputs on big data were 0.621%. The highest output by the LIS research on big data was in 2017, which accounted for 22.671%, which is very low relative to other outputs from other fields. The distribution of outputs based on the language of publications were English, 292 (90.683%), Spanish, 14 (4.348%), Portuguese, 5 (1.553%), German, 4 (1.242%), Catalan and Hungarian, 3 each (0.932%), and French, 1 (0.311%).

Bibliometric Mapping and Network Clustering

The two predominant models used to present the result visualisation were clusters and network mapping. Thus, the visualisation of the bibliometric mapping and network analysis were presented based on the following models.

Mapping of Network:

$$v\left(x_{i,\,\ldots\ldots,}\,\mathrm{x}_n\right) = \sum_{i<j} s_{ij}\ ||\ x_i - x_j\ ||$$

base on the following constraint

$$\frac{2}{\mathrm{n}\left(n-1\right)}\sum_{i<j} ||x_i - x_j\ ||= 1,$$

Using this metric, the *n* represents the network node, and the x_i represents the location of node i, and $||x_i$-$x_j||$ denotes the geometry of distances between the nodes *i* and *j*.

Clustering:

For the clustering, techniques adopted for this paper, data are denoted as follows:

c_i represents assigned nodes, δ(c_i, c_j), function = 1 if *ci = Cj and* 0 other, and *γ* denotes the resolution parameter that determines the details of the clustering (Van Eck and Waltman, 2014), meaning the value of *γ* determines the level of clustering details; hence, the model is expressed as:

$$V\left(c_i,\ldots,\mathrm{c_n}\right) = \sum_{i<j}\delta\left(c_i,c_j\right)\left(s_{ij}-\gamma\right)$$

As δ (x_i, x_j) equals 1 if x_i = x_j and 0

$$V(x_1,\ldots,x_n) = \frac{1}{\gamma}\sum_{i<j}(1-\delta(x_i,x_j))\left(\frac{1}{\gamma}\frac{2mc_{ij}}{c_ic_j}-1\right)$$

and because δ (xi, xj)equals 1 if xi = xj and 0

$$\hat{V}(x_1,\ldots,x_n) = -\frac{\gamma^2}{2m}V(x_1,\ldots,x_n)+\frac{1}{2m}\sum_{i<j}\left(\frac{2mc_{ij}}{c_ic_j}-\gamma\right)$$

Outputs Institutions using Clustering Network Analysis

Previous studies have established that articles within the same areas of study or focus of interests are often or are likely to be cited together (Hjorland, 2013). Liao, et al. had carried out a visualisation analysis of big data and medical research (Liao et al., 2018). However, since there are 189 clusters in this study, the authors only examined the top clusters which are outputs in red, dark blue, purple, sky blue, green and light brown below (see figure 1). Data in Table 1 suggests that the top two countries accounted for forty-one and a half percent (41.5%) of the output, yet the same two countries (the USA and China) had 51.5% of the total citation per countries with the USA in the first place (43.8%), and China outputs accounting for almost eight percent (7.7%). Although Canadian outputs were ranked 7th in the total contributions, they accounted for 12.4% of the total citation. Therefore, we concluded that citation analysis presented significant statistical inferences to measure outputs visibility, relevance and impact by the citation analysis.

Table 1: Co-authorship Outputs by Countries and Institutions

Co-authorship by Country				Co-authorship by Institution			
countries	TCC	TCPP (%)	TLS	Institutions	TCI	TCPP (%)	TLS
USA	112	2671 (43.8)	53	Wuhan University	8	37	5
China	45	472 (7.7)	32	City University Hong Kong	6	50 (0.9)	4
India	25	143 (2.3)	15	Massey University	6	41	2
Spain	20	66 (1.1)	1	Kent State University	5	73 (1.3)	6
England	18	240 (3.9)	11	Sun Yat Sen University	5	76 (1.3)	2
Germany	14	153 (2.5)	12	University of Illinois	5	15	4
Canada	13	759 (12.4)	7	MIT	4	26	2
South Korea	12	239 (3.9)	10	Nanjing University	4	23	0
Taiwan	12	43 (0.7)	6	San Diego State University	4	186 (3.3)	2
Brazil	10	36 (0.6)	2	University Cincinnati	4	1293 (23.0)	7
Netherlands	10	122 (2.0)	7	University Malaya	4	231 (4.1)	0
New Zealand	10	59 (1.0)	5	Copenhagen Business Sch.	3	76 (1.3)	0
Australia	9	137 (2.2)	8	Delft University Technology	3	22	0
Denmark	7	135 (2.2)	4	Drexel University	3	5	0
France	7	61 (1.0)	7	Erasmus University	3	15	1
Italy	7	168 (2.8)	3	Georgia State University	3	1266 (22.5)	2
Finland	6	44 (0.7)	5	king Abdulaziz University	3	20	1
Pakistan	6	51 (0.8)	14	NYU	3	136 (2.4)	0
Ireland	5	57 (0.9)	3	Simon Fraser University	3	43	1
Malaysia	4	231 (3.8)	5	Southern Lazio	3	110 (2.0)	6
Saudi Arabia	4	26 (0.4)	8	Stanford University	3	28	1
South Africa	4	4 (0.1)	0	University Arizona	3	1308 (23.2)	2
Algeria	3	3 (.0)	1	University Carlos iii Madrid	3	25	0
Liechtenstein	3	39 (0.6)	3	University Cassino	3	110 (2.0)	6
Portugal	3	21 (0.3)	0	University Liechtenstein	3	39	0

