Summer 8-30-2020

Big data and openness: a big issue with librarians

Oluwaseyi Wusu
drwusu1@gmail.com

Big data and openness: a big issue with librarians

James Oluwaseyi Hodonu-Wusu[1,2,5]**, Nneka G. Lazarus**[2,3]**, Mumeen Omoniyi Otun**[4,5]**,**
Olaniyi Basheer Arekemase[5]**,** and **Tokunbo Taofeek Olayiwola**[6]

[1]Department of Library and Information Science,
Faculty of Computer Science and Information Technology,
University of Malaya, Kuala Lumpur, MALAYSIA
[2]Department of Library, Archival and Information Studies,
Faculty of Education, University of Ibadan, Ibadan, NIGERIA
[3]Adeniran Ogunsanya College of Education, Lagos, NIGERIA
[4] Department of Library and Information Science,
Faculty of Education, University of Calabar,
Cross River, PMB 1115, Calabar Municipal, NIGERIA
[5]African Regional Centre for Information Science, University of Ibadan, Ibadan, NIGERIA
[6]Information Technology, Information Technology & Engineering, Management and Science University, 40100
Selangor, MALAYSIA
Corresponding Author e-mail: wususong@siswa.um.edu.my[1,2,5],

Abstract

Big data refers to the explosion of available information and knowing how to handle it is a big challenge to librarians in this age. The massive sample size and high dimensionality of big data introduce unique challenges to librarians, including scalability, sharing, intellectual properties and storage bottleneck. This paper overviews the opportunities and challenges brought by big data to librarians and presents ways of making data available for everyone to use without limitations, with emphasis on the distinguished roles librarians need to play in the era of Big Data and Openness to scholarly communications and, also presents the precision and recall scores of the articles. Future works can extend the search domain to predict the accuracy of the information retrieving processes.

Keywords: *Big Data, massive data, high dimensional data, librarians, Open Science, Scholarly Communication.*

1 Introduction

In today's scholarship, data is ubiquitous and librarians as well as researchers are finding it difficult to cope with the varied amount and availability of data everywhere. Big data refers to the explosion of available information, such a Big Data movement is driven by the fact that massive amounts of very high-dimensional or unstructured data are continuously produced and stored with much cheaper cost than they used to be [1]. From the time immemorial, man has been dealing with invention of reading, writing, storing, retrieving as well as disseminating information. However, as innovation continues to thrive, man's ability to capture and analyze data has been challenged. With the availability of computer and novel forms of information and communication technologies as well as the internet connecting many computers together worldwide with exceptional results that ordinary human brains cannot comprehends. [2] argues that Big data applies to data sets of extreme size (e.g. exabytes, zettabtes) that ordinary software of personal computer cannot cope with too. In his work, [3] believes that we are in the petabyte era "– the era of Big data, where more is not just more, more is different". It is a situation whereby large data sets are big in volume, velocity, veracity and variability [2]. This data is too big, too fast, or does not fit the regular database architecture and so, for librarians, a special ability or requirement is needed for creating, processing, using, storing, retrieving, assessing and disseminating information to the users [1][4]. Also, "data quality is a concern of everyone in the Big Data space. The quantity of data is growing at a geometric pace (increased volume, velocity and variety - the very description of Big Data) and library and information centres are not left behind. According to IBM, 90% of all data was created in the past 2 years. Therefore, making sense of all of this data starts with the need for assurance that the data is accurate - a very real concern [5], [6] to librarians."

Data librarians need to consider how to relate researchers and their data to established support and management frameworks such as copyright and intellectual property rights (IPR). "These are rights acquired over any work created, or invented, because of the intellectual effort of an individual. The copyrights of IPR are traceable to the institutional repositories of the researchers in the library, other forms of IPR include patents, trademarks, geographical indications, industrial design rights, design layouts and confidential information or trade secrets [7]." Researchers have a clear understanding on how IPR affect their publications or research outputs which can define their position as far as commercial exploitation and academic attribution is concerned. IPR should be covered by law and message to communicate in the mechanisms for defining how data may be managed, disseminated and re-used [7]. Databases may be protected by the 'database right' arising from the European Database Directive 1996, in recognition of work involved in their creation, structuring and arrangement. The anxiety of who owns the rights to a data and the feelings of lack of clear guidance or policy from the institution is a challenge to the researchers [40].

However, in the era of big data, the relationship between libraries and researchers has taken a new dimension and its evolving drastically. Technology is decreasing user's appreciation to library's services. Taking steps to strengthen and reinforce the role of librarians and information professionals, can revitalize their relationships with users and help create a library that users need and want. What librarians need is an innovative relationship between libraries and the people who use them, and the key to that relationship is the librarians and information professional — the people who are engaged in the pursuit and sharing of knowledge; the people who develop and deliver library and information services. This relationship holds the key to creating a library that users want – a library that anticipates needs rather than just responding to them, A library that moves with time, a library on the go – a real time library, a library that is communal as well as social, a library that has envoys and enthusiasts and cultivates loyalty, and a library that meets the users in their comfort homes. Librarians need to make this happened and knowledge of doing this essential for library services [8]. According to [9], see this trend to have a deep impact on science. For example, scholarly advances are becoming more and more data-driven and librarians and researchers will more and more think of themselves as consumers of data. Hence, the massive amounts of high dimensional data bring both opportunities and new challenges to data librarians.

