April 2021

Bibliometric Review of Digital Archive Research: Contemporary Status, Research Hotspots and Future Trends

ABU KS
abumutd@gmail.com

Bibliometric Review of Digital Archive Research: Contemporary Status, Research Hotspots and Future Trends

Abu KS

Abstract

Digital archiving has been practised more and more in recent years, because of its significant potentials and long-term benefits. Bibliometric based review was applied to examine the digital archive research publications from the web of science from 1989 to 2019. A total of 639 publications were obtained with a steady increase of publications every year. There were diversified research topics, which were categorized as computational aspects of digital archiving, libraries role in digital archiving and digital archiving practices in medicine. Keyword analysis highlighted the interdisciplinary approach of digital archiving practices. The study also aids in summarizing the status quo and growth patterns of digital archive research.

Keywords: Digital Archive, Bibliometrics, Web of Science, Vosviewer, Research Hotspots, Growth Pattern

1. Introduction

Digital objects created and disseminated by various stakeholders have been increasing exponentially on a daily basis. Social media has facilitated this ease of dissemination without much emphasis on long-term preservation. In a single minute, millions of digital objects are created and lakhs of contents are being consumed[1]. This enormous growth of digital content and also the fragile nature of digital objects have imposed the need for long term preservation strategies. Digital Archiving is a commonly practised technique for preserving digital objects. A digital archive is a storehouse of digital objects with the aim of preserving and disseminating to the beneficiaries as and when needed[2]. Physical aspects, long term retention and easy accessibility are kept in purview while preserving digital objects[3]. Though digital preservation and digital archiving are used synonymously, they are different in their purposes, where the former is focused on providing access to digital information for longer duration[4] and the latter gives more emphasis to description and access by providing appropriate findings aids. Digital archiving is practised in diverse domains for preserving digital objects; some of the domains are education[5-6], linguistics[7], forensic science[8], library collections[9], literature[10], history[11], astrophysics[12], physics[13], humanities[14], computing [15] and so on. The use of the archival practice in different domains indicates their apparent need to preserve their digital content for long term storage and access.

Bibliometric is a significant tool for assessing the research productivity of institutions, countries, authors and sources[16]. There are several bibliometric studies pertaining to the present research. These studies focused on related domains such as digital literacy[17], digital innovation[18], digital libraries[19], institutional repositories[20], digital humanities[21] and digital repositories[22]. According to our knowledge, no bibliometric study focused specifically on digital archives research productivity and considering this research gap, the present study was carried out to review digital archives research using bibliometric methods. Usually, bibliometric studies depict publication counts, languages, sources etc., without giving much significance on research contents[23]. Combining traditional bibliometric measures with research content analysis will help the researchers in understanding the progress of that particular field over different periods of time[24]. This study aims at filling this void by integrating general bibliometric measures and also giving emphasis on key research contents to assist the readers to understand the progress of the digital archive research at different stages.

The remainder of the paper is structured as follows: Section 2 provides methods and basic setup by presenting the source of data, method of data collection and the extent of analysis. Section 3 presents the detailed discussion of the results and finally, section 4 summarizes the findings and concludes the paper.

