Archiving Archives

Literature Review

Callum Fraser
Computer Science
University of Cape Town
Cape Town South Africa
FRSCAL004@myuct.ac.za

ABSTRACT

Digital archives are crucial in preserving digital items of the past and present for use in the future. However, they are harder and more complex to build than most people believe, and much must go into their development. Digital archives must overcome very different challenges to traditional ones, such as creating trust in their preservation and building to evolve with technology. This literature review performs an analysis on the current developments occurring in the digital archiving space and what it takes to build one. Concentrating on the aspects of preservation, modern digital archiving, and the tools and cases of actual application. From this, the key takeaways of this review indicate the untapped area of possibility regarding the creation of digital archives that can preserve other digital archives. This solution may alleviate many of the concerns currently inherent in the current process by adding additional redundancy and consolidation of archives.

KEYWORDS

Archiving, Archives, Digital Preservation, Web Archives, Digital Libraries, Clients-side Archives

1 INTRODUCTION

Digital Archives preserve digital documents of historical importance for future generations. They have many benefits over traditional archives but also bring their own challenges [5, 22]. A lot of work has been done to try and establish the most effective way to digitally preserve knowledge and to do so for a period at least comparable to existing preservation methods. Modern computing is very new, and the dramatic speed of change is creating uncertainty around the digital preservation process and which methods will last the test of time [1]. Digital archives do not only face concerns around the underlying technology but also on the organisations maintaining them, as they have great liberty in the methods and accuracy of the preservation applied [5]. This results in the existence of the archive depending on continued funding and commitment of its maintaining organisation.

A digital archive must provide a way of adding, organising, cataloguing, and easily retrieving large amounts of information through a user-interface, which overlaps with the services of a Digital Library [3]. Archives are separate to digital libraries in their explicit mission of preservation, which typically involves the use of well-defined techniques to prevent the original artifact from deteriorating over time [5]. Digital preservation encompasses the managed activities required for maintaining the computer bytes required to effectively reproduce the original content in an accessible way that can continue long into the future and with changing technology [4]. However, concerns still exist whether true digital preservation can be a reality [1]. This puts digital archives at the forefront of computer science research.

Archives need to create trust in their ability to preserve information, no matter the funding afforded to the managing organisations [2]. Thus, the possibility arises for establishing research around the creation of a new breed of digital archives that preserve other digital archives in the hope of creating trust, versioning, and long-term persistence of data through the public conservation of other archives. This literature review attempts to establish an understanding of the state-of-the-art practices around digital archiving to determine the potential to create and build an archive of other digital archives. First, an overview of the current methods used in digital preservation is considered. Second, the key aspects and innovations around the use of digital archives is discussed. Finally, a breakdown is done of the current tools used to create digital archives as well as case studies of what is done in practice.

2 DIGITAL PRESERVATION

2.1 Preservation Consistency

The core function of an archive is to persist historical information for as long as possible and many new forms of doing so have been developed [1, 22]. The aim for digital preservation is for information to be consistent and generate trust in the accuracy of its stored information. Trust is what makes people willing to store historically significant data in a particular archive and use its content without worry of compromise [2, 22]. Building trust for digitally stored data to ensure the reliability of the historical information is no easy task as it can be very tricky to determine the original object once data been copied, which is important for tracking objects changes over time and detecting modification [2]. Jantz and Giarlo [2] have presented methods around using digital signatures to detect modification and specify criteria for determining an original to ensure consistency of the item stored. To determine originality their method requires defining special characteristics in an original objects metadata, which can be