Furthermore, with the proliferation of data on the internet and whole wide world (www), librarians everywhere are starting to grapple with massive data sets, encountering challenges with handling real time information, processing and moving information that were once in the university's domain or library's community server to known and unknown servers. The issue of know-how as well as data storage, analysis and sharing of data is another challenge to librarians in this period. More so, storing and interpreting big data takes both real and virtual bricks and mortar. For instance, on the European Institutes (EI) campus, construction is under way to house the technical command center of ELIXIR – a project to help librarians and other scientists across Europe and other places safeguard and share their data, and to support existing resources such as databases and library and computing facilities in individual countries [10]. Also, CERN has one supercollider producing data in one location, librarians and libraries generate high volumes of research data, institutional repositories, theses, journals and other publications are distributed across many libraries' portals — highlighting the need to share resources. In terms of copyrights and intellectual property, librarians can create virtual spaces for publications, data, software and results that anyone can access, or they can lock the spaces up behind a firewall so that only a select group of collaborators can get to them [10].

In today scholarship, many of the scientific studies created or designed are shared online and so more emphasis are on cloud computing – which housed data and software in a huge, off – site centers that any user can access on demand. Librarians today, don't need to buy their own hardware in order to use and maintain it on site in the era of big data. They can partner with IT firms that work on such a service after necessary policies and implementation phases have been met [11]. These can be done via online, a lot of academic activities and other related information can be accessed and done virtually. In terms of copyrights and intellectual property, librarians can create virtual spaces for research findings and publications, data, software and results that anyone can access, or they can lock the spaces up behind a firewall so that only registered researchers or selected group of collaborators can get to them. The aim of this paper is to present an overview, the opportunities and challenges brought by big data to librarians as well as presents ways of making research findings, data, books and other electronic resources available for everyone to use without limitations, with emphasis on the distinguished roles librarians need to play in the era of big data and openness to scholarly communications.

2. *Literature Review*

2.1 The Era of Big Data and Open Scholarly Communication

The term big data was coined in the 2000s and migrated to all disciplines about 10 years later [12]. Big data can be defined as the data generated using digital technologies [13]. It has also been described by three Vs namely, volume, velocity and variety [14]. It is a process through which great volume of data that is too large for standard computer memory and software [15], [16]. Due to the increase in digital interference which led the big data era, scholarly communication and data-driven researchers have become popular. According to [17], it is easier to find data these days, but researchers need to select, analyze and compare and finally publish for other to use and redistributed. It was found out that the top three major problems of data are data preservation (data curation), which includes accountability for publicly funded research, inspiration for scientific advancements and reanalysis of previously generated data [18]. The Table 1 below highlights areas where data has a link with scholarly communication of librarians.

Table 1: Area where data has a link with scholarly communication of librarians

Main areas	Secondary Categories	References
Data as a source	Accuracy/Objectivity, Big Data, Big Data/Social network, data analysis, data reporting, database, digital data reporting, data ethics, data mapping tools, open data, data privacy, social networks	[19]
Data dissemination	Accuracy/Objectivity, big data, book review, case study, data security, data-driven researchers, data reporting, data practices, data awareness, Open data	[20]
Informational, consulting –type services	Institutional repository, data management, technical plans, hands-on services, data referencing, data training and practices	[19]
Research data management services	Researchers perception on data science, data development and management, data training and practices, grant proposal support, data management planning, locating data-related services, publication support, data management assistance.	[19]
Data repositories/data practices	Library responsibility, librarian roles, data training and data curation, advising, preservation of research output,	[21]; [22]; [23]; [24]; [25]
Data Awareness	Adopt data archiving, preservation, training data awareness	[19]; [25]; [26]
Research data preservation	Accountability, scientific advancement, analysis and reanalysis of generated data	[18]

2.2 Application of Open Scholarly Communication to Libraries

The role of libraries on open science has been recognized and discussed at multiple fora. There are increase number of scholarly societies and institutions developing open scholarly communication journals as part of contribution towards open science. Academic librarians manage the development of OA institutional repositories that houses theses, dissertations, institutional documents and data, as well as other files that may likely be accessible by the public [27]. Interestingly, [28] not in relation to open scholarly communication movement as follows:

> "Repositories as a system for collecting, publishing, disseminating and archiving digital scientific content have become one of the most prominent types of digital library applications. Especially with respect to Open Access publishing, repositories today serve as a platform for acquiring and disseminating scientific content, which before had been almost exclusively released by commercial publishers (para 1)."

Making resources available through open scholarly communication portend their readiness and practices towards openness in science, this equally helps the researchers as well as other academicians get scholarly information. The availability of preprint and post-print, dissertation, theses, dissertation research reports and other scholarly resources shows their practices towards expanding open scholarly communication as well as sharing scientific and knowledge sharing. To collaborate their readiness and practices of Open Science, [29] stated that academic libraries have embraced digital publishing to provide digital resources for both faculty and students or other users.