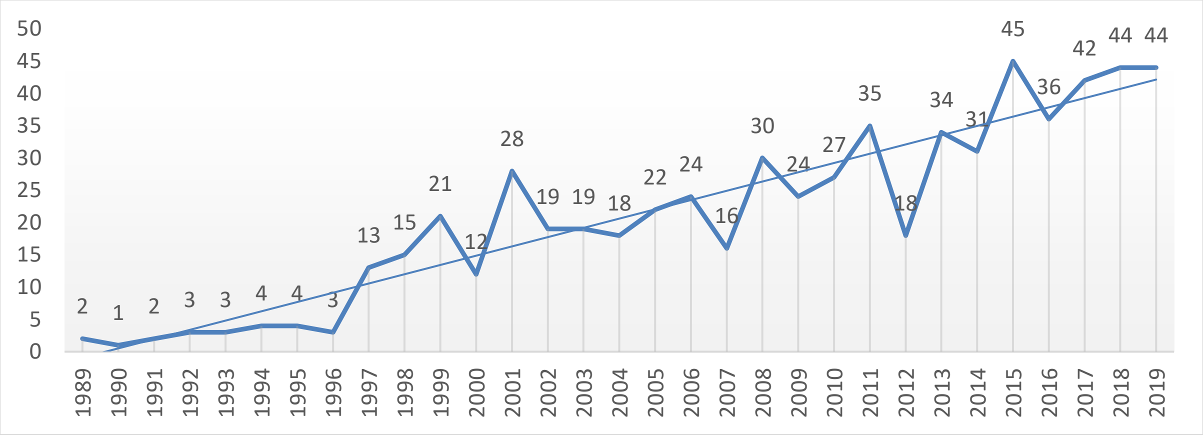


Fig.1: Number of publications

2. Fundamental Setup and Methods

The research adopted a bibliometric analysis method for its meritorious nature of highlighting the publication process and identifying research hotspots[24-25]. Since there is a clear conceptual definition of the digital archive, the same keyword (without any variations) was used in the title field to search the web of science database. The researcher chose the web of science database primarily because of its rigorous peer-reviewing process and also considering its reputation in the academic journal community[26].

For ensuring relevancy of publications, several criterions were imposed. First, all the publications of the researchers should be included in the web of science core collection. Secondly, the publications should be indexed in any one of the following indices: Science citation index-expanded (SCIE), Social science citation index (SSCI) and Arts & Humanities citation index (A&HCI). The publications fall in the range from 1989 to 2019.

A total of 639 publications have been included in the study. The year-wise distribution of publications has been plotted in Figure 1. A strong mounting trend can be seen during the study period. Apparently as divergent to slow and steady growth before 1996, we can see the sharp enhancement in the number of publications after 1996, which indicates significant interest from the researchers on digital archives since the report of the task force on archiving of digital information in 1996.

3. Results and Discussion

The present bibliometric study is focused on answering the following four questions: In which sources these publications are published? Where these researchers come from (country)? How these publications are cited? What are the prominent research hotspots and future trends?

Source Titles	**Number of Publications**

Lecture notes in computer science	19
Journal of digital imaging	15
Electronic library	13
Library journal	13
Zeitschrift fur bibliothekswesen und bibliographie	13
Archives and records the journal of the archives and records association	12
Library hi-tech	10
Program electronic library and information systems	9
Convergence the international journal of research into new media technologies	8
Digital journalism	8
Journal of the society of archivists	8
Digital scholarship in the humanities	6
Library quarterly	6
Library resources technical services	6
Moving image	6
Research and advanced technology for digital libraries	6
Smpte motion imaging journal	6
Stahl und eisen	6
Information research an international electronic journal	5
Journal of academic librarianship	5
Journal of documentation	5
Journal of the audio engineering society	5
Nfd information wissenschaft und praxis	5
Profesional de la informacion	5

Table 1: Source Distribution (Sources with publications over 4)

3.1. Distribution of Sources

Table 1 presents the list of sources according to their number of publications (more than four publications per source) in our sample. Clearly, digital archive is a field of computer science, there is no surprise in the top main sources as they turn out to be from the discipline of computer science and digital imaging. Moreover, one interesting aspect is that, out of these 24 journals, more than 15 journals deal with libraries, documentation and archival studies. This clearly indicates that the libraries are strongly associated with the preservation and archival process of digital objects. These top 24 journals have contributed 30% of the total publications. To get an overview of the major research terms published by these journals, term analysis was carried out using vosviewer. The analysis showed that totally 3284 terms were included in both the title and abstract field of these journals. To identify the key research terminologies, the threshold was set as 5, which implies a term must have at least 5 occurrences to be included in the analysis. As a result, only 140 terms comply with this parameter and all the 140 terms were included for further analysis. These 140 terms are grouped into 4 clusters based on their weight (links) and score attributes (Fig 2). Cluster 1 consists of 48 terms, cluster 2 with 38 terms, cluster 3 with 33 and cluster with 21 terms respectively. In cluster 1, terms like digital library, library, librarian, research, study and practice are grouped; In cluster 2, terms such as access, archive, archives, digital archive and digital preservation have the strong association with each other. Similarly, in cluster 3, communication, communication systems, data, digital imaging, medicine, constructions are grouped and finally in cluster 4, application, benefit, framework, strategy and long-term preservation are grouped together. Based on the term analysis of these sources, we can say that these sources have published research related to digital archive in the following four broad categories: roles of libraries in digital archiving, methods of preserving and accessing digital objects, applications of digital archive in diverse fields and the framework and strategies for ensuring long-term preservation of digital collections.