reverified and provide assurances that the original object has not been changed. This can allow for the establishment of versioning of objects. This work also leverages development around Persistent Identifiers, which can allow stored objects to remain accessible via citation, allowing academics to maintain faith in their choice of archive [2]. Audit Trails have also been presented as a way of maintaining an uninterrupted record of a digital object's life [2]. Recently, blockchain has also been presented as a different novel way of providing trust and consistency in preservation [7, 23]. Blockchain has become exceedingly popular and offers certain use cases around being tamper-proof, easily traceable, and decentralised, making it suitable for the protection and persistence of digital archives [7]. A potential blockchain solution has been presented by Lin et al. [7], which provides the foundation for a public and dependable digital archive system that removes the dependence on centralised cloud platforms or client hardware. This is achieved by focusing on the blockchain characteristic of digital file safety and consensus of an item, removing the need for typical backup processes. The archive works by having two blockchain - one chain that records all the archive's management information and another that tracks all archive usage data [7]. This system is implemented by getting groups of archive managers to work together to further persist their archives by having all metadata on a consortium blockchain. An alternative use case of blockchain is presented with the TrustChain model that focusses on certifying the contents of an archive using a digital signature certification chain [23]. This approach came from a desire to maintain the long-term preservation of digitally signed records by addressing the short-term validity of the current European standard for digitally signed documents [23]. The TrustChain model was established to confirm that the digital signature used on data remains valid and could be verified for extended periods of time. It successfully created a method of guaranteeing reliability through a single chain but only if the format of objects remains constant.

2.2 Extending Digital Preservation

Over time more diverse types of data is being persisted and new approaches for guaranteeing the continuous development and technical maintenance of this data is being investigated [22]. A typical problem of digital archives is that in the long run the underlying systems that power them can become unsupported as technology develops, presenting high security risks and making the software incompatible with modern hardware [6]. This can risk losing valuable historical information and makes it imperative that archives are able to develop over time while maintaining the accessibility and preservation of their stored digital objects [22]. Fedora is an architecture system that was designed for extensibility [20]. It provides an alternative to monolithic digital archive systems and is implemented as a set of web services that provide full programmatic management of digital objects and search [20]. It uses a service-oriented architecture (SOA) that defines a way to make software components reusable and interoperable via service interfaces. Instead of an SOA, Mayo et al. [6] have suggested a solution to the problem of extensibility by using microservices as an alternative to traditional monolithic systems, as this promotes customisation and decreases risks of failure by implementing independent applications. By applying this modular approach, parts of the system can get replaced by newer systems that have more active development without compromising the overall system. Microservices can allow for long-term support, transparency, and customisability, while meeting all functional requirements [6]. SOA is a means of breaking up monolithic applications into smaller components, while microservices provide a smaller, more fine-grained approach to accomplishing the same objective [21]. The microservices architecture is generally simpler to manage than a SOA due to having multiple independent services that work together rather than a single overarching scope to connect the components [21].

2.3 Preservation Policies

The active maintenance of digital archives is crucial for persistence and dictated by the policies implemented by the managing organisation [1]. Sustainability has recently become a major criterion for persistence because of its environmental impact and need to store knowledge in a way that is inclusive of all [8, 5]. New methods for creating environmentally sustainable preservation are being proposed. These approaches look at novel ways of making preservation more efficient to prevent excessive hardware infrastructure requirements [8]. Solutions include ways of reducing the number of redundant copies of data and setting criteria for determining and monitoring acceptable data loss levels, given the limited resources and environmental impact of perfect preservation. It has also been suggested to use efficient hardware which atomically checks for data integrity rather than having continuously running software doing it. On top of environmental sustainability, many other ethical responsibilities are defined in the policies of digital archives that consider how information is stored, distributed, and respected, and will empathise with the communities the historical information came from [5]. Digital Preservation faces unique constraints in Africa due to its differences in access to technology, and infrastructure compared to the rest of the world [12]. Africa has specific challenges around the threat of artefact deterioration, which could wipe out certain histories, and previous issues around suppressed history by colonial governments [12]. Many African universities that host archives try to implement long term digital preservation and develop their own policy to do so [9]. Yet almost none have the long-term funding, or the necessary full-time technical staff required for the long-term management of their archives in line with their own policy [9, 22]. This creates risks to these archives and highlights the need for long-term funding to maintain and support digital preservation.

2.4 Digital Preservation Discussion

Clearly digital preservation is a broad area that encompasses many different topics and issues, such as trust, evolution, and sustainability. Two novel blockchains have been presented for preserving data to provide trust through reliability. One focuses on

tracking all activities of the blockchain and having a consortium to manage it [7]. The other focuses on the signing of documents and tracking it to ensure long-term reliability [23]. Additionally, a non-blockchain method of signing documents has been mentioned as an alternative that can also track an object over its lifespan using audit trails [2]. This all show that guaranteeing trust in a digital archive is easier than ever and many practices exist for implementing it.

