Seeking Sustainability

Developing a Modern Distributed Digital Preservation System

|  |  |  |
| --- | --- | --- |
| **Nathan Tallman** |  | **Hannah Wang** |
| *Penn State University**United States**ntt7@psu.edu**0000-0002-5308-4100* |  | *Educopia Institute**United States**hannah.wang@educopia.org**0000-0002-6676-1254* |

**Abstract – As modern commercial developments in storage infrastructure mature and become increasingly available through popular open-source projects, there are important opportunities for digital preservation communities to leverage the increased efficiency and flexibility that these technologies offer. Not only do these developments offer a way to “modernize” the digital preservation technology stack and make it more efficient, but they also may allow digital preservation communities to seek increased sustainability per the triple bottom line: reduce the costs of operations, reduce required labor to maintain, and reduce the environmental impact. The twin values of affordability and sustainability are core to the mission of digital preservation, and the MetaArchive Cooperative is pursuing the research and development of a modern distributed digital preservation system to better practice these values.**

**Since its inception in 2004, MetaArchive has used Stanford University’s Lots of Copies Keep Stuff Safe (LOCKSS) software for its core infrastructure to achieve distributed digital preservation. While LOCKSS has been reliable software for many years, recent evaluations by MetaArchive and a desire to make its practices more sustainable have led to MetaArchive partnering with Keeper Technology to explore software-defined storage and serverless computing technologies for digital preservation. The results of this multi-phase project will be shared with the digital preservation community, with the hopes that it will encourage other digital preservation technological developments with a similar sustainability mindset.**

**Keywords – distributed digital preservation, sustainability, software defined storage, cloud infrastructure, inclusion.**

**Conference Topics – Innovation; Environment.**

# Introduction

The MetaArchive Cooperative is an international distributed digital preservation (DDP) network based in the United States and hosted by the Educopia Institute [1]. MetaArchive started in 2004 as a project with funding from the National Digital Information Infrastructure and Preservation Program (NDIIP) initially involving six southern United States universities and the Library of Congress. After initial success, MetaArchive transitioned from a project to an independent network open to any cultural heritage organization in 2007 [2].

Since its inception, MetaArchive has relied upon Stanford University’s Lots of Copies Keep Stuff Safe (LOCKSS) software [3] for its core infrastructure to achieve distributed digital preservation in a Private LOCKSS Network (PLN) [4]. While LOCKSS has been reliable software for many years, MetaArchive has experienced challenges in meeting network and member needs. In 2020, MetaArchive conducted an evaluation of LOCKSS to identify strengths, weakness and challenges, and opportunities [5]. Findings verified that LOCKSS (1.x) was a viable, demonstrated cost-effective solution for bit-level preservation that is still used within the community. However, LOCKSS’ high maintenance costs (direct and indirect) in MetaArchive’s implementation, issues with scaling, and difficulties in content management for members inhibit the growth of the MetaArchive Cooperative. Among other recommendations, two parallel paths were established: continue with LOCKSS & explore potential benefits of the LOCKSS Architected as Web Services project [6] and explore alternative options for DDP that leverage modern technological innovations.

In January 2022, MetaArchive began a multi-phase research and development project to create a modern DDP system. Along with leveraging technical infrastructure advancements in the commercial sector and simplifying digital preservation systems, MetaArchive’s goal is to make digital preservation more sustainable per the triple bottom line [7]: reduce the costs of operations, reduce required labor to maintain, and reduce the environmental impact.

# Triple Bottom Line Sustainability and the UN Sustainable Development Goals

As a member of the Digital Preservation Services Collaborative, MetaArchive’s mission to preserve cultural heritage and research is rooted in a set of shared core values [8]. The research and development described in this paper is being undertaken in pursuit of greater adherence to these values, particularly technological diversity, inclusion, and, most immediately, affordability and sustainability. While sustainability may be a buzzword, it’s a critical concern for digital preservation. If sustainability is ignored, only the most well-resourced organizations will be able to put forward the money, labor, and infrastructure to steward digital content into the future. Not only are individual resources at risk of being lost, but if only the collections of the largest organizations persist, the consequences to representation in the academic, scholarly, and cultural record are severe.