Academic librarians and their libraries are seen as resources as well as publishers all at once, considering their practices in publishing and disseminating knowledge. They are seen as very useful resources for research supports and scholarly communication [30]. Academic librarians are seen as promoters of Open Science initiatives, they serve as librarians, researchers, reviewers, editors and provide access to research output and other documents from their individual institutional repositories. "As an enabler, librarians and libraries have adapted their roles as a preserver, curator, disseminator of digital scientific findings if the form of publications, data and other research – related content. Libraries and repositories constitute the physical infrastructure that allows researchers to share use and reuse the outcome of their work, and they have been effective in the synthesizing open science movement" [31]. In fact, librarians contribute immensely to the scholarly and scientific communication by providing and marketing the resources instead of keeping them away from the users, this they do by opening doors of scholarly communication.

Besides urging the journal articles to be accessible online and free, the OA movement and Open Science initiatives have brought repositories to academic institutions. The universities authorities, as well the funding agencies have mandated all researchers under their watch to make known their research outputs or results to the public which is the major aspect of open scholarly communication and librarians have led the ways to the movement of Open Science [32]. However, this movement does not exist universally, in developing countries for example, the movement has been slow. Many libraries still consider their repositories as an important asset of the universities and some librarians keep them away from the public and still consider themselves as the custodian of repositories. More so, lack of infrastructure for online access is another challenge faced by them and this has slow pace open scholarly communication the third world nations. The open scholarly movement is a movement that makes repositories available online and freely accessible via the internet [33] [31]. Academic librarians need to be aware of this fact, and key into the vision and mission of open science initiatives which among others is to provide resources in order to boost the growth of scientific knowledge. All librarians at all levels need to key into this vision and be well prepared for the challenge ahead. Also, the perception of librarians needs to change concerning how science is being carry out today and they need to move with time else they would be left behind by technology. Libraries should not be a close access to institutional repositories, rather should be open access repositories. Every hand must be on deck to drive in this vision across the breadth and length of our institutions as librarians and researchers need a lot to do to achieve this. Also, data awareness of the librarian can go a long way in extending traditional information literacy and bibliographic instruction programmes. In some ways existing forms of library instruction lend themselves easily to the addition of concepts of data management and re-use. For example, in teaching about doing a literature search in each discipline, librarians may give instruction in using standalone or online reference management tools, such as EndNote, Reference Manager, Zotero or Mendeley [34]. However, some disciplines used data in textual form (such as law and history), this method could be the best ways to conduct data management through a research project and other discipline that uses other data type, data librarian could further explore other options such as those described in the Research Data MANTRA training course Organizing Data available at[1]. Scholarly communication could also be given by the data librarian through bibliographic search techniques that may be useful in teaching data discovery skills by the librarian which would eventually aid discovery of published literature and data [34]. Furthermore, many online databases, licensed products or datasets required permission to access, content cannot be indexed by Google, and so understanding of the potential sources is paramount. Training in data sources may be offered by the data librarian as part of information skills programmes, alternatively, online -based resources may equally be offered for example the Bodleian Data Library Web Page[2]. Researchers can also learn skills for citation trails to key databases through publication lists of seminar articles, for instance, where a publication has been written that is based on an original dataset, the author should provide instructions for accessing the dataset for re-use, if not a complete citation to it, else, the author's details may be used to track down an incomplete reference to a dataset or better still the data librarian can help contacting the author on behalf of the researchers. More so, data citation could also be done through the awareness given to the researchers, staff and students of the institutions in order to give credits to whom it due [35] [36].

As shared datasets, images, video clips and other non-textual digital objects become more valued in exchanges of scholarly communication, the provenance of these objects gradually becomes as important to the scholarly record as the peer - reviewed, published papers which describe analyze them. This is not only important for the reader who wishes to track down a copy of the original object; it is equally essential for the object's creator who wishes to receives to receive career rewards based on the academic value of their work, as measured through citation counts and other impact measures that show the data have been recognized, consulted, downloaded or cited in other studies (including replication studies). therefore, data or academic librarian wishing to support best practices of scholarly communication can equally advocate proper data citation along with a well-established bibliographic citation practices, stressing not only that it should be done but also advising on how it can be done [34].

2.3 Ways by which Libraries can fulfilled their role as enablers of Openness in Science

1. **Changes in the environment:** Change is ubiquitous either politically, socially, educationally, economically or technologically. Many things required change and librarians cannot be left behind in their areas of job scheduled, most particularly in the roles and tasks of those involved in the preservation and transmission of cultural heritage, and interpersonal information intervention. Sharing, storing and retrieving digital information are commonplace activities in today's world, and now formally built digital libraries constitute an important component of this virtual information and dissemination environment [37]. Like conventional physical libraries, virtual libraries are created to serve clienteles or to collect and provide access to selected information resources (whether text documents or artefacts). As information and communication technologies open up entirely new opportunities which are not achievable using physical information resources. It should be remembered, in the face of such change, that there is no immutable set of fixed principles for librarians and therefore no need to cling to those which were determined by their prevailing social, cultural and technological contexts some time ago: these must be reinterpreted. According to

[1]

[2]

[37] “librarians have long come to constitute themselves within specific ideological paradigms and social programs, but often without any real critical engagement of those contextual framings, or consciousness of any need to maintain currency with ongoing social change”. [38] warn that, “When simple change becomes transformational change, the desire for continuity becomes a dysfunctional mirage,” and the dysfunctionality of librarianship was expressed by [39] as “its perennial and increasingly well-founded anxiety of irrelevance”. Therefore, librarians are to learn this new skill and disseminate in a scholarly manner in order to fulfill their roles as enablers of openness in their profession, also to transform information environment by displaying competencies as the custodian of knowledge to their users.