Planning for changing technologies is important for a digital archive. Two architectures have been examined that focus on ensuring that a digital archive can be extended over its lifetime so that it can manage technological changes. One applies a service-orientated architecture using the Fedora system to allow modularity of different components, while the other achieves the same goal with a simpler microservices architecture where is component is completely independent [6, 21].

The overview of digital preservation ends by reviewing preservation policies. This brought to light the issue that archives may face issues around sustainability by not making optimal use of the resources they have available, which could be bad for both the environment and archives in under resourced countries that want to reduce the hardware and technical skills required [8, 9, 22]. These two issues of creating inclusive and environmentally safe archived will become bigger in the future and should be implemented in modern undertakings.

3 Modern Digital Archiving

3.1 Web Archiving

A vast amount of digital information exists on the internet, and all webpages of historical significance need to be saved [10]. Web archiving is the process of gathering up data that has been recorded on the World Wide Web and then collecting, cataloguing, storing, and preserving the data [10]. The preservation of web resources makes use of the same principles of the preservation of other digital resources and most of the work on web archives can be reapplied to other archives. The Internet Archive is a massive web archive which preserves petabytes of data while being managed by a tiny team of under 10 people [24]. A core feature of the Internet Archive is its Wayback Machine that provides an access tool with the ability to retrieve stored web pages through URL search [25]. New research emphasizes the importance of the Internet Archive because of its historical fact checking capabilities that it has provided while continually operating for more than 20 years [26]. The long life of the archive in part because of its simple design that has allowed it to be maintained with less hassles [24]. However, there are costs to its simplicity including the need for system administrators to perform many manual tasks, such as the copying files to new nodes when there is higher demand for it. It is estimated that the archive handles more than 10 TB of data per day. To keep handling the large daily load, research has been done suggesting that a reverse proxy cache could be used to improve the current architecture, which could be applied over a distributed system [24]. Rather than improve the Internet Archive an alternative to it is presented by the ArchiveWeb project that builds on it by trying to focus more on the consumer of the archive [25]. It attempts to differentiate itself from the Archive-It service, which the Internet Archive provides as a subscription to preserve an organisations digital content, by focusing on the end user and how they explore the archive [25]. It establishes a collaborative exploration method which allows search across many collections as well as the live web and provides collaboration capabilities to improve the archive through commenting and tagging items [25]. Instead of only focusing on archiving web pages research has also been done on using the data preserved. Web archive analytics makes use of publicly accessible web pages and their evolution for research purposes [27]. Analytics allows the ability to relate important and similar data together and a new system is presented that takes a subset of the Internet Archive data for processing, storing, and analysing [27]. The Webis research group has developed a custom analytics stack, for the data, which contains multiple layers including for data consumption, analytics, management, and acquisition, and then makes the output available for vendors to consume [27]. The intention is to find important insights from the large amount of information that no human could go through or discover by themselves.

3.2 Client-side Archiving

Most digital archives, particularly older ones, were created on some form of client-server basis where a prospective viewer would request content from a centralised server [10]. However, as everyday computers have increased in capacity, less need exists for centralised servers. Servers have problems around being single fault points, require stable network connections for clients to interact with them, and often have relatively high costs associated with them [22]. Thus, new work is being done on trying to distribute archives to be viewed on clients with less dependence on networks, however this has presented concerns [13, 15]. Nevertheless, web technologies continue to rapidly develop and achieve widespread adoption and Web Archiving has led to many ways of incorporating client-side technologies to view archived websites on a local environment with fewer server requests [16, 30, 31]. One attempt at creating better client-side archiving has been to develop a system that focuses on reproducing and storing the functionality of an archived web application instead of only preserving its data [30]. At its core the service decouples the web sites data from the functionalities it supports and moves the web application to a new simplified hosting environment that can be accessed and ported to other hosts across time and varying platforms [30]. Decoupling occurs by first extracting the current code and files of the deployed archive, then identifying the dependencies of the existing application, next redeploying it as a new application in a sandboxed environment, and finally creating a portable format to share the application and its hosting environment [30]. The ServiceWorker web API has been suggested as another way of maintaining online or linked functionality without the need for the original hosed environment