Sweden	3	60 (1.0)	4	University North Carolina	3	102 (1.8)	0
Switzerland	3	4 (0.1)	6	University Oberta Catalunya	3	9	0
UAE	3	55 (0.9)	2	University Roma tor Vergata	3	110 (2.0)	6
				University Tennessee	3	65 (1.2)	3
				University Virginia	3	29	1
				University Waikato	3	35	2
				University Wisconsin	3	29	0

TCPP = Total Citations per Paper;
TLS = Total Link Strength;
UAE = United Arab Emirates
TCC = Total Output Count per Country
TCI = Total Co-Authorship Output Count per Institution

Global Trend of LIS Co-authorship Distributions by Countries and Institutions

Table 1 shows the percentage distribution of the top 15 co-authorship contributions from countries and top 5 Institutions. The citation analysis is a vital measure to evaluate the outputs and performances as a quantitative metric to rank visibility, influence and impact of institutions. Mishra et al. (2018) note that to rank journal significance, citation analysis is effective. The University of Arizona output (n=3) was ranked in the sixth place based on the total outputs. However, when the citation metrics were used as a unit of impact analysis, the same accounted for more than twenty-three percent (23.2%), thus ranking the university contributions in the first place. The University of Cincinnati outputs (n=4) was in the fourth place but ranked second (23%) using the citation index. In the third place was the Georgia State University outputs (n=3) which accounted for twenty-two and a half percent (22.5%) of the LIS outputs on big data. Institutions from the 80th to 189th clusters had one output each in the WoS outputs. In South Africa, there were only three institutions with one output each, and the Cape Peninsula University of Technology was in the 8th cluster possibly because of its co-authorship with an institution in that cluster. The University of KwaZulu-Natal was in the 170th clusters, and the University of Stellenbosch was in the 184th cluster as a single entity.

Outputs by Journal Sources for Big Data in the LIS Field

The clustering of bibliometric data aggregates and groups articles with the same aims and area of focus together (Mishra et al., 2018). The Big Data publications within the LIS discipline were published from 88 journals which comprised 21 clusters and 1255 network links. Out of these, 76 journals items were not connected or linked within the cluster. This suggests that authors had not cited the outputs in these 76 journals from the other journal. However, based on the argument of Horjland (2013), this could be because the area of focus was not closely related, or as a result of low outputs on big data research in LIS. Nevertheless, the library profession cannot be exempted from researching these technologies as its impact in the LIS profession is inevitable. The different network colour indicates journals publishing outputs in a similar or the same areas of interest (Hjorland, 2013; Anne, 2019). Unfortunately, none of the known South African journals were visibly represented in these clusters.