2. **Propagating and raising awareness:** promotion of the benefits of open science should take place in parallel with the development of tools and services, the incentives and recognition mechanisms that support excellence in science. Librarians can propagate within institutions to develop open access policies and roadmaps. This advocating will serve not only the researchers but the patrons and stakeholders at institutional level, the whole community and promote participation of citizens towards open science.
3. **Contributing to the development of research data management (RDM):** Policies and strategies to enable openness can be easily carried out by the home institutions through RDM.
4. **Giving Support to the Infrastructure:** libraries are to share research articles or data, including repositories; keeping with their involvement and responsibilities in the development and governance of repositories of publications and data, in regards to appraisal, selection, description and metadata applications, curation and preservation; information retrieval; monitoring data reuse, citation and impact.
5. **Giving Training and Support Researchers:** librarians are known to train scholars to open up their research workflows, share and reuse the research findings produced by other researchers. Aside the necessary research infrastructure, scholars need support at a practical level throughout the whole research cycle. Librarians offer themselves as a counsellor, guidance, trainer and provide services to the community which they served. The provision of information through exploratory stage of research; funding opportunities and requirements; bibliography and data management; applying metadata; identification of open research methods and tools for analysis; output sharing and publication; data citation, licensing and other intellectual property and copyright issues; preparation of data for deposit and long – term preservation of data among others. With this in mind, the librarians need to know their community research practices regarding information use, production, and sharing platforms, tools and services they use.
6. **Librarians should enlighten the public and users (scholars inclusive) to submit preprints to publicly available repositories:** In today scholarly communications, many journal editors allow the posting of researcher’s pre-prints to open repositories for instance, arxiv.org pending the submission and peer reviewing, comments and other improvements can be done to the paper before final publications [40].
7. **Scholars should be encouraged to publish in Open Access journals where possible:** today, many subscriptions based – journal options are available to authors to make their findings or data more accessible and discoverable [40]. For instance, some journals allow authors to post published articles in a public repository (such as Pubmed Central) and typically between 6 to 12 months after publications [41].
8. **Share Data and Materials:** librarians have the ability to change the perception of the users to share their data and materials for others to benefits from their works and findings for reproducibility and reusability. The code, methods and data to produce findings in researcher’s work should be made open and available for other to use so as to avoid duplication of research works [40].

The above roles require libraries and librarians to develop novel processes and skills to fulfil their functions in a digital age especially in the era of big data. Since the movement of open access in 2002 at the Budapest Open Access Initiative and Berlin Conference in 2003, the discussion on the roles of libraries, challenges of librarians on the current scholarly environment has been ongoing in the global arena. The debate of librarians and libraries to identify, define and defend their profession in the digital disruption era. Librarians need to move quickly with technology as constant needs to adapt to a changing environment of big data and the era of internet of things in order not to be left behind from what they are known for.

According to the study of [42], librarians need to fill the following skills gap in order to perform their roles in today era of big data and open science:

a. Ability to advise on preserving research outputs.
b. Knowledge to advise on data management and curation, including ingest, discovery, access, dissemination, preservation and portability.
c. Knowledge to support researchers in complying with the various mandates of funders, including open access requirements.

d. Knowledge to advise on potential data manipulation tools used in disciplines or subjects.
e. Knowledge to advise on data mining.
f. Knowledge to advocate and advise on the use of metadata.
g. Ability to advise on the preservation of project records (e.g. correspondence).
h. Knowledge of sources of research funding to assist researchers to identify potential funders.
i. Skills to develop metadata schema, and advise on discipline/subject standards and practices, for individual research projects.

There are calls to develop skills and career paths for various data – related professions that are essential to research institutions in a data intensive age: these include data analysts, data managers, data scientists, data curators and data librarians [43].

1.4 Challenges and Prospects of Libraries partnering with Open Science and Big Data Trends

a) **Changes in leadership and management to support Open Science and Big Data transition:** leadership of libraries and librarians must as a matter of urgency bring in the cultural change that will reflect open scholarly communication experts in their leadership roles so as to usher in the needed change libraries need to meet with open science policy. Leadership and management of staff and resources should happen in close, everyday interactions and learning is essential in keeping up with real time progress of open science and big data.

b) **Systematic Work and Determination is essential:** professionals that are essential to research institutions in a data intensive age are needed to make universities libraries multi-professional working communities. These include researchers in scholarly communications, teachers, data analysts, library programmers, data managers, data scientists, data curators and data librarians [43]. Involvement of these actors from various aspects of open science and big data will make cultural change to happen in the libraries and among the librarians.

c) **Collaboration with other universities** about happenings in the realm of open scholarly communications and big data issues in libraries. Strategic roles on openness in the library and new challenges must be worked out by the management of libraries [44].

d) **Moving to the Library Cloud:** in the era of big data and scholarly communications, clouds are seen as a solution, but they also throw up fresh challenges to librarians. Ironically, their proliferation can cause a bottleneck if data end up parked on several clouds and thus still need to be moved to be shared. And using clouds means entrusting valuable data to a distant service provider who may be subject to power outages or other disruptions which may causes problems for libraries' services to their clienteles [10].