[31]. Specifically, this solution applies to the copying of web data called Composite Mementos, which are archived web pages that include both HTML pages and embedded resources [31]. A service worker is a part of a website that intercepts networks requests made by its users. By using a service worker, it is possible to create a client-side solution that does not require rewriting the URL references to resources of an archived web site, as is usually the case, by instead rerouting requests using a service worker. Thus, archived content can be presented rather than a dead link or incorrect content [31]. Service workers are advantageous as they can constantly be updated as required without any user interaction and have been shown to be more effective than having the server perform URL rewriting [31]. Another new development focuses on a search solution to address concerns around unreliable network access by running client-side [13]. The complexities of a server can be reduced with a simple browser-based search powered by JavaScript and HTML [13]. Browser based search can be achieved with a once-off pre-indexing of the collection. Tests performed illustrate that search can occur fully on the client within a reasonable amount of time for most queries even on relatively larger data sets of up to 32000 items. There is even the possibility of future improvement using JSON for the indexing [13]. Lastly, it is also possible for clients to perform more of the general processing. Ajax is a set of web development techniques that empower modern browsers to do many of the services previously required by a server [16]. Implementing Ajax practices can enhance user interaction, while reducing bandwidth usage and increasing scalability by making the client responsible for the bulk of processing [16]. This has been shown to be applicable for RSS feeds and even in-browser search by storing the indexed data set in static XML files.

3.3 Web and Client-side Archiving Discussion

The Internet Archive has been examined as an important web archive for its size, popularity, and how it shows what is possible to achieve with a small team and good design [24]. However, suggestions have been made on how to build on the Internet Archive's architecture by using a reverse proxy cache and the ArchiveWeb system tries to create a newer alternative to it by focusing on collaboration to improve the experience of the archive's end-user [25]. This shows that even well-established archives have the potential to be improved with newer options. Finally, an additional use case of the preserved web information was shown with the Webis analytics system that indicates that there are constantly new ways of making use of an archives stored information [27].

Four methods for improving client-side archiving have been introduced. Each focused on a particular aspect of the digital archive and on improving or converting the system for running locally on a client. The first method involved decoupling the components of a web archive to convert it to a format which can more easily be transported and self-contained [30]. The second method, specifically focused on creating a client-side way to address the presentation of embedded resources for archived web pages and not the full experience [31]. It presented a solution using the ServiceWorker web API. The third technique showed that it possible to locally replicate search on the client for smaller data sets using JavaScript [13]. Finally, the fourth method suggesting the client could perform much of the processing currently done by the server using the Ajax web techniques [16]. In total, other than the first method that is self-contained, these present complementary methods to make it more effective for archive users to be able to locally run and store the archives that they access. This shows that the preservation of online information can remain functional when in a local environment and that this aspect of archiving is likely to grow.

4 TOOLS AND CASE STUDIES

3.1 Tools

New digital archives can be created using previously developed and distributed tools. These tools make the process easier and attempt to address some of the standard problems of the archiving process [22]. Many such state-of-the-art tools are being consistently developed and are imperative in the current state of digital archiving. DSpace is a popular tool used in South Africa and provides a self-contained solution to implement all that is required of a typical archive including the collecting, managing, indexing, and distributing of digital items [4, 14]. Managing persistence is dependent on the policy of the implementing organisation, while DSpace itself stores all objects in a way which "digital archaeologists" could re-access if the current formats lose support [4]. On the other hand, Simple DL has been created as an alternative to address some of the shortcomings of tools, including DSpace, specifically for the creation of simpler archives which can also be accessed offline [11]. Simple DL is a software toolkit that focuses on prioritising simplicity, offline low-resource support, and the use of static data rather than having a broad scope [11]. It can facilitate preservation without active management and avoids the need for a database or software installation for usage. It stores unstructured data as flat files that are then referenced by metadata stored in XML documents [11]. It displays content through a local web page, which can also be hosted. The system has been applied to several in use archives, including that of the Digital Bleek and Lloyd archive. The Google Cultural Institute Platform is an archiving system which provides an alternative to other tools by reapplying the same processes and services needed for creating new archives [19]. It is unique in that all archives made through the tool are created and managed by Google without the same developer freedom as open-source options [19]. However, this caters to less technical archive administrators and allows the platform to easily scale. The main goal of the tool is contextualising assets to create storytelling that presents a compellingly experience of the information stored [19]. To achieve this, the platform is designed to go beyond preservation and convey the context and importance of a particular object [19]. The tool performs additional automated extraction of metadata to deduce understanding of the content stored and even executes computer vision to allow the analysis of images. These measures