## Environmental Sustainability

Ben Goldman points out that data centers have a significant environmental impact consuming large amounts of energy and water in “It’s Not Easy Being Green(e)” [9]. The 2019 NDSA Storage Infrastructure Survey, with only 85 responses to the question, indicates that cultural heritage organizations are storing at least 51 petabytes and as much as 114 petabytes of unique digital content [10]; since there are hundreds or thousands of cultural heritage organizations worldwide this number is likely a substantial underestimate. As Keith Pendergrass and his coauthors point out in “Toward Environmentally Sustainable Digital Preservation,” even conservative estimates indicate that the world’s cultural heritage organizations consume over five exabytes of storage, without even accounting for the fact that most organizations replicate content multiple times [11]. Furthermore, if digital preservation practices and policies do not include reappraisal and deaccessioning of preserved content, digital storage consumption will only grow unchecked in perpetuity. Without even considering the practices that produce physical and virtual servers and other infrastructure components, digital preservation itself has a significant environmental impact.

## Economic Sustainability

Despite overall price trends in commodity storage, preservation storage is expensive. The Digital Preservation Storage Criteria project’s 61 criteria are an indication of why this is the case [12]. Cultural heritage organizations need to be strategic about digital infrastructure and not operate in old paradigms. MetaArchive members are small and large, DDP technology needs to be as low-cost as possible to be accessible to all.

## Labor Sustainability

Many digital preservation systems continue to perform preservation actions in the application level of the technology stack. The application layer is often the most time-consuming and expensive level to maintain. Moving preservation actions into lower levels of the infrastructure, such as the underlying storage system, can reduce the amount of labor. Additionally, by using commercial-sector infrastructure, it’s easier to recruit skilled professionals; conversely, it can be challenging to recruit developers to maintain bespoke cultural heritage applications.

## UN Sustainable Development Goals

The UN Sustainable Development Goals are a framework for improving life for all living creatures. Goal 9 specifically calls for building resilient infrastructure, promoting inclusive and sustainable industrialization, and fostering innovation [13]. While most efforts in Goal 9 are focused on manufacturing and industry, digital infrastructure in cultural heritage organizations needs to embrace these principles if our collections are to persist for future users. Innovating our technology to adopt strategies used by the world’s leading corporations will make our infrastructure less expensive, more resilient, and available to more organizations.

# A Framework for Modern Distributed Digital Preservation Systems

MetaArchive anticipates implementing technologies created in the past two decades when building our next-generation DDP system and re-thinking how digital preservation occurs from the ground up. This paper will focus on only two of those technologies: software-defined storage and serverless computing. For a more detailed look at modern technology for digital preservation, see Nathan Tallman’s 2021 article in Information Technology and Libraries [14].

## Software-Defined Storage

Software-defined storage is an application that operates in-between physical disks and the operating system. It replaces traditional storage management with a dynamic, flexible, and resilient system that can help manage basic preservation activities. Instead of using a file system managed by an operating system or a storage appliance that presents a file system to the operating system, software-defined storage lets you build your own storage appliance or storage network that supports multiple protocols for access including file, object, and block storage [15]. Popular open-source software-defined storage solutions include Ceph [16] and Gluster [17] and there are several commercial offerings as well. A software-defined storage network can be configured within a single data center or multiple data centers anywhere in the world.

Software-defined storage has several features that support digital preservation. First, because you can build your own storage appliance/network, you can achieve hardware diversity to mitigate risk related to single points of failure. Combining different hard drive manufacturers, batches, and hard drive technology, as well as the server components themselves provides a broad spectrum of hardware diversity. Second, like RAID, software-defined storage builds in erasure encoding for parity. This helps to protect against loss when a hard drive or even an entire node or cluster in a storage network fails. The level of protection can be configured as needed. Third, software-defined storage offers many options for achieving geographic-redundant replications. Software-defined object stores can be set up in different availability zones with bucket-level policies to ensure replication to as many locations as needed and supported by the storage network. If you have ever used AWS, Azure, or Google cloud storage, you have almost certainly been using software-defined storage.

When using traditional file storage (SMB/CIFS) in a software-defined storage network, it’s also possible to leverage modern file systems such as OpenZFS [18] and BTRFS [19] that have built-in features to ensure integrity. This is achieved by tracking checksums for the underlying blocks in file storage and leveraging parity to repair blocks found to have loss. Managing fixity in the storage-level simplifies higher-level applications and significantly reduces the environmental impact when compared to ongoing file-level fixity checks. For a fuller explanation of ZFS and digital preservation, see Alex Garnett, Mike Winter, and Justin Simpson’s 2018 iPRES paper [20].