Several other new applications that are becoming possible in the Big Data era and openness for librarians include [45]:

e) **Personalized services**: With more personal data collected, commercial enterprises can provide personalized services adapt to librarian preferences. For instance, Librarians can get all the targeted information at their disposal by predicting user's information need in a real time as well as analyzing the collected records.

f) **Internet security:** When a network-based attack takes place, historical data on network traffic may allow us to efficiently identify the source and targets of the attack. Librarians are more into people's oriented, but with the collaborations of IT guards, so a headache will be over.

g) **Digital humanities:** Currently numerous archives are being digitized. For instance, Google has scanned millions of books and identified about every word in every one of those books. This produces massive amount of data and enables addressing topics in the humanities, such as mapping the transportation system in ancient Roman, visualizing the economic connections of ancient China, studying how natural languages evolve over time, or analyzing historical events.

However, most researchers tend to download remote data to local hardware for analysis. But this method is "backward", says Andreas Sundquist, chief technology officer of DNAnexus. "The data are so much larger than the tools, it makes no sense to be doing that." The alternative is to use the cloud for both data storage and computing. If the data are on a cloud, researchers can harness both the computing power and the tools that they need online, without the need to move data and software (see 'Head in the clouds[3]'). "There's no reason to move data outside the cloud. You can do analysis right there," says Sundquist. Everything required is available "to the clever people with the clever ideas", regardless of their local computing resources, says Birney. Various academic and commercial ventures are engineering ways to bring data and analysis tools together — and as they build, they have to address the continued data growth. Xing Xu, director of cloud computing at BGI (formerly the Beijing Genomics Institute) in Shenzen, China, knows that challenge well. Therefore, librarians need to utilize

[3]

these tools for their organizational or institutional advantage.

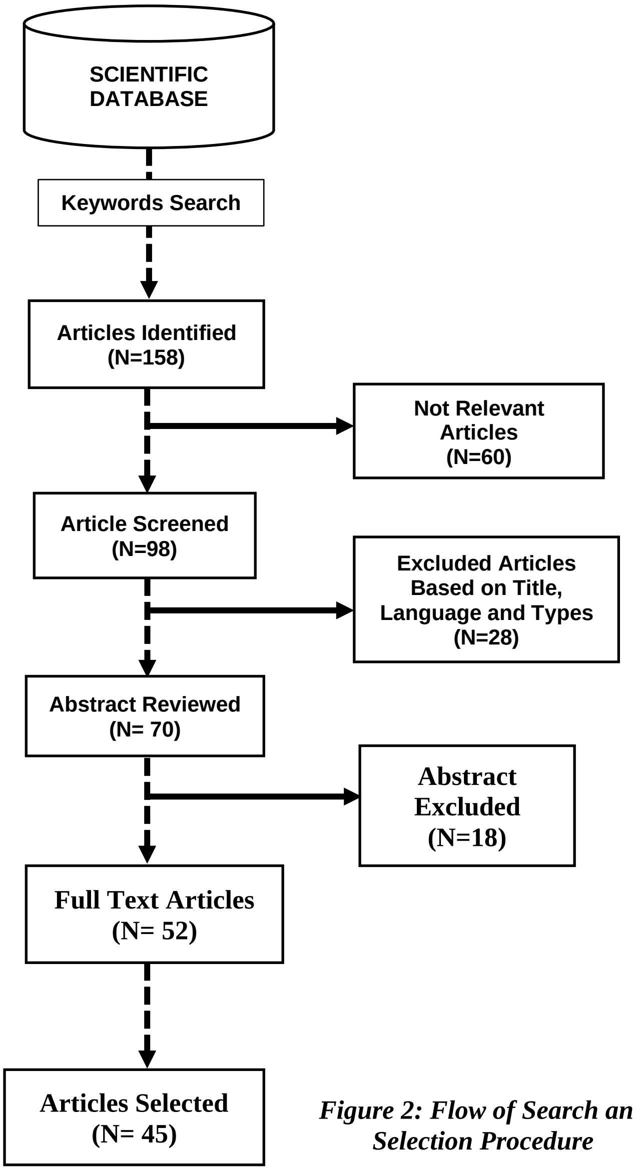
SCIENTIFIC DATABASE
Keywords Search
Articles Identified (N=158)
Not Relevant Articles (N=60)
Article Screened (N=98)
Excluded Articles Based on Title, Language and Types (N=28)
Abstract Reviewed (N= 70)
Abstract Excluded (N=18)
Full Text Articles (N= 52)
Articles Selected (N= 45)

Figure 2: Flow of Search and Selection Procedure

3. *Method and Materials*

3.1. Analytics of Big Data, Open Science with Librarians on Web of Science Database

To analyze the evidence of the flow of search and selection procedure for big data and openness in librarianship, we presented a tree table of big data-openness and librarianship based approach after rapidly examined 98 articles related to big data, open science and librarianship in Web of Science (WoS) (search was done in June 2019), with limitation to date of publication (2009 - 2019). The data was collected from the “Web of Science Core Collection” that constitutes of the Social Sciences Citation Index “SSCI”, Science Citation Index Expanded “SCI-EXPANDED”, Conference Proceedings Citation Index- Science “CPCI-S”, Conference Proceedings Citation Index- Social Science & Humanities “CPCI-SSH”, Arts & Humanities Citation Index “A&HCI”, and newly included Emerging Sources Citation Index “ESCI” that consists of articles of standard and acceptable quality [50], [51]. The first search attempt with the keyword “Big Data, librarians, Open Science, Scholarly Communication” as the title returned 158 articles, 98 articles were screened in this study while others are excluded. The citations source items indexed within the Web of Science Core Collection are reflected in this following report. 52 full-text articles were reviewed but 45 articles in all were relevant to our target, and others that are not so much relevant to Big Data, librarians, Open Science, Scholarly Communication were removed (*Figure 2*). Many of these articles are in (IEEE publications) and have passed through rigorous peer review. A classification tree table of big data-openness and librarianship-based approach was developed, and we later explained the rationale behind this novel approach in the subsequent section.