allow associations to be gathered between items that enable connections to be found between assets instead of individual separation.

3.2 Case Studies

Many case studies exist of institutions creating their own digital archives. These cases all present certain themes and problems faced by its implementers and how they tried to create solutions to solve them [22]. These cases provide information on what is currently being performed in practice and what can be reapplied and standardised. The New Zealand Digital Library established its own architecture around the issue of having an archive made up of multiple distributed collections of data [28]. Its collection focused architecture works with multiple collections, with many document formats, that have different search and indexing engines. In addition to collections the project also worked on building a digital library that could simplify library administration and was developed to be capable of constantly evolving to support modern mediums as they come available [28]. Rather than focus on distributed collections the Bleek and Lloyd collection showed that an XML centric solution could replace the need for a database [15]. This was achieved by pre-processing data and then storing the whole collection into a single XML source document. Using XML over databases was able to increase the device support and removed the need for special software to run the archive [15]. A static XHTML powered website provided a scalable solution that could be set up to have specific resource bounds [12]. In this case, an AJAX-based local, browser search engine allowed querying that worked on all devices and could be distributed online and offline [12]. Similarly, the 500 Hundred Year Archive attempted to address concerns around stable internet access and reduce the required resources when retrieving content from its archive [14]. Additionally, it tried to overcome the common problem of tools being too complex for the needs of smaller archives. The end archive was able to reduce the need for network connection and large data storage, while achieving data preservation [14]. It achieved this by focusing on minimalism to create a higher chance of long-term support, offline distribution and client-side running [14]. In the opposite manner, the National Documentation Centre in Greece developed a large cloud first digital repository service that could archive the work of domestic cultural and science organisations and has been used to successfully create 28 repositories after rollout. The main objective of the project was to comply with the European Union's priority to provide open access to large amounts of intellectually important information while complying with the proper licensing [29]. This required addressing community needs to provide information as well as major international standards for managing cultural and scientific data of a high-quality data. The DSpace archiving tool was used as the data hosting and management system, due it being well established and compatible with the intended size [29]. The service was built upon a cloud infrastructure to support the desired large scale and due to the organisations previous experience building cloud-based administration systems [29]. While not being as large, the Bushman Online Dictionary (BOLD) is a system that also focused on local culture and was developed for archiving the cultural heritage of the Khoisan people through a visual dictionary of their |xam language [17]. It was able to achieve high user satisfaction for browsing its archive through enhanced contextualisation of content and a good viewing experience.

3.3 Tools and Case Studies Discussion

Three tools have been mentioned and each presents its own pros and cons. The Google Cultural Institute Platform offers a great way for non-technical archivists to have a fully managed system particularly when the archive is focused as much on engagement as in preservation [19]. This is a very different use case to that of Simple DL which focuses on preserving data in an offline and portable way which is not complicated and can be performed on a much smaller scale with less resources [11]. In contrast, DSpace provides a mixture of both approaches by providing developer control as well as a focus on large scale and complex archiving capabilities [4]. These tools together offer the variety to cater to almost any archivist and their priorities.

The case studies show that there are many different desires of archiving managing organisations and various methods of carrying it out. However, it appears that the local African archives are more likely to focus on utilising lower resources, being more portable, and client-side ready [12, 14, 17]. On the other hand, the National Documentation Centre in Greece created a cloud-based archive which made use of the DSpace tool, and the New Zealand Digital Library created a system which focused on distributed collections [28, 29]. Both of which show little constraint of resources or priority for offline support.

