Submitted on: 08.06.2016

Disasters at any scale: The MetaArchive Cooperative's community-based approach to risk mitigation and disaster preparation in the 21st century

Sam Meister, Educopia Institute, Columbus, Ohio, USA
sam@educopia.org

Gabrielle V. Michalek, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA
gabrielle@cmu.edu

Abstract:

Integration and automation between multiple acquisition, analysis, and collection management software systems is key to the sustainability of digital preservation within and across institutional boundaries amongst cultural heritage organizations. The MetaArchive Cooperative distributed digital preservation network has continually sought to simplify integration with multiple systems and has engaged in research and development activities to achieve these goals. To illustrate these activities we will provide a case study of one member's (Carnegie Mellon University) experience integrating digital preservation workflows between MetaArchive and a vendor-based digital repository platform. Additionally, we will describe current efforts to evolve our network infrastructure to provide increased opportunities for smaller, under-resourced cultural heritage institutions to participate in the MetaArchive Cooperative, and therefore protect their digital collections against disaster-related risk and loss.

Keywords: disaster recovery, digital preservation, cultural heritage, system integration

Introduction

Disasters in the 21st century take many different shapes and sizes. In addition to high profile natural disasters such as floods, fires, and earthquakes, there are numerous small, discrete, and seemingly invisible disaster events that are increasingly likely to occur. These tiny disasters are the types of risks that occur within the realm of digital cultural heritage, at the level of the bits and bytes that constitute digital information. Mitigating against these risks has become a part of standard operating procedures for disaster preparation and recovery in many cultural heritage institutions. As more cultural heritage institutions increase their digital holdings, whether

through digitization or acquisition of born-digital materials, the need for digital preservation solutions that can protect against the range of risks that digital information are susceptible to is paramount.

The MetaArchive Cooperative was founded on the premise that by actively engaging in community-based collaboration, cultural heritage institutions can best position themselves to overcome the challenges of preserving digital information over the long-term. Utilizing a network-based distributed digital preservation approach that replicates copies of digital collections geographically, MetaArchive Cooperative members collaborate to preserve each other's digital content, achieving fundamental bit-level digital preservation and protecting against a range of natural and man-made risks. Since its founding in 2004, the MetaArchive Cooperative has encountered multiple risk scenarios at both the institutional and network-wide level, and has successfully recovered digital content without loss.

MetaArchive Cooperative Overview

As a Private LOCKSS Network (PLN)[1], the MetaArchive Cooperative utilizes a specialized application of the LOCKSS protocol and function, using the same software as the public Global LOCKSS network. While the Global LOCKSS network focuses on the preservation of electronic journals, MetaArchive is format agnostic, and has preserved a wide variety of content including newspapers, electronic theses and dissertations, photographs, audio, video, and datasets. There are multiple information resources that describe the technical and organizational infrastructure of the MetaArchive Cooperative.[2] For the purposes of this paper, we will focus on two specific areas where MetaArchive has worked to support member institutions in preserving, protecting, and recovering digital content from 21st century disasters; system integration; and support for small cultural heritage organizations.

System Integration

From its inception, the MetaArchive Cooperative has sought to develop, implement, and evolve its technical and organizational infrastructure to support the preservation of a wide variety of digital content from multiple organization types. In relation to technical infrastructure, this has entailed developing and modifying mechanisms to ingest content from a number of digital collections management and repository systems. Currently, MetaArchive member organizations use a wide range of platforms to acquire, manage, describe, and provide access to their digital content. These include open source (e.g. DSpace, Hydra/Fedora), commercial (e.g. bepress Digital Commons), and homegrown software applications. Each of these applications may be built on different code bases, and may organize and structure digital content in different ways. All of these factors must be taken into account when building and implementing connectors between these systems and the MetaArchive PLN to ingest and preserve digital content in a secure and automated manner. This building and customization process entails rigorous testing to ensure that the digital content is accurately and completely ingested to allow for future recovery, restoration, and access.

As with any development process, the more customization that is needed, the more resources (both human and technical) that will be required. Additionally, as digital collections and repository systems are updated and upgraded, these changes may require additional technical modification to allow for continued ingest and preservation into the MetaArchive network. To

meet its goal of keeping digital preservation costs low for members, the MetaArchive Cooperative has continually strived to achieve a balance between supporting the ingest and preservation of digital content from multiple systems and working towards simplified, common ingest pathways that entail less customization and troubleshooting. Below we present a case study illustrating the efforts undertaken to successfully integrate two proprietary digital collections and repository platforms at a MetaArchive member institution, Carnegie Mellon University.