To calculate the results and to give us the precision and recall level for this study, we deployed a valuation measurement postulated by [47]. The assessment measurements are seen in F1 and accuracy. To appreciate these metrics, there is a 2x2 possibility table, which classify evaluation into two divisions a) (true positive – a section that was correctly selected/reviewed) or (false negative – a section that was incorrectly not chosen) and b) (true negative – a section that was correctly selected/reviewed) or (false positive – a section that was not correctly selected). Precision is the measurement of chosen review papers that are correct while recall is the opposite gauge, it is the measurement of correct papers chosen. Using Precision and Recall, the fact of high rate of true negative is not important any longer [47].

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive+False Negative}} \text{ ----- Formula 1}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive+False Negative}} \quad \text{----- Formula 2}$$

Analyzing Precision and Recall and knowing that they are opposite, the key concept here is the trade – off researchers must do in each measure, looking for the best metrics to evaluate their systems. Most of the information science researchers used Recall as a measure metric since it does not matter how the false positive rates are, if there are high true positive rates, the result will surely be good. However, in this study, maybe Precision would be better, to balance the trade off, the f measure is presented [47] [48] in the figure below:

$$F_1 = \frac{2\text{ (Precision .Recall)}}{\text{Precision+ Recall}} \qquad \text{------ Formula 3}$$

This measure implements a weighted score of assessing the Precision and Recall trade-off. According to [47] [48], the following are the metrics used to evaluate any systems, such as Mean Average Precision (MAP) – which is a standard measure for information retrieval,

$$\text{MAP}= \sum_{Q}^{P=1} \frac{\text{Ave P(p)}}{Q} \quad \text{------- Formula 4}$$

Where the AveP, Average precision, given by the Eq. 5.

Mean Reciprocal Rank (MRR) shows below and it is used to calculate the relevance

$$\text{AveP}= \sum_{N}^{k=1} \frac{\text{P(k) X rel (q)}}{(relevant)} \quad \text{------ Formula 5}$$

$$MRR \ = \frac{1}{N}\sum_{i=1}^{N} RR(qi) \quad \text{------ Formula 6}$$

3.2 Evaluation of Results

Table 2: Evaluation of Results

Criteria	Big data, openness and librarians	Data archiving, cloud computing and others	Total Relevant
Total number of Review papers	68	30	98
Review papers with keywords	60	8	68
Review papers with correct keywords	45	7	52
Precision	75.0%	87.5%	76.5%
Recall	66.2%	23.3%	54.7%

The system gave 98 reviewed papers downloaded from Web of Science in June 2019, of which 45 are relevant and correct on big data and openness of librarians, thus achieving 54.7% recall and 76.5% average precision. We noticed that big data/openness/librarian/open scholarly communication gives more precision than data archiving, cloud computing and other categories. This can be justified through the fact that big data and openness with librarians defines possible means of identifying it entities which helps in classifying the keywords used in the study. Surprisingly, Data archiving, cloud computing and others returns higher precision of 87.5% while recall was at 23.3%. This was in line with a similar study by Alagha [46]. However, his data is not available publicly to compare the performance of the two systems. Though, this approach is centred on the same method with Alagha [46].

4. Discussion

This section discusses the classification of the analysis presented in Table 3.2. Out of the 98 reviewed papers, only 60 mentioned Big data, openness and librarians while 8 emphasized on data archiving, cloud computing and others (See Table 2). From the 98 papers reviewed, 45(45.9 percent) of the papers implemented big data, open science and librarianship-based paradigm, while 7(7.1 percent) mentioned data archiving and related topics.

5. Conclusion and future works

In this study, we carried out an overview on the opportunities and challenges brought by big data to librarians and presents ways of making data available for everyone to use without limitations, with emphasis on the distinguished roles librarians need to play in the era of Big Data and Openness to scholarly communications. The result of the reviewed articles showed that 98 out of 158 papers have been identified as describing a big data and open science in relation to the librarians' roles which was evaluated using precision and recall methods. Future works may look at different domains that predict the accuracy of the information retrieving processes.

References

[1]. L. Stein "The case for cloud computing in genome informatics," *Genome Biol*, Vol. 11 p 207, 2010.

[2]. A. Swan, Y. Gargouri, M. Hunt, and S. Harnard, "Open access policy: Numbers, analysis, effectiveness. Preprint 2015." https://arxiv.org/abs/1504.02261. Accessed September 2017.

[3]. C. Anderson, "The end of theory: will the data deluge makes the scientific method obsolete?" Wired, 2008. Retrieved November 2019 from http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

[4]. E.N. Baker, "Data archiving and availability in an era of open science," (in English), *Iucrj,* Editorial Material vol. 4, pp. 1-2, 2017.

