TOWARDS LEVELS OF DIGITAL PRESERVATION AS A SERVICE

Building an extensible, affordable digital preservation service

Jefferson Bailey
Internet Archive
USA
jefferson@archive.org
https://orcid.org/0000-0002-0830-6325

Peggy Lee
Internet Archive
USA
peggy@archive.org

Abstract – Since 1996, the Internet Archive (IA) has provided storage, preservation, and access infrastructure and services to over 1,000 cultural heritage organizations around the world. It has also provided customized digital preservation services on a contractual basis to a handful of large institutions. In 2020, IA began building a more generalized digital preservation service in response to the needs of a broader range of institutions and to leverage IA's self-owned data centers, non-profit cloud services, and demonstrated expertise in both small and petabyte-scale digital stewardship. This system is being developed in direct dialogue with 30+ organizations, including universities, public libraries, arts organizations, and cultural heritage organizations, over the course of the 2021 - 2022 year. This paper shares key takeaways from the information collected from this pre-pilot phase and serves as a lean landscape review of the current gaps within the digital preservation landscape, particularly as they relate to the distinct needs and goals of nonprofits, libraries, and cultural heritage organizations, that this service aims to address.

Keywords – digital preservation, product development, archiving, open infrastructure, sustainability

Conference Topics – Scanning the New Development; Building the Capacity.

I. INTRODUCTION

I. Since its inception as a non-profit digital library in 1996, Internet Archive has focused on ensuring the continued availability and accessibility of human knowledge by creating a digital library to permanently store digital content. The Internet Archive is the world's largest public web archive, with hundreds of petabytes of data stored within its independently owned and operated, not-for-profit data centers. Currently 1000+ partners, including national libraries, universities, and cultural heritage organizations, collaborate with the Internet Archive on various archiving, access, open source technology development, and digital library projects with the shared mission of ensuring perpetual preservation and access to diverse, cultural, and historically-relevant digital collections from around the world.

Internet Archive is building a new general purpose digital preservation services to complement and extend its existing suite of free and paid, cost-recovery services for digitization, web archiving, general data storage, and web and access services. The new Digital Preservation Service suite will be built on existing Internet Archive infrastructure and open-source software and incorporate the feedback of pilot partners and peer stakeholders as it progresses through the product life cycle. Dozens of institutions are participating in a pilot project that includes iterative development cycles informed by pilot partner usage and by broader input from the community of users of IA services. One of the goals of the services is to replicate the simplicity, extensibility, and available to organizations of any type and size that led to the creation of the original NDSA Levels of Digital Preservation. [1][2] One of the co-creators of the Levels of Preservation is the Director of the service and the service aims to take the guidance and principles of the Levels of Preservation and translate them into a best-of-class service for the cultural heritage, non-profit, and social impact sectors. Engaging users at all stages of development will help ensure the service's fidelity to the goals and needs of mission-aligned organizations and, in turn, further the capacity of these non-profit organizations to preserve and protect valuable materials for the public good.

II. CENTRAL FEATURES OF THE SERVICE

At its core, the Digital Preservation Service Suite will allow users to deposit any digital content of any

17th International Conference on Digital Preservation
iPRES 2021, Beijing,China.

DOI: 10.1145/nnnnnnn.nnnnnnn

size, specify what geographical location their data will be stored (across multiple locations in 3 different countries), set how many copies of the data will be replicated and their distribution across various data centers in various regions, select if they want their collections stored in different technical architectures, and select the frequency of audit and repair operations such as fixity checking and digital object correction. In line with IA's user-centric design philosophy, a key success metric for the digital preservation service suite is its ability to accommodate the diverse preservation goals of organizations of various sizes, locations, and expertise in digital preservation management. The suite is intended to be responsive and customizable to various use cases, with partners able to select custom numbers of copies, specify desired storage locations, and schedule multiple fixity occurrences with service levels from basic storage services to highly-replicated, full-features distributed digital preservation services.

The service has an interactive dashboard to view the real-time status of all preserved data, including storage location, fixity reporting, manifests, analytics, and other transactional metrics, so that partners will be able to actively monitor their data and make timely decisions about its organization or the what various service features should be implemented for specific collections within their overall account. Reports and metadata will also be available through APIs, with additional plans for integration with peer services, repositories, and preservation systems in progress.

Mindful of the resource constraints of nonprofits, the digital preservation service suite will also benefit from the Internet Archive's efficiencies of scale to offer storage and preservation solutions at minimal cost so that mission-aligned organizations, particularly those who have heretofore been unable to participate in digital preservation practice, no longer have cost or technology as a barrier for entry, a common finding in IA's regular "State of WARC" survey amongst IA partner organizations. [3]

III. Stakeholder Needs Assessment

From continued conversations with stakeholders, the Internet Archive product team has learned a great deal about the current digital preservation service gaps experienced by libraries, universities, and non-profit organizations. Many organizations have also used the Levels of Preservation guidance as an assessment tool for analyzing or planning for their own digital preservation activities. [4] An early, welcome surprise to the team's call for participants was the high amount of enthusiasm and demand for this service, suggesting that current service providers are not meeting the variety of needs of many heritage organizations, especially smaller or more unique libraries and archives. Our intended 3 group calls more than doubled to 8 to accommodate the growing number of organizations wanting to participate in these initial conversations. Several organizations were keen to incorporate the features of the service in their long term organizational preservation planning and plan to develop their digital preservation strategy alongside our service's product development. Additionally, several organizations voiced dissatisfaction with current commercial solutions (detailed further below). The data amassed from our needs assessment form and early conversations guide our belief that the digital preservation service suite successfully taps into a high need area for mission-aligned, memory organizations. In all, over 50+ organizations engaged with us, including college or university libraries, public libraries, religious, specialty, or research libraries and archives, arts and museum institutions, multiple consortia entities, and international organizations.