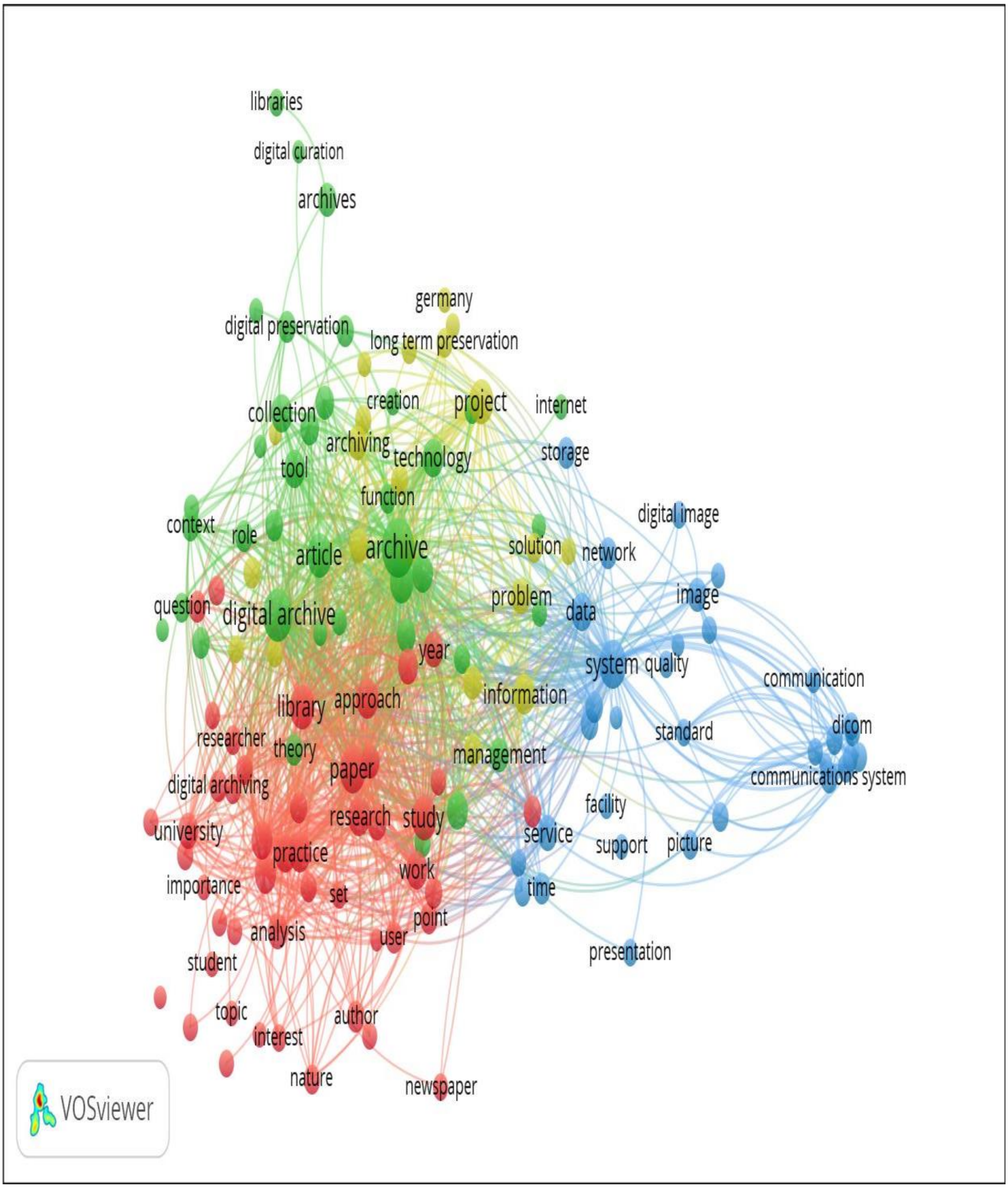