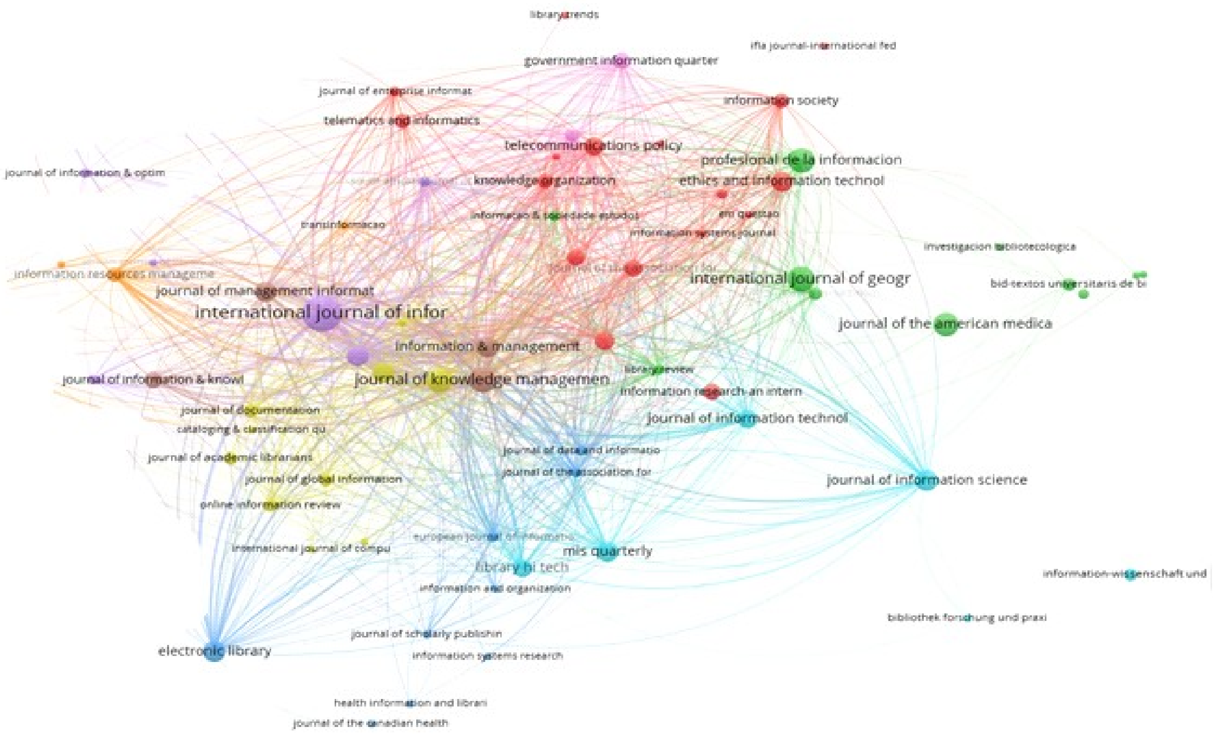
library trends
ifla journal-international fed
government information quarter
journal of enterprise informat
telematics and informatics
information society
telecommunications policy
profesional de la informacion
journal of information & optim
knowledge organization
ethics and information technol
informacao & sociedade-estudos
em questao
transinformacao
information systems journal
investigacion bibliotecologica
information resources manageme
journal of the association for
international journal of geogr
bid-textos universitaris de bi
journal of management informat
international journal of infor
journal of the american medica
information & management
library review
journal of information & knowl
journal of knowledge managemen
information research-an intern
journal of documentation
cataloging & classification qu
journal of information technol
journal of academic librarians
journal of data and informatio
journal of the association for
journal of global information
journal of information science
online information review
european journal of informatio
international journal of compu
mis quarterly
library hi tech
information-wissenschaft und
information and organization
bibliothek forschung und praxi
journal of scholarly publishin
electronic library
information systems research
health information and librari
journal of the canadian health

Fig. 1: Outputs based on the Bibliographic Coupling Network (the network colour indicated journal which articles have been cited within the network

Bibliographic coupling analysis is an important technique for identifying collaboration, rankings, clusters of contributions by the most influential and active authors and journals (Ferreira, 2018). Co-citation and bibliographic coupling are essential to tracking research focus and changes in a field of study (Chang, Huang and Lin 2015). Despite the usefulness of co-citation and bibliographic coupling, the visualisation depicted an indirect relationship; thus making it less accurate in comparison with a citation analysis (Van Eck and Waltman, 2017). Nevertheless, co-citation and bibliographic coupling can be useful for understanding the intellectual structure of research trajectories.

Citation Analysis

The citation analysis showed that from the 8, 415 global outputs on big data, 48,917 citing articles had cited these outputs 81, 645 times. However, the total output by the library and information science outputs were 322 articles. The citing articles could be used as a measure to find out the prominence of a study's areas on a particular subject and how other researchers perceived the authority of such a particular topic. This analysis indicated that in the citation report, other researchers citing articles had cited one or more of these articles (items) in the citation report analysis. The system extrapolated that out of 38,116 total citing articles, Library and Information Science accounted for 3.030% which is 1155 total record (citing article). This is in comparison to Computer Science which was in the first place with 11,262 records representing 29.547% outputs of the total 38,116 total citing articles. Engineering was second place with 8,161 other articles citing their outputs, which accounted for 21.411%, and telecommunications was in the third place with 3,754 citing articles that accounted for 9.849% of the total citing articles. The Science and Technology research area was in the fourth place with 2,548 citing article, accounting for 6.685% of the total 38,116 citing articles.