## Serverless Computing

A Serverless computing, sometimes also called function-as-a-service, allows you to offload certain tasks. Instead of running commands on the same server as the repository, you call a function-as-a-service that executes the action on a different server. It’s the ultimate microservice that can be finely tuned to consume only the required resources; instead of having a monolithic stack that is highly resourced for peak activity, the repository stack can be optimized for repository management while the serverless computing platform is optimized for high performance and throughput, often managed by another entity, and paid for based on usage. Serverless computing can be invoked by an application to perform tasks like file format characterization, format migrations, replications, and more. Depending on the platform and your needs, a serverless function may use its own container or simply run on its host platform. Using a serverless computing for basic preservation activities simplifies the digital preservation stack that needs to be maintained, functions can be called as needed. If using a commercial FaaS platform, this reduces the need to maximally resource servers and maintain more infrastructure, which in turn simplifies labor requirements.

# Research and Development

MetaArchive is partnering with Keeper Technology (KeeperTech), a Virginia-based storage and data solutions company with deep expertise in software-defined storage [21], in carrying out this work. The work is structured in three phases, with off ramps at the end of each stage for both parties. Phase 1 is to collaboratively develop a high-level design document and feature requirements for a modern DDP system. Before Phase 1 commencement in January 2022, MetaArchive members used the Digital Preservation Storage Criteria to prioritize, justify, and refine initial requirements [12]. Additionally, OSSArcFlow diagrams were developed for select basic network tasks [22]. KeeperTech reviewed these inputs and developed several questions to explore with MetaArchive members. KeeperTech distributed a questionnaire to all MetaArchive members and a more detailed set of questions that were explored in focus groups. Based on responses to these questions, the provided inputs, and previous conversations between MetaArchive, KeeperTech will craft a white paper or high-level design document. This document will be shared with the digital preservation community.

At the conclusion of Phase 1, if mutually agreeable to both parties, MetaArchive and KeeperTech will begin to implement the high-level design in a working prototype in Phase 2. The working prototype will include the development of architectures, infrastructure, and applications that will be capable of demonstrating proof-of-concept. Phase 2 will also include creating testing requirements to ensure success. Again, outputs from this phase will be shared with the digital preservation community as open source.

Phase 3 will explore a fully operational DDP network. Like Phase 2, it will only be pursued if both parties agree, though either may choose to work independently. Phase 3 tasks will include blueprints for central and network operations, deployment guides, documentation, software packaging, and other items necessary to move into production. Although not yet determined, MetaArchive may choose to contract with KeeperTech for ongoing support or even to have them host centralized components of the system.

# Conclusion

While the concept of DDP emerged in the 1990s, digital preservation communities don’t have to continue to rely on 1990s technology, nor should they. As Trevor Owens, quoting Martha Anderson in *The Theory and Craft and Digital Preservation*, says, digital preservation is a relay race [23]. It is a chain of hand-offs between mediums, systems, and stewards. Preservationists should expect to update their technology as time goes on, the same way they forward-migrate digital content itself. Doing so ensures that communities are using the most sustainable, efficient, and affordable means to achieve DDP goals. As the MetaArchive Cooperative aims to put digital preservation in reach for organizations of any size, it is vital to ensure it is fiduciarily responsible in meeting that goal. Ultimately, MetaArchive hopes that the results of this project will not only improve services for its own members, but also will encourage other digital preservation communities to adopt and pursue technological developments with a similar sustainability mindset.