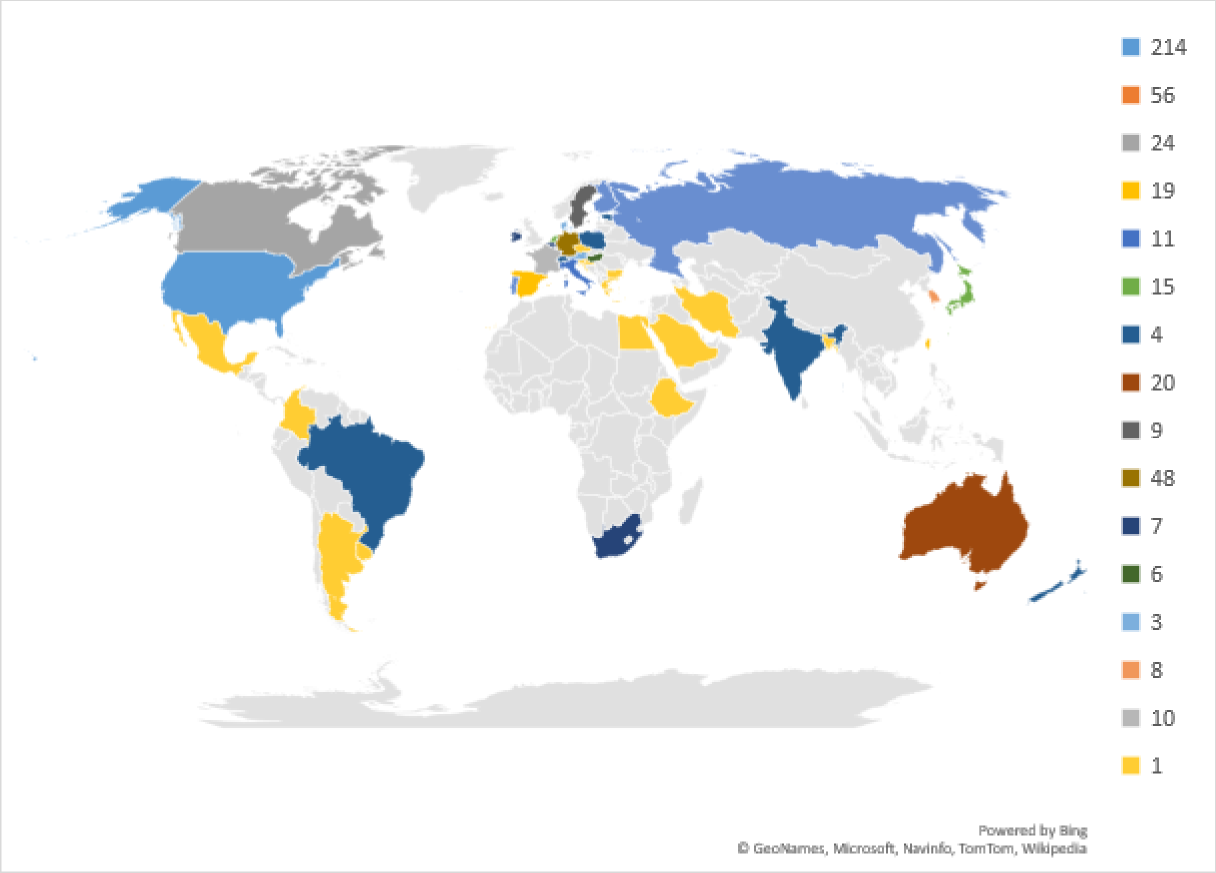