Data regarding South Africa scholars' contribution suggested an uptake in research interest in big data as some institutions were collaborating with other researchers worldwide based on 147 outputs from Scopus database (see table 2, co-authorship). The justification for including this dataset was that, while South African scholars had 40 outputs from WoS, their outputs from the Scopus were more. However, there was a limitation to the Scopus dataset, because it did not show LIS contributions specifically, but Social Sciences and Humanities. The output was cited 1,266 times from the Web of Science Core Collection citation reports, the library and information

Co-authorship by institutions: (selected institutions with at least three outputs on big data)

Scopus Big Data Collaboration with South African Scholars

One of the innovative attempts adopted by this study was to use field-weighted citation impact (*fci*) to test the co-authorship distribution of countries' outputs as a unit of analysis using a small sample to see if it would serve as an insightful indicator to measure collaboration. Although Reller (2016) used it (*fci)* to measure the field-weighted citation impact as an indicator for field-based differences in citation, we inferred that the same could be used to measure collaboration impact, as we replaced the field of study with countries, but still used the citations distributions as the unit of analysis (see table 4). Our findings concluded that *fci* is a valid indicator to measure the performance and influence of an output based on co-authorship and collaboration. The data indicated that although South Africa's outputs were many, the FCI of Netherlands were three (*fci* = 3) indicating that the Netherlands outputs on big data in the LIS had been cited 200% more than the world average.

Table 2: Top Countries collaborating with South African institutions

Rank	Countries and territories	Outputs (TPCC)	Citations per Publication	Field-Weighted Citation Impact	Citation Count (TCCP)
1.	South Africa	111	4.0	0.88	439
2.	United Kingdom	10	16.3	2.00	163
3.	Australia	7	9.3	1.83	65
4.	Netherlands	5	15.4	3.08	77
5.	United States	5	9.2	1.20	46
6.	Belgium	2	0.5	0.00	1
7.	Germany	2	2.5	0.41	5
8.	Sweden	2	1.0	0.64	2

Table 2 shows countries collaborating with South Africa based on the top 100 countries output in Scopus analysis. Other collaborating countries had one output each with South African scholars these were not included. Furthermore, beyond the citation analysis, we examined the field-weighted impact of the citation (see table 2) to determine the ratio of the total citations received by each country's contributions vis-à-vis their collaborative outputs with South African institutions and scholars. Other studies (Salimi, 2017) used the field-weighted citation impact (FCI) as a metric to present the impact of citation ratio as denominator's outputs vis-à-vis the total expected citation based on the average of the selected unit of analysis (subject field, citation, outputs etc.). Based on these metrics (FCI = 1), where the FCI is equal to one, we suggested that the citation analysis be performed as expected. Consequently, a figure below one indicated that the outputs performed relatively lower below average for the global average as would the case for South Africa citation analysis. However, the output by Germany, Belgium and Sweden can be ignored due to low co-authorship outputs with South Africa. Nevertheless, we inferred that the FCI for South Africa was lower because big data research uptake in South Africa is relatively low especially in the LIS field, in comparison with South Africa and the UK, Netherlands and the USA.

In contrast to the Scopus data, there were only 40 outputs (n = 40) from the WoS, with h-index = 8; STC = 993; CA= 956; ACI = 24.83 in total. However, the citation distribution from 2008 to 2016 were one hundred and fourteen (STC =114), while 2017's citations surpassed that with (STC = 204), 2018 (STC-227), and 2019 (STC =194), and the yearly total average citation per year from 2008-2019 was above eighty (ACY = 82.75). Although the analysis showed an increasing interest in big data, the only limitation to this metric was the inability to generalise as statistically significant due to the limited sample.