[5], IBM: What is Big Data?, 2015

[6] G. Moran. "We're in a Data Literacy Crisis. Could Librarians Be the Superheroes We Need?" *Tech: Human Intelligence* 2019.

[7]. R. Rice, and J. Southall. "The data librarian's handbook." *Facet Publishing.* Pp. 19-33, *2016.*

[8]. ICoLIS. "Romanticizing the library: creating what users need and want." *6th International conference on libraries, information and society,* ICoLIS, 2016.

[9]. J. Fan, F. Han, and H. Liu, "Challenges of Big Data Analysis." *Natl Sci Rev*., 1(2): 293–314., doi:10.1093/nsr/nwt032., 2014.

[10]. Nature, "The Big Challenges of Big Data." *NATURE*. Macmillan Publisher Limited. Accessed at http://pic.b.qs1401.com/42548/pdf/bigbioldata_nature13.pdf, 2013.

[11]. D. Donoho, "High-dimensional data analysis: The curses and blessings of Dimensionality," The American Mathematical Society Conference; Los Angeles, CA, United States, 2000.

[12]. V. Mayer-Schönberger, and K. Cukier. “Big Data: A revolution that will transform how we live, work, and think.” Boston: Houghton Mifflin Harcourt 2013.

[13]. H. Margetts, “Data, data everywhere: Open data versus Big Data in the quest for transparency.” In Transparency in Politics and the Media: Accountability and Open Government, edited by Nigel Bowles, James T. Hamilton, and David A. Levy, 167-78. New York: I.B. Tauris & Co. Ltd., in association with the Reuters Institute for the Study of Journalism, University of Oxford, 2014.

14]. D. Laney, “3D Data management controlling data volume, velocity and variety” | *BibSonomy.* Gartner Blog. https://www.bibsonomy.org/bibtex/263868097d6e1998de3d88fcbb76ca6/sb3000 2001.

[15]. C. S. Lewis, and O. Westlund, “Big data and journalism. Epistemology, expertise, economics and ethics. Digital Journalism 3(3): 447-66, Doi:10.1080/21670811.2014.976418, 2015.

[16]. S. Parasie, “Data –driven revolution? Epistemological tensions in investigative journalism in the age of “big data”. *Digital Journalism*, 3(3). 364-80, 2015.

[17]. E. M. Ferreras-Rodriguez, ‘Nuevos perfiles professionals: El periodista de datos. In Actas –IV conngreso International Latina de Communicacion Social –IV CILCS –Universidad de La Laguna, Diciembre 2012, 1-19, 2012.

[18]. T. Kuipers, and J. Van der Hoeven, “*Insight into digital preservation of research output in Europe: Survey report* (D3.4)”. Didcot, UK: PARSE.Insight p.37. Retrieved from http://www.parse-insight.eu/downloads/PARSE-Insight_D3-SurveyReport_final_hq.pdf, 2009.

[19]. E. Brown, “I know what you researched last summer: How academic librarians are supporting researchers in the management of data curation.” *The New Zealand Library & Information Management Journal*, 52(1), 55–69, 2010.

[20]. L. Markauskaite, M. A. Kennan, J. Richardson, A. Aditomo, and L. Hellmers, “Investigating eResearch: Collaboration practices and future challenges.” In A. Juan, T. Daradoumis, M. Roca, S. Grasman, & J. Fauli (Eds.), Collaborative and distributed e-research: Innovations in technologies, strategies and applications (pp. 1–33), 2012. http://dx.doi.org/10.4018/978-1-4666-0125-3.ch001.

[21]. G. Steinhart, J. Saylor, P. Albert, K. Alpi, P. Baxter, E. Brown, et al. “Digital re-search data curation: Overview of issues, current activities, and opportunities for the Cornell University Library. A report of the Cornell University Library DataWorking Group. Retrieved from http://ecommons.library.cornell.edu/bitstream/1813/10903/1/DaWG_WP_final.pdf (2008).

[22]. Association of Research Libraries. “E-science and data support services: A study of ARL member institutions.” Washington, *Association of Research Libraries*, 2010. Retrieved from http://www.arl.org/storage/documents/publications/escience-report-2010.pdf

[23]. F. M. Cheek, and P. S. Bradigan, “Academic health sciences library research support.” *Journal of the Medical Library Association*, 98(2), 167–171, 2012. http://dx.doi.org/10.3163/1536-5050.98.2.011

[24]. M. P. Newton, C. C. Miller, and M. S. Bracke,. Librarian roles in institutional repository data set collecting: Outcomes of a research library task force, 2010.

[25]. W. G. Potter, C. Cook, and M. Kyrillidou, “ARL profiles: Research libraries. Washington, D.C”: *Association of Research Libraries*, 2010. Retrieved fromhttp://www.arl.org/storage/documents/publications/arl-profiles-report-2010.pdf

[26]. A. Creamer, M. E. Morales, J. Crespo, D. Kafel, D., & Martin, E. R. “An assessment of needed competencies to promote the data curation and management librarianship of health sciences and science and technology librarians in New England. *Journal of eScience Librarianship*, 1(1), 18–26, 2012. http://dx.doi.org/10.7191/jeslib.2012.1006.