Fig 2: Co-occurrence map of sources based on terminologies.

3.2. Researchers' country of Origin



Powered by Bing
© GeoNames, Microsoft, Navinfo, TomTom, Wikipedia

Fig 3: Researchers' country of origin

It is additionally worth investigating where the research on digital archive originates from. It is done by associating the authors with their countries of publications. The sample revealed that a total of 46 countries contributed to digital archive research publications during the period of study. It provides a clear picture that a major contributor to digital archive research is USA (Fig 3). This is not a surprising fact as USA played a prominent role in the development of digital preservation and archiving. Moreover, they have many associations specifically dedicated to the development and growth of digital archive research and also acting as the headquarters for many influential international organizations. The second highest contribution is from England which is not near to half of the contributions of USA. This clearly indicates the dominance of USA in the field of digital archive research. The co-authorship analysis of authors based on their countries (Fig 4) reveals that England has the maximum collaboration with 9 countries, followed by USA with 8 countries. In terms of citations, it not surprising that USA has the maximum citations but with an average citation of 5.2 per publications. One interesting analysis is that though India has published only 4 publications, they have secured 62 citations with an ACP of 15.5.

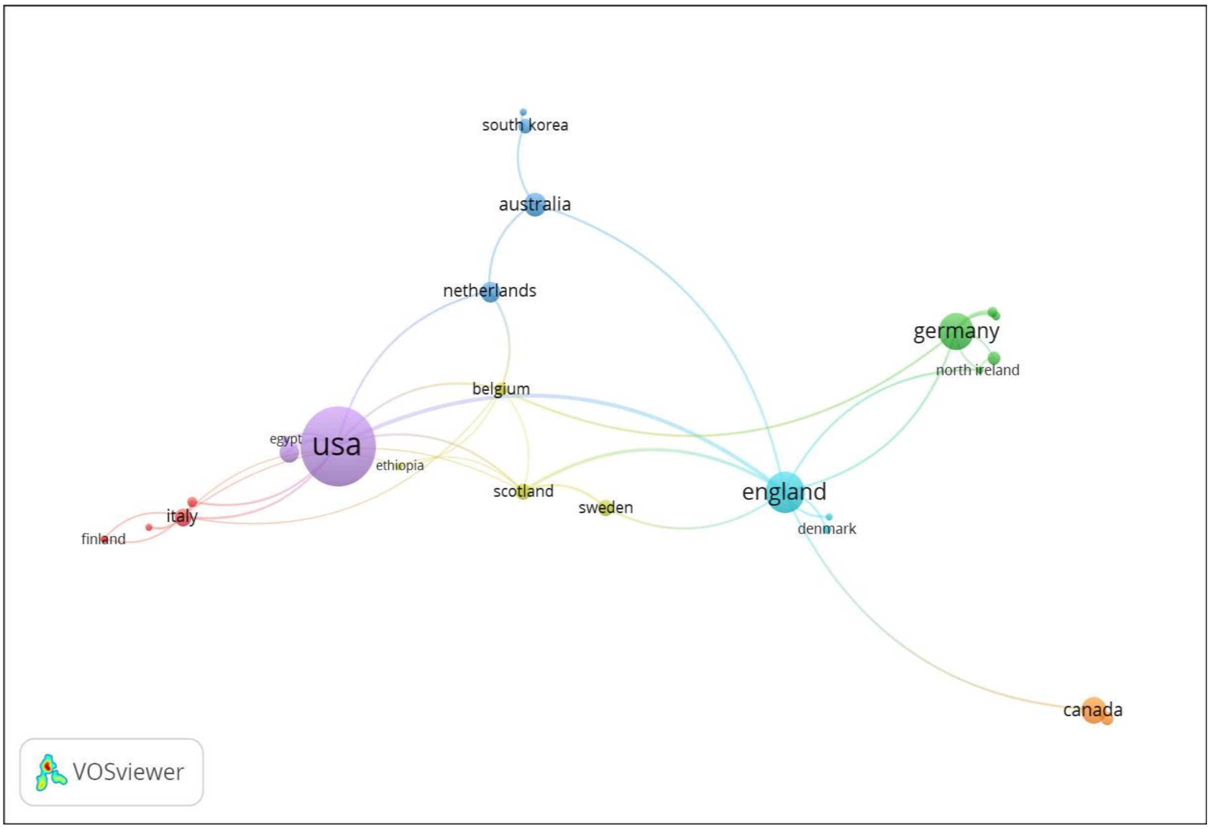


Fig 4: Co-authorship analysis of researchers based on their countries

3.3. Citation Analysis

Publications on digital archive have received decent citations. The total number of citations received by 639 publications is 2364 with an ACP of 3.69. Interestingly, only 296 documents have acquired all the 2364 citations and the remaining 343 publications didn't receive even a single citation. Although citations take considerable time to accumulate, 200 publications out of 343 were published before 2014, which indicates these publications have been overlooked consistently. The highest citation is 274 on Springob, CM et.al (2005). The top ten highly cited publications are listed in Table 2. These ten publications indicate the application of digital archive in diverse subjects, which denotes the interdisciplinary nature of the domain.