# REFERENCES

1. Educopia Institute, "Educopia Institute | Empowering Collaborative Communities," 2022. [Online]. Available: <https://educopia.org>. [Accessed 21 January 2022].
2. K. Skinner and M. Halbert, "The MetaArchive Cooperative: A Collaborative Approach to Distributed Digital Preservation," *Library Trends*, vol. 57, no. 3, pp. 371-392, 2009. [Online] Available: <https://doi.org/10.1353/lib.0.0042>. [Accessed 28 January 2022].
3. Stanford University, "LOCKSS," 2021. [Online]. Available: <https://www.lockss.org>. [Accessed 21 January 2022].
4. V. Reich and D. Rosenthal, "Distributed Digital Preservation: Private LOCKSS Networks as Business, Social, and Technical Frameworks," *Library Trends*, vol. 57, no. 3, pp. 461-475, 2009. [Online] Available: <https://doi.org/10.1353/lib.0.0047>. [Accessed 28 January 2022].
5. N. Tallman, Z. Vowell, B. Robbins and M. Schultz, "MetaArchive LOCKSS Evaluation," 2020. [Online] Available: <https://doi.org/10.26207/3acj-bs29>. [Accessed 4 February 2022].
6. Stanford University, "Frequently Asked Questions | LOCKSS," 2021. [Online]. Available: <https://www.lockss.org/about/frequently-asked-questions#laaws>. [Accessed 21 January 2022].
7. K. Miller, "The Triple Bottom Line: What It Is & Why It's Important," 08 December 2020. [Online]. Available: <https://online.hbs.edu/blog/post/what-is-the-triple-bottom-line>. [Accessed 28 January 2022].
8. Digital Preservation Services Collaborative, "Digital Preservation Declaration of Shared Values," 12 April 2018. [Online]. Available: <https://dpscollaborative.org/shared-values_en.html>. [Accessed 18 February 2022].
9. B. Goldman, "It's Not Easy Being Green(e): Digital Preservation in the Age of Climate Change," in *Archival Values: Essays in Honor of Mark A. Greene*, Chicago, American Library Association, 2018, pp. 274-295. [Online]. Available: <https://scholarsphere.psu.edu/resources/381e68bf-c199-4786-ae61-671aede4e041>. [Accessed 28 January 2022].
10. 2019 NDSA Storage Infrastructure Survey Working Group, "2019 Storage Infrastructure Survey Report," 24 February 2020. [Online]. Available: <https://doi.org/10.17605/OSF.IO/UWSG7>. [Accessed 28 January 2022].
11. K. Pendergrass, W. Sampson, T. Walsh, L. Alagna, “Toward Environmentally Sustainable Digital Preservation,” *The American Archivist*, vol. 82, no. 1, 2019. [Online] Available: <https://doi.org/10.17723/0360-9081-82.1.165>. [Accessed 18 August 2022].
12. S. Schaefer, N. McGovern, A. Goethals, E. Zierau and G. Truman, "Digital Preservation Storage Criteria," 10 December 2018. [Online]. Available: <https://doi.org/10.17605/OSF.IO/SJC6U>. [Accessed 28 January 2022].
13. United Nations, Department of Economic and Social Affairs, "Goal 9 | Department of Economic and Social Affairs," 2021. [Online]. Available: <https://sdgs.un.org/goals/goal9>. [Accessed 28 January 2022].
14. N. Tallman, "A 21st Century Technical Infrastructure for Digital Preservation," *Information Technology and Libraries*, vol. 40, no. 4, 2021. [Online] Available: <https://doi.org/10.6017/ital.v40i4.13355>. [Accessed 4 February 2022].
15. M. Carlson, A. Yoder, L. Schoeb, D. Deel, C. Pratt, C. Lionetti and D. Voigt, "Software Defined Storage," January 2015. [Online]. Available: <https://www.snia.org/sites/default/files/SNIA_Software_Defined_Storage_%20White_Paper_v1.pdf>. [Accessed 4 February 2022].
16. "Ceph.io," 2022. [Online]. Available: <https://ceph.io/>. [Accessed 2022 February 2022].
17. Red Hat, Inc., "Gluster," 2019. [Online]. Available: <https://www.gluster.org/>. [Accessed 2022 February 2022].
18. "OpenZFS," 2021. [Online]. Available: <https://openzfs.org/wiki/Main_Page>. [Accessed 4 February 2022].
19. "btrfs Wiki," 4 February 2022. [Online]. Available: <https://btrfs.wiki.kernel.org/index.php/Main_Page>. [Accessed 4 February 2022].
20. A. Garnett, M. Winter and J. Simpson, "Checksums on Modern Filesystems, or: On the Virtuous Consumption of CPU Cycles," in *iPres 2018 Conference [Proceedings],* Boston, 2018. [Online] Available: <https://doi.org/10.17605/OSF.IO/Y4Z3E>. [Accessed 4 February 2022].
21. Keeper Technology, LLC., "KeeperTech," 2022. [Online]. Available: <https://www.keepertech.com/>. [Accessed 4 February 2022].
22. K. Skinner, S. Meister and C. Lee, "OSSArcFlow," 2020. [Online]. Available: <https://educopia.org/ossarcflow/>. [Accessed 4 February 2022].
23. T. Owens, *The Theory and Craft of Digital Preservation*, Baltimore, MD: Johns Hopkins University Press, 2018.