5 SUMMARY

This paper started by introducing archives and their current relevance in computer science. This was followed by the body of the review which firstly discussed the core archiving principle of preservation. Many new ways of improving the preservation process were shown from using a microservices architecture for modularity, to using blockchains to build trust, and even taking steps to increase environmentally sustainable. Secondly, web and client-side archiving were discussed. The web archiving section focused on the Internet Archive by highlighting improvements and alternatives for it, as well as how its data can be used for analysis. For client-side archiving, multiple methods of increasing its capability were suggested that included decoupling, service workers, browser-based search, and increased client processing. In addition, a few current archiving tools were considered with each having a different use case. Finally, the implementation of specific archives was examined for insight into the practical building of them. From all these sections, this literature review makes clear the importance of archiving. However, as can be seen, there is not yet a solution for ensuring the conservation of archives themselves. While aggregation occurs, there is no archive that provides an open access to other archives and versions their history overtime. Research has shown that archives themselves

experience existential risks to their persistence, particularly for those in African. There is the clear need to conserve information that is already archived, as there is never a guarantee that it is completely safe.

REFERENCES

[1] Hedstrom, M., 1997. Digital Preservation: A Time Bomb for Digital Libraries. *Computers and the Humanities*, 31(3), pp.189-202

[2] Jantz, R. and Giarlo, M., 2005. Digital Preservation: Architecture and Technology for Trusted Digital Repositories. *Microform & Imaging Review*, 34(3), pp.135 - 147.

[3] Pandey, R., 2003. Digital Library Architecture. *DRTC Workshop on Digital Libraries: Theory and Practice*, 25, p.2012. *Dlissu.pbworks.com*

[4] Smith, M., Barton, M., Branschofsky, M., McClellan, G., Walker, J., Bass, M., Stuve, D. and Tansley, R., 2003. DSpace: An Open Source Dynamic Digital Repository. *D-Lib Magazine*, 9(1).

[5] Carbajal, I. and Caswell, M., 2021. Critical Digital Archives: A Review from Archival Studies. *The American Historical Review*, 126(3), pp.1102-1120.

[6] Mayo, C., Jazairi, A., Walker, P. and Gaudreau, L., 2019. BC Digitized Collections: Towards a Microservices-based Solution to an Intractable Repository Problem. *Code4Lib Journal*, 6(44). *Journal.code4lib.org*

[7] Lin, W., Zuo, J., Su, S. and Chen, C., 2019. A double-blockchains based Digital Archives Management Framework and Implementation. In *2019 IEEE 14th International Symposium on Autonomous Decentralized System (ISADS)*. Utrecht, Netherlands: IEE, pp. 1-6.

[8] Pendergrass, K., Sampson, W., Walsh, T. and Alagna, L., 2019. Toward Environmentally Sustainable Digital Preservation. *The American Archivist*, 82(1), pp.165-206.

[9] Anyaoku, E., Echedom, A. and Baro, E., 2019. Digital preservation practices in university libraries: An investigation of institutional repositories in Africa. *Digital Library Perspectives*, 35(1), pp.41-64.

[10] Niu, J., 2012. An Overview of Web Archiving. *D-Lib Magazine*, 18(3/4).

[11] Suleman, H., 2021. Simple DL: A Toolkit to Create Simple Digital Libraries. *Lecture Notes in Computer Science*, pp.325-333.

[12] Suleman, H., 2011. An African Perspective on Digital Preservation. *Multimedia Information Extraction and Digital Heritage Preservation*, pp.295-306.

[13] Suleman, H., 2019. Investigating the Effectiveness of Client-Side Search/Browse Without a Network Connection. *Digital Libraries at the Crossroads of Digital Information for the Future*, pp.227-238.

[14] Suleman, H., 2019. Reflections on Design Principles for a Digital Repository in a Low Resource Environment - UCT Computer Science Research Document Archive. *Pubs.cs.uct.ac.za*.