Author	Journal	Year	Cited Count	Title
Springob, CM et.al	Astrophysical Journal Supplement Series	2005	274	A digital archive of HI 21 centimeter line spectra of optically targeted galaxies
Beer, D et.al	Theory culture & society	2013	93	Popular Culture, Digital Archives and the New Social Life of Data
Tivy A et.al	Journal of Geophysical Research-Oceans	2011	86	Trends and variability in summer sea ice cover in the Canadian Arctic based on the Canadian Ice Service Digital Archive, 1960-2008 and 1968-2008

Gutman DA et.al	Journal of the American Medical Informatics Association	2013	57	Cancer Digital Slide Archive: an informatics resource to support integrated in silico analysis of TCGA pathology data
Huisman, A et.al	Human Pathology	2010	52	Creation of a fully digital pathology slide archive by high-volume tissue slide scanning
Panagos, P et.al	International Journal of Digital Earth	2011	51	European digital archive on soil maps (EuDASM): preserving important soil data for public free access
Wang, JH and Langer, S	Journal of Digital Imaging	1997	47	A brief review of human perception factors in digital displays for picture archiving and communications systems
Deacon, D	European Journal of Communication	2007	42	Yesterday's papers and today's technology - Digital newspaper archives and 'Push button' content analysis
Gehl, R	International Journal of Cultural Studies	2009	41	YouTube as archive Who will curate this digital Wunderkammer?
Given, LM	Library Quarterly	2010	37	What's Old Is New Again: The Reconvergence Of Libraries, Archives, And Museums in The Digital Age

Table 2: Top cited articles in digital archive research

3.4. Research Hotspots and Trends

The growth and development trends of a particular field can be obtained by analysing the keywords[27]. Similarly, keyword co-occurrence is on the prominent ways to perceive the research hotspots and future instances of a particular field of study[24]. Hence, keyword co-occurrence analysis was carried out using vosviewer software to investigate the existence of correlation among keywords. Vosviewer can be used for developing and visualizing bibliographic networks[28]. There were totally 1102 keywords in all the publications for the given time period. All the 1102 keywords were taken for co-occurrence analysis, which was grouped into 84 clusters. Specifically, there were 3 clusters with more than 35 items which implied the recent research hotspots. The maximum number of keywords in the highest cluster is 43. The largest cluster was labelled as computational aspects of digital archiving, focusing on keywords such as algorithms, big data, content analysis, machine learning and so on. The second cluster was categorized as libraries role in digital archiving, as it focuses on academic libraries, digital repositories, digital storage, library and information management, national libraries and public libraries. The third cluster was branded as digital archiving practices in medicine, emphasising on breast cancer, chest images, cancer, digital radiography, coronary angiography, radiotherapy, radiology, DNA and lung cancer.