Conclusion

The contributions of the library and information studies field and the contribution of the scholars in the field to big data are presented in this study. Our findings showed the extent and contributions of journals in the LIS field and the focus of publications on big data. Most importantly, the findings highlighted the diminished research trends on certain aspects of big data technologies such as cloud computing, internet of things, amongst other key variables. The big data outputs from the African scholars were limited in scope and volume. Most of the LIS big data outputs have not addressed techniques and technologies to facilitate big data collections, mapping, filtering and reduction. Nevertheless, the ability of the LIS scholar to map, filter and reduce big data content will assist them in creating faceted classifications, automatic controlled indexing terms and big data metadata management. The study recommends LIS focus on big data metadata management for solving some of the challenges of controlled indexing terms. Furthermore, ontology and classification and how it might be applicable in the big data should be examined. This may require the ability of scholars in the LIS to use some of the big data analytics software to perform data scrubbing, mapping, and processing. The study reveals that the contribution of South Africa to studies on big data was very low.

References

Ajibade, P., and Mutula, S. M. (2019). Bibliometric Analysis of Citation Trends and Publications on E-Government in Southern African Countries: A Human-Computer Interactions and IT Alignment Debate. *Library Philosophy and Practice* (E-Journal), 2234(Winter 2019), 1- 19. Retrieved From Https://Digitalcommons. Unl. Edu/Libphilprac/2234/

Albort-Morant, G., Henseler, J., Leal-Millán, A., and Cepeda-Carrión, G. (2017). Mapping the Field: A Bibliometric Analysis of Green Innovation. Sustainability, 9 (6), 1011.

Aydin, B., Akkineni, V., and Angryk, R. A. (2016). Modeling and Indexing Spatiotemporal Trajectory Data in Non-Relational Databases. In Managing Big Data in Cloud Computing Environments (pp. 133-162). IGI Global.

Blummer, B., and Kenton, J. M. (2018). Big Data and Libraries: Identifying Themes in the Literature. *Internet Reference Services Quarterly*, 23 (1-2), 15-40.

Chang, Y. W., Huang, M. H., and Lin, C. W. (2015). Evolution of Research Subjects in Library and Information Science Based on Keyword, Bibliographical Coupling, and Co-Citation Analyses. *Scientometrics*, 105 (3), 2071-2087.

Christensen, H. D. (2017). Rethinking Image Indexing? *JASIST*, 68 (7), 1782-1785.

De Mauro, A., Greco, M., and Grimaldi, M. (2016). A Formal Definition of Big Data Based on its Essential Features. *Library Review*, 65 (3), 122-135.

Di Pretoro, E., and Geeraert, F. (2019). Behind the Scenes of Web Archiving: Metadata of Harvested Websites. ABB: Archives Et Bibliothèques De Belgique - Archiefen Bibliotheekwezen In België, Brussel: Archief- En Bibliotheekwezen in België, N.D., In Press, Trust And Under-standing: The Value Of Metadata In A Digitally Joined-Up World, Ed. By R. Depoortere, T. Gheldof, D. Styven and J. Van Der Eycken, 106, pp.63-74. Ffhal-02124714

Chen, H. L., Doty, P., Mollman, C., Niu, X., Yu, J. C., and Zhang, T. (2015). Library Assessment and Data Analytics in the Big Data Era: Practice and Policies. Proceedings of the Association for Information Science and Technology, 52 (1), 1-4.

Cobo Serrano, S. (2018). Producción Científica Internacional Sobre Gestión De Proyectos En El Área De Información Y Documentación: 1996-2015. *Investigación Bibliotecológica*, 32 (75), 125-144.

Cuifeng, H. (2013). Library Services Innovation and Development in the Era of Big Data [J]. The Library, 1.

Demchenko, Y., De Laat, C., And Membrey, P. (2014). Defining Architecture Components of the Big Data Ecosystem.

Ferreira, F. A. (2018). Mapping The Field of Arts-Based Management: Bibliographic Coupling and Co-Citation Analyses. *Journal of Business Research*, 85: 348-357.

Golub, K., And Hansson, J. (2017). (Big) Data in Library and Information Science: A Brief Overview of Some Important Problem Areas. *Journal of Universal Computer Science* (Online), 23 (11), 1098-1108.

Gulgec, N. S., Shahidi, G. S., Matarazzo, T. J., And Pakzad, S. N. (2017). Current Challenges with Bigdata Analytics in Structural Health Monitoring. *In: Structural Health Monitoring and Damage Detection*, Volume 7 (pp. 79-84). Springer, Cham.

Hjørland, B. (2013). Citation Analysis: A Social And Dynamic Approach to Knowledge Organization. *Information Processing and Management*, 49(6), 1313-1325.

Hu, J., And Zhang, Y. (2017). Discovering the Interdisciplinary Nature of Big Data Research Through Social Network Analysis and Visualization. *Scientometrics,* 112 (1), 91-109.