Carnegie Mellon University's Digital Collections

Carnegie Mellon University (CMU) has been building digital libraries for more than twenty-five years. CMU Libraries currently licenses two major proprietary platforms, each supporting a distinct set of collections and functions. ArchivalWare, by PTFS[3], is the platform that supports homegrown digital archives collections of text, photographic images, audio and video files. Bepress' Digital Commons[4] is the platform for the university's institutional repository (IR), which consists primarily of journal articles published by faculty. For the vast majority of the articles, the full text is available in the IR; for the other articles, the IR provides only a link to the full text that exists behind a paywall. The university's electronic theses and dissertations (ETDs), Honors Theses, and technical reports are also available via the IR. The two platforms provide free online access to approximately two million digital items. Given the considerable amount of human and technical resources devoted to the creation of these digital assets over the past two and half decades, Carnegie Mellon is understandably concerned about their long-term maintenance and preservation.

Rationale for Joining MetaArchive

Today, concern for long-term maintenance and preservation of digital collections informs our routine digital curation practices, but this was not always the case. Best practices for digital preservation are now just emerging. In the past, system administrators learned how to protect data from corruption or loss through trial and error. In 2003, a system administrator in the CMU Libraries wrote a faulty backup script for the libraries' digital collections, resulting in the loss of approximately eight months' worth of digitization and description work. Thereafter, project personnel became determined to find a more reliable solution to long-term preservation of data. A program was developed to help ensure that the data belonging to homegrown digital collections was routinely checked, looking for backup and server errors, corruption of files, and data loss[5]. While this in-house program proved to be successful for a time, it did not offer the level of data redundancy required to meet emerging best practices. With digital libraries still in their early stages, the need for a more comprehensive solution to the problem of long-term data preservation was needed.

Early in the implementation of the LOCKSS software, CMU Libraries was invited to serve as a LOCKSS test site. CMU Libraries set up a beta server and began ingesting a very limited amount of subscription journal content into a dark archive. Although Carnegie Mellon did not formally join the LOCKSS Alliance until 2012, working on the LOCKSS beta site allowed one library developer to become familiar with the software and the technical requirements for the LOCKSS system. By 2012 the MetaArchive Cooperative had matured enough to become a viable community interested in digital preservation. At that time, any institution wishing to join the MetaArchive Cooperative needed to also join the LOCKSS Alliance because the MetaArchive is

a private LOCKSS network. Deciding to join the MetaArchive Cooperative and the LOCKSS Alliance allowed CMU Libraries to begin working on a more comprehensive digital preservation strategy for digital assets across the entire library system, including homegrown digital collections, content in the university's institutional repository, and open access and licensed subscription journals for which the publisher has signed a LOCKSS agreement.

CMU Libraries joined the LOCKSS Alliance and the MetaArchive Cooperative in 2012 because they offer valuable resources that CMU Libraries could not have accessed if working independently. Making the data from digital collections redundant across multiple servers in remote locations, housed at other like-minded institutions with tools to verify the integrity of the data, is part of a series of best practices related to digital curation.

ArchivalWare plugin Development and Testing

After joining these two organizations, the next challenge was to determine how best to develop the plugin[6] for each system that would allow for the harvesting of data from two different proprietary platforms into the MetaArchive Network. The plugin allows each system to export data in a manner that it could be ingested into the MetaArchive Network. Because the systems use different metadata schemas and support different file types, each system required its own unique plugin.

Although the ArchivalWare platform is fairly open, and staff members at PTFS were willing to provide support to help CMU Libraries' developers build the plugin, library developers resisted doing any development to a proprietary system not owned by the university. As a result, PTFS was contracted to create the plugin and make adjustments during the testing phase of the project. Developers at PTFS worked closely with MetaArchive staff members throughout much of 2013. Staff members at the CMU Libraries learned how to define CMU collections and identify appropriate archival units using the Conspectus collections management tool provided by MetaArchive.

The Carnegie Mellon collections in ArchivalWare are organized and described in multiple ways. Archival collections of digitized faculty papers have a highly extensible version of Dublin Core, stretched to accommodate EAD metadata. The file formats for these collections include the black and white, 600 DPI master TIFF files and derivative PDFs. Archival journal collections have a Dublin Core metadata schema and color, 300-400 DPI master TIFF files and derivative PDFs. The rare book collections have a MARC metadata schema cross-walked to extensible Dublin Core and color, 300-400 DPI master TIFF files and derivative PDFs. Initially, the project manager made the decision to ingest all the metadata and PDFs from all of the collections. However, after calculating the cost of ingesting and distributing the 21 terabytes of color master TIFF files from the journal and book collections, CMU Libraries' administration decided not to include them in the MetaArchive network. Expectations are that if the cost of storage continues to drop, or new storage options become available, the decision to include all master files in the MetaArchive network will be revisited. In the meantime, the CMU Libraries also keeps multiple back-ups of all data on storage tape.