As potential users, many organizations had distinct ideas for how they would like to use the digital preservation service and shared how their current process or solutions are in need of improvement. These findings include:

Priority Features: We asked each potential pilot partner to indicate which feature they were most interested in for their organization. Features deemed most desirable were 1) geolocation options that would allow partners to select between 3 countries in 2 continents for storage, 2) dashboard tools that provide clear data monitoring, provide simple visual representation of the content, status, and activity related to preserved data organized in various collections, and ready access to audit and repair reports, and 3) replication functions that allow partners a flexible means to manage content replication according to various criteria related to types of digital objects and to initiate more or less copies for different subsets of their content. As several organizations desired to store audio-video files within the service, preview and appraisal tools were of higher need than initially anticipated, and highlighted an aspect of digital collections management that suggest a need for temporary basic storage that can be easily connected with preservation storage with more flexibility for administering collection status.

Nascent Digital Preservation Practice: Roughly a third of the pilot participants indicated a need for more support for building their digital preservation strategy. In addition, these organizations lacked external digital preservation tools or service providers. These organizations were in the early stages of developing a digital preservation strategy, had been attempting to DIY solutions at a level that were deemed unsatisfactory, or, in one case, had their existing digital preservation service decommissioned within the last few years. Amongst this particular group of organizations, there was a large spectrum of technical proficiency; some had attempted to build a patchwork of services in house while most had generally kept digital records on a server without a comprehensive digital preservation plan. All organizations within this category came to the product team hoping for ready-made solutions for active monitoring of their data, for

replicating digital materials as needed, and ensuring perpetual access.

Large-Scale Grants and Acquisitions: Multiple organizations viewed the digital preservation service as a solution for an anticipated influx of digital materials in conjunction with recent grants for digitization efforts or for new acquisitions to their collection. One potential partner reported that they will be acquiring an additional 100 TB of data from grant-funded projects within the year and another institution shared that they were in need of 500 TB of storage for a new film and media archive. The large scale of these new acquisitions warranted a digital preservation solution that can accommodate the size of these collections and provide the adequate tools for organizational oversight, including large-scale fixity checks and comprehensible reports.

Consortial Considerations: Large consortial organizations, many of which engaged in early conversations with the product team, described the necessity of nimble solutions that can sufficiently address the needs of their various participating member organizations. Responding to the divergent needs of many affiliated organizations tests the strength of the service's customizable controls and options. Such organizations present valuable operational opportunities to apply both consortial and individual organizational digital preservation strategies to diverse use cases.

Dissatisfaction/Cost Constraints with Commercial Services: Many of our potential partners shared difficulties and limitations with current service providers. The overriding difficulties related primarily to 1) unintuitive interfaces not mapping to desired workflows, including a lack of options with fixity checks, 2) high expense associated with commercial services, with maintenance worries if organizations experience lapses in funding, 3) cost constraints associated with commercial services relegating such options to one of many patchwork services that do not add to a comprehensive, end-to-end solution for organizations. Most organizations within our pilot partner group are not able to afford more holistic service offerings which results in considerable operational investment from staff and additional difficulties when managing expanding collections.

IV. NEXT STEPS

With the wealth of information provided by the initial cohort group, the product team plans to incorporate the prioritized features into the initial product design of the service. Mindful of the expressed need for a comprehensive service with clear and accessible controls, the team will continually validate feature development and service offerings in direct collaboration and dialogue with the pilot cohort.

The pilot is currently underway and intended to run throughout 2021 with the goal to provide early, no-cost access to the service's core features for 30+ selected organizational partners in exchange for their use, input, and feedback on ongoing technical development. Pilot partners will deposit multiple terabytes of data at no charge to them and will be guaranteed perpetual preservation and access to the data and ongoing access to the service in an ongoing manner. In return, pilot partners will provide their feedback to the product team in quarterly check-in calls, meetings, survey forms, and other communication instruments.

The iterative, co-creation design principles of the service's development lifecycle bolsters the Internet Archive's capacity to build relevant, accessible, and sustainable preservation solutions for mission-aligned organizations. The success of such efforts will be measured through stakeholder feedback sessions, user testing, and, eventually, in the uptake of the service's offerings within the digital preservation ecosystem. In line with the open source ethos of the Internet Archive, findings and lessons learned from the development and launch of the digital preservation service will be shared with the larger research community to further the field. In addition, the Internet Archive will continue to pursue collaborations, integrations, and/or future testing opportunities with diverse, mission-aligned organizations to ensure services developed are as inclusive as possible of various cultural and technical contexts

REFERENCES

[1] Bailey, et al, "The NDSA Levels of Digital Preservation: An Explanation and Uses," Proceedings of the Archiving (IS&T) Conference, 2013, https://web.archive.org/web/20190207222334/http://www.digitalpreservation.gov/documents/NDSA_Levels_Archiving_2013.pdf.

[2] Levels of Digital Preservation, https://ndsa.org/publications/levels-of-digital-preservation/.

[3] Archive-It, "State of the WARC," 2020, https://archive-it.org/blog/category/state-of-the-warc-reports/.

[4] Baucom, et al, "Using the Levels of Digital Preservation as an Assessment Tool," NDSA, 2019, https://osf.io/m6j4q/.