Jeble, S., Kumari, S., and Patil, Y. (2018). Role of Big Data in Decision Making. *Operations and Supply Chain Management-An International Journal*, 11 (1), 36-44.

Mishra, D., Gunasekaran, A., Childe, S. J., Papadopoulos, T., Dubey, R., and Wamba, S. (2016). Vision, Applications and Future Challenges of Internet of Things: A Bibliometric Study of the Recent Literature. *Industrial Management and Data Systems*, 116 (7), 1331-1355.

Mishra, D., Gunasekaran, A., Papadopoulos, T., and Childe, S. J. (2018). Big Data and Supply Chain Management: A Review and Bibliometric Analysis. *Annals of Operations Research*, 270 (1-2), 313-336.

Munshi, U. M. (2016). Management of Phytophthora–A Deadly Plant Pathogen. *Current Science*, 110 (12), 2213.

Patel, A. B., Birla, M., and Nair, U. (2012, December). Addressing Big Data Problem Using Hadoop and Map Reduce. In: 2012 Nirma University International Conference on Engineering (Nuicone) (Pp. 1-5). IEEE.

Liao, H., Tang, M., Luo, L., Li, C., Chiclana, F., and Zeng, X.-J. (2018). A Bibliometric Analysis and Visualization of Medical Big Data Research. Sustainability, 10 (1), 166.

Shan, J., and Gang, W. (2013). The Enlightenment of Big Data For Library [J]. Library Work and Study, 4.

Simoviæ, A. (2018). A Big Data Smart Library Recommender System for an Educational Institution. *Library Hi Tech.*

Sun, G., Chang, V., Guan, S., Ramachandran, M., Li, J., and Liao, D. (2018). Big Data and Internet of Things – Fusion for Different Services and its Impacts. *Future Generation Computer Systems*, 86: 1368-1370.

Reller, T. (2016). Elsevier Publishing – A Look at the Numbers, and More. *Key Journal*. Pp. 1-7.

Salimi, N. (2017). Quality Assessment of Scientific

Outputs Using The BWM. *Scientometrics*, 112 (1), 195-213.

Teets, M., and Goldner, M. (2013). Libraries' Role in Curating and Exposing Big Data. *Future Internet*, 5 (3), 429-438.

Van Dijck, J. (2014). Datafication, Dataism and Dataveillance: Big Data Between Scientific Paradigm and Ideology. *Surveillance and Society*, 12 (2), 197-208.

Van Eck, N. J., and Waltman, L. (2017). Citation-Based Clustering of Publications Using Citnetexplorer and Vosviewer. *Scientometrics*, 111 (2), 1053-1070.

Wang, C., Xu, S., Chen, L., And Chen, X. (2016). Exposing Library Data with Big Data Technology: A Review. Paper Presented at the 2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS).

Wang, X., Wang, W., Yang, L. T., Liao, S., Yin, D., and Deen, M. J. (2018). A Distributed HOSVD Method with its Incremental Computation for Big Data In Cyber-Physical-Social Systems. *IEEE Transactions on Computational Social Systems,* 5 (2), 481-492.

Weihong, F., Chenhui, L., Xingwang, Z., Xiaozhu, Q., and Zikuan, G. (2012). How do Libraries Need" Big Data"? [J]. *Library Journal*, 11. [Online]Available: Http://En.Cnki.Com.Cn/Article_En/Cjfdtotal-TNGZ201211017.Htm

Yýldýrým, A. A., Özdoðan, C., and Watson, D. (2016). Parallel Data Reduction Techniques for Big Datasets. In: Big Data: Concepts, Methodologies, Tools, and Applications (pp. 734-756). IGI Global.

Dr. Patrick Ajibade is currently a National Institute for the Humanities and Social Sciences NIHSS CODESRIA African Pathways Scholar in the Information Studies Programme at the University of KwaZulu-Natal, South Africa. He holds MSc (MLIS-Cum laude) and PhD from the University of Fort Hare. He obtained M.Sc. Information Studies, (BIS) Business Information Systems from Universiteit van Amsterdam, and Vrije Universiteit.

Prof. Stephen Mutula is Professor in the Information Studies Programme in the School of Social Sciences, at the University of KwaZulu-Natal, and Immediate past Deputy Vice Chancellor of College of Humanities of the same Institution. Currently, Professor Mutula is the Dean and Principle, School of Management, Information Technology and Governance in the same University.