[27]. R. Cullen, and B. Chawner, “Institutional repositories, open access, and scholarly communication: A study of conflicting paradigms.” Journal of Academic Librarianship,37(6), 460–470., 2011.

[28]. R. S. Jeffrey, “The Open Science E-Framework: Improving Science by Making it Open and Accessible.” A Dissertation presented to the Graduate Faculty of the University of Virginia in Candidacy for the Degree of Doctor of Philosophy 2013.

[29]. B. Hunter, “The effect of digital publishing on technical services in university libraries.” The Journal of Academic Librarianship Volume, 39(1), 84–93., 2012.

[30]. McKee, Stamison and Bahnmaier, 2014, p.190).

[31]. OECD. Making Open Science a Reality. OECD Science, Technology and Industry Policy Papers, No. 25, *OECD Publishing*, Paris. Pp55-63. http://dx.doi.org/10.1787/5jrs2f963zs1-en.

[32] P. Suber, (2008). “Strong and weak OA. Retrieved from http://legacy.earlham.edu/~peters/fos/2008/04/strong-and-weak-oa.html 2012.

[33]. FOSTER. “Open science definition (2015). Available at = https://www.fosteropenscience.eu/taxonomy/term/100 Accessed: 10 March 2018.

[34] W. Gregg, C. Erdmann, L. Paglione, J. Schneider, and C. Dean, “A literature review of scholarly communications metadata.” *Research Ideas and Outcomes* 2019.

[35] Smith. D. W. (2013). Phenomenology. Stanford Encyclopedia of Philosophy. Retrieved from http://plato.stanford.edu/entries/phenomenology/

[36] OECD. OECD Reviews of Innovation Policy: Sweden (2013), *OECD Publishing*, Paris, http://dx.doi.org/10.1787/9789264184893-en.

[37] S. S. Myburgh, & M. A. Tammaro, "Education for Digital Librarians: Some European Observations" *In Library and Information Science Trends and Research*: Europe: 217-245, 2015. Available at https://doi.org/10.1108/S1876-0562(2012)0000006013.

[38] B. L. Hawkins, & P. Battin, (Eds.). "The mirage of continuity: Reconfiguring academic information resources for the 21st century." *Washington, DC: Council on Library and Information Resources and the Association of American Universities,* 1998.

[39] B. Frohmann, "Discourse analysis as a research method in library and information science." *Library and Information Science Research*, 16, 119–138, 1994.

[40] J. O. Hodonu-Wusu, "Open Science: A Review on Open Peer Review Literature. Library Philosophy and Practice (e-journal).2018, accessed at http://digitalcommons.unl.edu/libphilprac/1874.

[41] Farnham et al., "Early career researchers want Open Science." Genome Biology 18:221., pp1-4., 2017.

[42] M. Auckland, "Re‐skilling for research: An investigation into the role and skills of subject and liaison librarians required to effectively support the evolving information needs of researchers." London: Research Libraries UK 2012. Retrieved from http://www.rluk.ac.uk/wp-content/uploads/2014/02/RLUK-Re-skilling.pdf

[43] Science International. "Open data in a big data world." Paris: International Council for Science (ICSU), International Social Science Council (ISSC), The World Academy of Sciences (TWAS), Inter Academy Partnership (IAP)2015. <https://twas.org/sites/default/files/open-datain-big-data-world_short_en.pdf>. Accessed 18 Feb 2018.

[44]. H. Silvennoinen-Kuikka, "A strategy look at research support and open science services at our library", *Library Connect, 2018.* Accessed at https://libraryconnect.elsevier.com/articles/strategic-look-research-support-and- open- science-services-our-library

[45] N. Meng. "The big challenges of big data," 2013.

[46] I. Alagha, "AR2SPARQL: An Arabic Natural Language Interface for the Semantic Web" in International Journal of Computer Application (0975-8887), vol.125, no.6, pp.19-27.

[47] X. Yao, "Feature -driven Question Answering with natural language alignment." John Hopkins, University (Ph.D. Thesis) 2014.

[48] X. Li, et al., "Automatic question answering from web documents." 2007. Wuhan University. J. Nat. Sci. 12(5), 875-880. https://doi.org/10.1007/s11859-007-0046-4.

[49] Zhang, Z & Liu, G. (2009). "Study of Ontology-Based Intelligent Question Answering Model for Online Learning". In Information Science and Engineering (ICISE), pp. 3443-3446.

[50] C. Zhu, T. Jiang, H. Cao, W. Sun, Z. Chen, & J. Liu, "Longitudinal analysis of meta- analysis literatures in the database of ISI Web of Science." International Journal of Clinical and Experimental Medicine, 8(3), 3559–3565, 2015.

[51] F. Mukhlif, O. J. Hodonu-Wusu, A. K. Noordin, and M. Z. Kasirun, "Major Trends in Device to Device Communications Research: A Bibliometric Analysis" 2018 IEEE 16th Student Conference on Research and Development (SCOReD), Bangi, Malaysia (26-28 Nov 2018).

Footnotes

1. http://datalib.edina.ac.uk/mantra/organisingdata.
2. www.bodleian.ox.ac.uk/data
3. 'Head in the clouds' https://wenku.baidu.com/view/ec5767c2960590c69ec376ff.html. *Nature*, p. 258, Vol. 498,