Deciding to include some, but not all, elements of a set of digital collections meant that the plugin had to be designed to know which parts of the collections to include and which parts to exclude from harvesting. In addition, the plugin had to support incremental crawls to capture

only new records, or records that had been recently modified, to avoid harvesting multiple copies of identical records on the same set of servers. After work was completed on the plugin the project moved into the testing phase.

The collections selected as test subjects were small in size and representative of each of the organizational structures within ArchivalWare. CMU chose small, representative collections so that they could be crawled quickly and the test would fail quickly if the plugin didn't work correctly. Testing these collections would provide insights as to what would happen when the entire set of collections were harvested. The initial export of data from ArchivalWare and its ingestion to the MetaArchive was successful. However, knowing that data is successfully stored in a dark archive is of little value if the organization cannot get it back out in a way that is useful. In order for CMU Libraries to feel confident that it could successfully provide long-term preservation using the MetaArchive network, the Libraries needed to export the harvested data from MetaArchive and restore it back into ArchivalWare. In running the restoration process, the project team discovered that the data exported from MetaArchive had to be preprocessed before being re-ingested into ArchivalWare. Developers at PTFS created an Export Package Preprocessor Tool and integrated it with the ArchivalWare system. After this tool was developed, the restoration process was successfully completed. CMU Libraries' homegrown digital collections now have a viable disaster recovery process. By working with the CMU Libraries and the MetaArchive Cooperative, PTFS learned how to improve their product in a meaningful way for their other digital library customers.

BePress Development and Testing

At the urging of the CMU Libraries, bepress agreed to cooperate with the staff at the LOCKSS Alliance and the MetaArchive Cooperative to develop a plugin that would allow for the harvesting of collections within CMU's institutional repository, called Research Showcase[7]. The development of the bepress plugin occurred in parallel with the creation of the ArchivalWare plugin.

The collections in the institutional repository consisted of Dublin Core metadata and a PDF file. Development of the plugin was completed by staff at LOCKSS and bepress. As bepress hosts Carnegie Mellon's data on their servers, testing was done at bepress' server facility. The work was completed in late 2013 and, as a result, bepress collections can now be added to the MetaArchive network without the additional expense of joining the CLOCKSS Network, another private LOCKSS network.

Project Resources

During both system integration efforts, each of the project partners contributed a significant amount of resources. CMU Libraries contributed staffing time and financial support for development work as well as membership fees in the LOCKSS Alliance and the MetaArchive Cooperative. During 2013 when the project was first launched and the development and testing was ongoing, the project manager spent approximately 20% of her time facilitating communications between partners. Currently, a CMU Libraries system administrator spends less than 5% of his time setting up and maintaining the MetaArchive and LOCKSS servers. The University Archivist spends less than 5% of her time adding collection information into the

Conspectus collections management software whenever a new collection is added to the MetaArchive Network.

The MetaArchive Cooperative and the LOCKSS Alliance contributed staffing to this project as well. Staff from MetaArchive trained staff at CMU Libraries and PTFS to use the system. They also participated in the testing of the plugins and contributed to project meetings. Staff from the LOCKS Alliance worked with members of bepress to develop and test the bepress plugin at no additional cost to Carnegie Mellon.

PTFS did charge CMU Libraries for the development work on the ArchivalWare plugin, but this cost was far less than if the Libraries had to develop the plugin themselves, in which case the library developers would have had to learn both the ArchivalWare and MetaArchive architectures.

Each project partner received something valuable from participating in this project. The LOCKSS Alliance and the MetaArchive Cooperative received two new plugins that they can use to attract other customers of bepress and PTFS who might be interested in working with them on digital preservation. PTFS gained a better understanding of what digital libraries need in terms of digital preservation. Bepress can now offer preservation functionality with their product. Finally, CMU Libraries is now confident that it can preserve important digital assets of the university in case of disaster.

Simplified Ingest Pathway

As described in the above case study, the resources required from multiple organizations to support the integration of multiple software platforms and systems with the MetaArchive PLN can be significant, but they function as valuable contributions to an overall effort to ensure successful preservation and disaster recovery. The MetaArchive Cooperative plans to continue support for such customized development activities, but is also in the midst of exploring alternative ingest pathways that would provide similar system integration opportunities for members. This simplified ingest pathway investigation is focusing on the utilization of the BagIt file packaging format for transfer and storage of digital content[8]. MetaArchive originally investigated the use of a BagIt-based ingest process during a series of recent research projects[9], and has been supporting the BagIt-based ingest pathway within the MetaArchive PLN for approximately the past five years.