[15] Suleman, H., 2007. Digital Libraries Without Databases: The Bleek and Lloyd Collection. In *Proceedings of Research and Advanced Technology for Digital Libraries*. ECDL 2007. Lecture Notes in Computer Science, 4675, pp.392-403. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-74851-9_33

[16] Suleman, H., 2007. In-Browser Digital Library Services. In *Proceedings f Research and Advanced Technology for Digital Libraries*. ECDL 2007. Lecture Notes in Computer Science, 4675, pp. 462-465. Springer, Berlin. https://doi.org/10.1007/978-3-540-74851-9_43

[17] Suleman, H. 2007. In-Browser Digital Library Services. In *Proceedings of Research and Advanced Technology for Digital Libraries*. ECDL 2007. Lecture Notes in Computer Science, 4675. Springer, Berlin. https://doi.org/10.1007/978-3-540-74851-9_43

[18] Lighton Phiri and Hussein Suleman. 2013. Flexible design for simple digital library tools and services. In *Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference* (*SAICSIT '13*). Association for Computing Machinery, New York, NY, USA, 160–169. https://doi.org/10.1145/2513456.2513485

[19] Seales, W., Crossan, S., Yoshitake, M. and Girgin, S., 2013. Google Cultural Institute. *Choice Reviews Online*, 50(07), pp.50-3627-50-3627. American Library Association. DOI: 10.5860/choice.50-3627

[20] Lagoze, C., Payette, S., Shin, E. and Wilper, C., 2005. Fedora: an architecture for complex objects and their relationships. *International Journal on Digital Libraries*, 6(2), pp.124-138.

[21] Amaral, M., Polo, J., Carrera, D., Mohomed, I., Unuvar, M. and Steinder, M., 2015. Performance Evaluation of Microservices Architectures Using Containers. In *2015 IEEE 14th International Symposium on Network Computing and Applications*. pp. 27-34. IEEE. http://10.1109/NCA.2015.49

[22] Yadav, D., 2016. OPPORTUNITIES AND CHALLENGES IN CREATING DIGITAL ARCHIVE AND PRESERVATION: AN OVERVIEW. *International Journal of Digital Library Services*, 6(2), pp.63-73.

[23] Bralić, V., Stančić, H. and Stengård, M., 2020. A blockchain approach to digital archiving: digital signature certification chain preservation. *Records Management Journal*, 30(3), pp.345-362.

[24] Elliot Jaffe and Scott Kirkpatrick. 2009. Architecture of the internet archive. In *Proceedings of SYSTOR 2009: The Israeli Experimental Systems Conference* (*SYSTOR '09*). Association for Computing Machinery, New York, NY, USA, Article 11, 1–10. https://doi-org.ezproxy.uct.ac.za/10.1145/1534530.1534545

[25] Fernando, Z., Marenzi, I. and Nejdl, W., 2017. ArchiveWeb: collaboratively extending and exploring web archive collections—How would you like to work with your collections?. *International Journal on Digital Libraries*, 19(1), pp.39-55.

[26] Harris, K., Beis, C. and Shreffler, S., 2022. The Internet Archive has been Fighting for 25 Years to Keep What's on the Web From Disappearing and You Can Help. *Ecommons.udayton.edu*.

[27] Völske, M. et al., 2021. Web Archive Analytics. *INFORMATIK 2020*, pp.61-72.

[28] Rodger, M., Witten, I. and Boddie, S., 1998. A distributed digital library architecture incorporating different index styles. In *IEEE International Forum on Research and Technology Advances in Digital Libraries-ADL'98*. IEEE, pp. 36-45. DOI: 10.1109/ADL

[29] Bartzi, K., Vasilogamvrakis, N., Lagoudi, E., Hardouveli, D. and Sachini, E., 2019. The Digital Repository Service of the National Documentation Centre in Greece: a model for Digital Humanities data management and representation. In AIUCD 2019-Book of Abstracts, pp. 65-71.

[30] Xu, W., Esteva, M., Beck, D. and Hsieh, Y., 2017. A Portable Strategy for Preserving Web Applications Functionality. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. IEEE, pp. 1-4.

[31] Sawood Alam, Mat Kelly, Michele C. Weigle, and Michael L. Nelson. 2017. *Client-side reconstruction of composite mementos using serviceworker*. In *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries (JCDL '17)*. IEEE Press, 237–240.