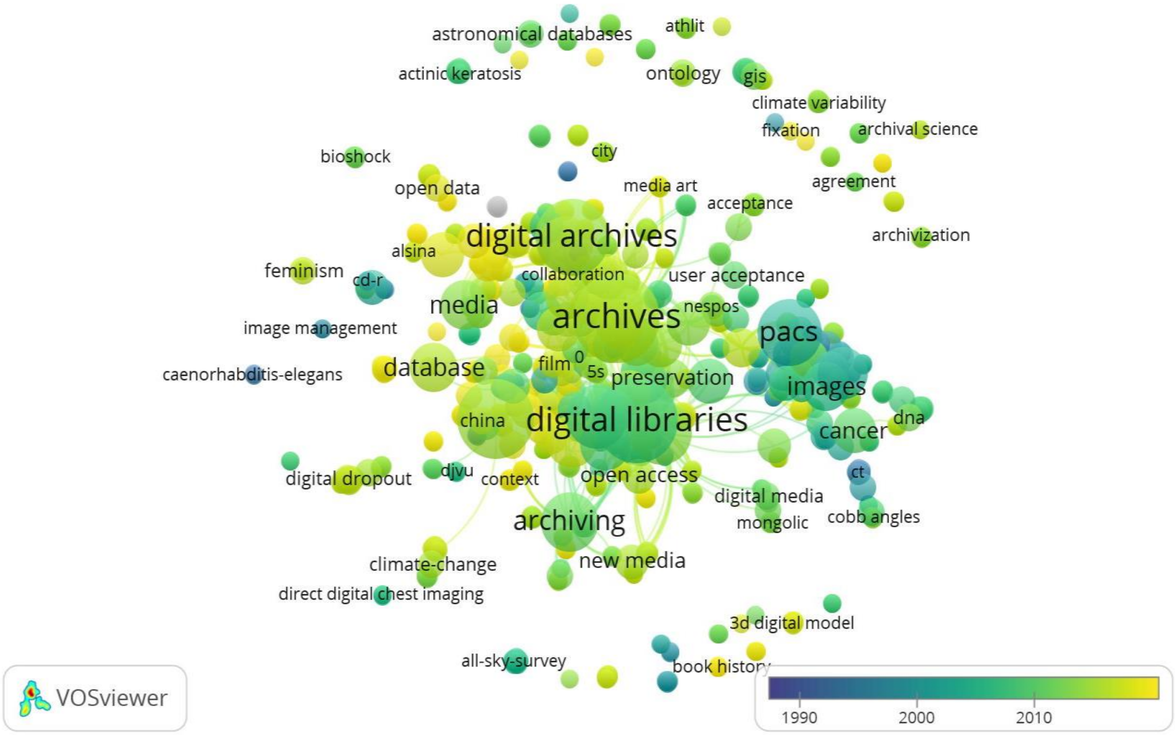


Fig 5: Keyword Co-occurrences of digital archive research

The growth and development of digital archive studies through the decades were examined through three stages: The first stage was from 1989-1999, with the burst keywords consist of chest images, radiology, picture archiving, implementation, health level and picture archiving and communication system (pacs). The second stage was from 2000-2009, with the keywords growing, apart from the health sector, the practices of digital archiving was extended to other domains such as astronomy, earth sciences and library and information sciences. In this stage, studies were focusing on the role of libraries in archiving digital collections and digital archival practices in domains such as astronomy and earth sciences. The final stage was from 2010 to 2019, other than the topics during 1989-2009, the study further expanded to the contribution of digital archive to big data analytics, machine learning, algorithms, data and text mining. During this period studies mainly focused on computational aspects of digital archiving practices.

4. Conclusion

The study employed a simple bibliometric method to examine the existing trends and research hotspots of digital archive research. Based on the ranking examination and visualization patterns on numerous aspects on the 639 publications, the study found some significant information that helps to draw a precise representation of digital archive research.

Examination of sources highlighted that the top contribution of publications was from computer science and majority of the top sources are from libraries, documentation and archival domains. The analysis of contributors' country of origin revealed that the research publications of digital archive are dominated by researchers of developed countries. Investigation of keywords on the publications prompted that, digital archive is of interdisciplinary nature by broadening its application in diverse subject domains.

The development of digital archive research was phased into three stages: During the initial stage digital archiving practices in medicine, in the second stage the research explored the role of libraries in archiving digital collections and finally, in the last stage the study deepened into the computations aspects of digital archives and their analytics in deep learning and artificial intelligence may be future research directions.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Statistia. Media usage in an online minute. *Statistia.,* 2019, https://www.statista.com/statistics/195140/new-user-generated-content-uploaded-by-users-per-minute/ (accessed on April 15, 2020).
2. University, SHS. What are Archives and Digital Archives?. *University, SHS.*, 2019, https://shsulibraryguides.org/c.php?g=86819&p=558330 (accessed on April 15, 2020).
3. CINES. What is: digital archiving?. *CINES.*, 2016, https://www.cines.fr/en/long-term-preservation/a-concept-problems-2/ (accessed on April 15, 2020).
4. Bodleian Libraries. Introduction to Digital Preservation: What is Digital Preservation?. *Bodleian Libraries.*, 2018, https://libguides.bodleian.ox.ac.uk/digitalpresrvation/what isdp (accessed on April 15, 2020).
5. Analeigh E.Horton. Kathryn Comer, Michael Harker & Ben McCorkle. The Archive as Classroom: Pedagogical Approaches to the Digital Archive of Literacy Narratives. Science Direct, Computers & Composition Digital Press, Colorado, 2020.
6. Lee, K. H. Y. & Patkin, J. G. Building the digital archive of Hong Kong english learning: Methodology, challenges and reflection. *System.*, 2017, **65,** 61-68.
7. DeWispelare, D. Cyberformalism: Histories of Linguistic Forms in the Digital Archive. *Modern Philology.*, 2019, **117**(1), E15–E18.
8. Montoya Mogollón, J. B. & Rodríguez, S. M. T. Diplomatic forensics science: historical review for approaching the born-digital archive record. *Investigacion Bibliotecologica.,* 2019, **33**(78), 47-62.
9. Heinmaa, H. Special collections of printed music in the digital archive of the National Library of Estonia. *Fontes Artis Musicae*., 2017, **64**(2), 175-192.
10. Dillon, E. M. Translatio studii and the poetics of the digital archive: Early American Literature, Caribbean Assemblages, and freedom dreams. *American Literary History.*, 2017, **29**(2), 248-266.
11. Bishop, C. The serendipity of connectivity: Piecing together women's lives in the digital archive. *Women's History Review.*, 2017, **26**(5), 766-780.
12. Springob, C. M., Haynes, M. P., Giovanelli, R. & Kent, B. R. A Digital Archive of H I 21 Centimeter Line Spectra of Optically Targeted Galaxies . *Astrophysical Journal Supplement Series.,* 2005, **160**(1), 149-162.
13. Crease, R., Graham, E. & Folsom, J. Database thinking and deep description: Designing a digital archive of the National Synchrotron Light Source. *Digital Scholarship in the Humanities.*, 2019, **34**(1), i46–i57.
14. Lee, B. C. G. Machine learning, template matching, and the International Tracing