Simply described, the BagIt-based ingest pathway still entails use of the LOCKSS plugin mechanism, but it's based on the premise that this BagIt plugin will remain fairly fixed and not require additional resource-intensive customization to connect multiple software platforms and systems. In certain contexts, this alternative ingest pathway would potentially require members to export content from repository platforms before ingest into the MetaArchive PLN could occur. The MetaArchive Cooperative has organized a working group to conduct testing and analysis to determine the viability and feasibility of this BagIt-based ingest approach. This working group will provide a set of recommendations to the MetaArchive Cooperative Steering Committee to inform any decision-making around future technical infrastructure changes.

Small Cultural Heritage Organizations

While still relatively limited in number, there are more options available today in the form of software tools, storage services, policy guidance, etc. than ever before, which enable and assist cultural heritage organizations in mitigating against the range of large and small disasters that can impact long-term preservation and access to their digital collections. Despite this growth in digital preservation solutions, small cultural heritage organizations still face many challenges that obstruct their efforts to preserve their digital content. Small museums, historical societies, and public libraries continue to increase their digital holdings, whether through digitization of existing collections, or acquisition of born-digital content. The digital collections within such organizations often document local and regional history, activities, and events and contribute significant value to our overall cultural knowledge and heritage, yet are vulnerable to the same types of risks - including natural and manmade disasters - as those managed by larger cultural heritage institutions.

The primary challenge faced by small cultural heritage organizations is the limited amount of resources available for digital preservation activities. Such organizations typically have small overall budgets and rely on external grant-based funding for digitization projects, which does not usually include funding for ongoing preservation activities. The number of full-time staff tends to be low, and these individuals often have numerous responsibilities, with the preservation of digital content being a small fraction of their immediate concerns. Additionally, volunteers often assist with collection management and/or information technology-related operational tasks. These characteristics make it difficult for small organizations to engage in digital preservation activities and necessitate the development of solutions for organizations with limited resources.

As a nonprofit, membership-based organization, the MetaArchive Cooperative has long embraced collaboration as an approach to increase the opportunities for organizations of all sizes to directly engage in the preservation of their digital collections. This principle of collaboration is embedded throughout both the organizational and technical infrastructure. On the organizational side, the MetaArchive governance model is designed to ensure that the needs and concerns of all members, regardless of organization size or membership level, are considered in strategic and operational decision-making. This governance model and shared decision-making process led to the creation of a Collaborative membership category, where multiple small organizations can band together to share the cost of membership fees in order to provide a lower cost alternative to the single institution membership categories. In regards to technical infrastructure, member organizations maintain control and ownership over the local hardware needed to operate the PLN, and equally participate in a collaborative effort to preserve other members' content. This local ownership and control allows member institutions to select from a range of hardware options, including low cost options that still meet baseline technical requirements.

Despite these efforts to make doing the work of digital preservation feasible and affordable, the barriers to entry and participation in preserving and protecting content from 21st century disasters are still high, particularly for smaller organizations with limited resources. In recognition of this state of affairs, the MetaArchive Cooperative has recently launched an effort to explore the feasibility and viability of making changes to its network technical infrastructure to lower the barrier to entry for small organizations. Specifically, this exploration is investigating the feasibility of a shift to a "supernode" network model, where a limited set of member

institutions would serve as geographically distributed replication sites for preserving content in the network. In this alternative network model, not every member institution would be required to purchase, configure, maintain, and refresh computer hardware to serve a replication node in the network. Instead, members not serving as a replication site would have an option to pay a separate fee to have their digital content ingested and preserved within the network. This fee would assist in supporting the costs incurred by those member institutions continuing to serve as network replication nodes. This exploration is still in an early stage, with any decisions around changes to network technical infrastructure to be decided upon by the MetaArchive Cooperative Steering Committee, but is conceptually promising as a potential approach to allow smaller organizations to start engaging in digital preservation activities.

References

1. Private LOCKSS Networks - https://www.lockss.org/community/networks/
2. Skinner, K., & Schultz, M., eds. (2010). A Guide to Distributed Digital Preservation. Atlanta, GA: Educopia Institute: http://www.metaarchive.org/GDDP
3. ArchivalWare PTFS renamed Knowvation in 2015 - http://www.ptfs.com/knowvation
4. BePress Digital Commons - http://digitalcommons.bepress.com/
5. Michalek, G. V., "Requirements and Characteristics of a Preservation Quality Information Management System," The Journal of Zhejiang University SCIENCE (2010), https://works.bepress.com/gabrielle_michalek/6/
6. LOCKSS plugin description - https://www.lockss.org/about/how-it-works/
7. Research Showcase - http://repository.cmu.edu
8. Bagit packaging format description - http://tools.ietf.org/html/draft-kunze-bagit-06
9. Chronicles I and II projects - https://educopia.org/research/chronicles