Service digital archive: Automating the retrieval of death certificate reference cards from 40 million document scans. *Digital Scholarship in the Humanities.,* 2019, **34**(3), 513–535.

15. Wu, Z., Xie, J., Lian, X. & Pan, J. A privacy protection approach for XML-based archives management in a cloud environment. *The Electronic Library., 2019,* **37**(6), 970–983 .
16. Tang, M., Liao, H. & Su, S. F. A Bibliometric Overview and Visualization of the International Journal of Fuzzy Systems Between 2007 and 2017. *International Journal of Fuzzy Systems.*, 2018, **20**, 1403–1422.
17. Alagu, A. & Thanuskodi, S. Bibliometric analysis of digital Literacy research output: A global perspective. *Library Philosophy and Practice*., 2019, 1-20.
18. Zhang, X. *et al.* A Bibliometric Analysis of Digital Innovation from 1998 to 2016. *Journal of Management Science and Engineering.*, 2017, **2**(2), 95-115.
19. Singh, G., Mittal, R. & Ahmad, M. A bibliometric study of literature on digital libraries. *The Electronic Library*., 2007, **25**(3), 342-348.
20. Bhardwaj, R. K. Institutional repository literature: A bibliometric analysis. S*cience & Technology Libraries*., 2014, **33**(2), 185-202.
21. Wang, Q. Distribution features and intellectual structures of digital humanities: A bibliometric analysis. *Journal of Documentation.*, 2018, **74**(1), 223-246.
22. Valetutti, L. Cultural Heritage Preservation in Digital Repositories: A Bibliometric Analysis. *SLIS Connecting.*, 2015, **4**(2), 9.
23. Wang, H., Liu, M., Hong, S. & Zhuang, Y. A historical review and bibliometric analysis of GPS research from 1991-2010. *Scientometrics.,*2013, **95**, 35–44.
24. Li, D. *et al.* Biochar-related studies from 1999 to 2018: a bibliometrics-based review. *Environmental Science and Pollution Research volume.*, 2020, **27**, 2898–290.
25. Persson, O., Glänzel, W. & Danell, R. Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies. *Scientometrics*., 2004, **60**, 421–432.
26. Wang, J. & Wang, S. Preparation, modification and environmental application of biochar: A review. *Journal of Cleaner Production*., 2019, **227**(1), 1002-1022.
27. Hotspots in research on the measurement of medical students' clinical competence from 2012-2016 based on co-word analysis. *BMC medical education.,* 2017, **17**(1), 162.
28. Wang, W., Liu, J., Xia, F., King, I. & Tong H. Shifu: Deep learning based advisor-advisee relationship mining in scholarly big data. *In* 26th International Conference on World Wide Web Companion, 3-7 April 2017, Perth, Australia. 2017. pp. 303–310. https://doi.org/10.1145/3041021.3054159 (accessed on May 20, 2020).