IFLA WLIC 2013
SINGAPORE

Submitted on: July 19, 2013

Chronicles in Preservation: Preserving Digital News and Newspapers

Nick Krabbenhoeft
Educopia Institute, Atlanta, Georgia, USA
E-mail address: nick@metaarchive.org

Katherine Skinner
Educopia Institute, Atlanta, Georgia, USA
E-mail address: katherine@metaarchive.org

Matt Schultz
Educopia Institute, Atlanta, Georgia, USA
E-mail address: matt.schultz@metaarchive.org

Frederick Zarndt
IFLA Newspapers Section
E-mail address: frederick@frederickzarndt.com

Abstract:

Since the mid-1990s, libraries and archives have been digitizing newspapers for preservation and access. The standards used for this work have evolved significantly during this time. Modern collections employ digitization techniques, metadata extraction and standards, and file formats that are very different compared to early collections. Increasingly, libraries and archives also include born-digital material. Whether collected as transfers of production masters directly from newspaper publishers or harvested from their websites, this material also differs greatly in its composition, metadata, and other collection characteristics. Given the importance of newspapers as primary documents of history, libraries and archives must preserve their digitized and born-digital collections carefully.

The National Endowment for the Humanities (NEH) has funded the Chronicles in Preservation project to study the preservation readiness of digital newspaper collections. Led by the Educopia Institute (www.educopia.org), the project has brought together seven American academic libraries and three distributed digital preservation (DDP) systems—MetaArchive, Chronopolis, and University of North Texas's Coda repository. Together, these partners are accomplishing a range of activities. First, they investigated community standards, specifications, and practices for digital newspaper collections and distilled this information into a set of Guidelines for Digital Newspaper Preservation Readiness. The Guidelines are now available for public review (publishing.educopia.org/chronicles).

Second, they staged test content exchanges, exporting collections from the libraries, and ingesting them into the DDP systems. The DDP systems are documenting this experience in a Comparative Analysis of Distributed Digital Preservation Frameworks. Finally, the project is augmenting a set of existing digital preservation tools to simplify the packaging and exchange of digital newspaper collections.

This paper provides a walkthrough of the structure and contents of the Guidelines for Preservation Readiness of Digital Newspapers, shares the evaluative metrics for the Comparative Analysis of Distributed Digital Preservation Frameworks, and discusses the implementations of the interoperability tools.

Keywords: Distributed digital preservation, digital newspapers, digital curation

---

1 INTRODUCTION

Libraries, archives, historical societies, and other organizations have digitized analog newspapers and have collected born-digital newspapers for decades. Over that period the best practices have gradually improved. As a result, curators must manage collections created with many different standards. Some newspapers might have been scanned as TIFF. Some might have been scanned as JPEG/2000. Descriptive metadata might be Dublin Core, METS-ALTO, or not recorded at all. Born-digital content could be given by the newspaper or harvested by the library. The Chronicles in Preservation project studies the preservation challenges of curating these digital newspaper collections. It is a partnership funded by the National Endowment for the Humanities (NEH) [1] between three distributed digital preservation (DDP) systems and seven U.S. libraries that manage digital newspaper collections.

The three DDP systems are MetaArchive[2], Chronopolis[3], and UNT[4] Coda. DDP systems assist with long-term preservation by storing, curating, and validating objects in multiple secure locations. Many different technologies can be used to accomplish this task. In the case of the Chronicles partners, MetaArchive uses LOCKSS (Lots of Copies Keeps Stuff Safe), Chronopolis uses iRODS[5] (Integrated Rule-Order Data System), and Coda uses a suite of microservices[6], some of which have been developed at UNT. The test exchanges for the Chronicles project are allowing the DDP partners to compare the relative advantages and disadvantages of their technical frameworks for particular content types and repository frameworks.

The libraries of University of North Texas (UNT), Penn State, Virginia Tech, University of Utah, Georgia Tech, Boston College, and Clemson University contributed a variety of digital newspaper collections for the project. Some of the collections were digitized as part of the National Digital Newspaper Program (NDNP)[7], an extremely successful grant program by the

---

[1] http://www.neh.gov/
[2] http://www.metaarchive.org
[3] http://chronopolis.sdsc.edu
[4] http://www.library.unt.edu/
[5] California Digital Library. "Microservices." UC Curation Center: Curation Wiki. https://www.irods.org/
[6] https://wiki.ucop.edu/display/Curation/Microservices
[7] http://www.loc.gov/ndnp/

NEH and Library of Congress[8] to digitize and federate newspaper collections for access using a standard set of specifications. Libraries with experience using these standards have also employed them in non-NDNP funded digitization projects. On the other hand, a number of libraries had digitized collections before NDNP began in 2003 or did not have the resources to achieve NDNP standards without NDNP grants. In terms of born-digital material, libraries also hold collections ranging from early HTML pages to PDF print masters. As a result of this collection diversity, the Chronicles project has been able to study a wider range of digital preservation challenges.

The Chronicles project is producing three deliverables.

1. Guidelines for Digital Newspaper Preservation Readiness – These documents record practical advice for organizations that curate digital newspapers. The Guidelines are based on interviews with stakeholders in the digital newspaper community, experience gained during the project, and the findings and recommendations documented in reports and standards like NDNP.
2. Comparative Analysis of Distributed Digital Preservation (DDP) Frameworks - The DDP systems are comparing their experiences with the project test exchanges by documenting workflows and recording a shared set of objective metrics. The goal of this analysis is for each DDP system to understand how their infrastructure handles collections of a specific kind of content at levels of varying preservation readiness.
3. Interoperability Tools – The Chronicles project tested, documented, and packaged a set of existing digital curation and preservation tools, making them easier to understand and implement in a wide variety of local technical environments.

2 GUIDELINES FOR DIGITAL NEWSPAPER PRESERVATION READINESS

Resources like the National Digital Newspaper Program (NDNP) provide guidance for libraries, archives, historical societies, and other organizations that want to digitize their collections. However, curators do not have a resource that describes how to improve the preservation conditions for existing collections created to varying standards. The Chronicles project is publishing the Guidelines for Digital Newspaper Preservation Readiness for this audience.

The information for the Guidelines came from many sources. Stakeholders in the newspaper industry, library community, and digital preservation community were interviewed. The project's library partners contributed case studies about specific topics in their programs. Standards such as OAIS (Open Archival Information System)[9], METS (Metadata Encoding and Transmission Standard)[10], and NDNP were referenced. The Chronicles project has now released the draft Guidelines for public review at http://publishing.educopia.org/chronicles. The goal of the public review period is for curators of digital newspaper collections to help evaluate, refine, and add to the documents. Community feedback will be incorporated into the Guidelines before their official publication at the end of the year.

The Guidelines include six main modules:

[8] http://digitalpreservation.gov/
[9] http://public.ccsds.org/publications/archive/650x0m2.pdf
[10] http://www.loc.gov/standards/mets/

1. Inventorying Digital Newspapers for Preservation – How to record what content an organization has and how it is stored
2. Format Management for Digital Newspapers – How to identify, validate, and migrate formats
3. Metadata Packaging for Digital Newspapers – How to choose metadata formats, export metadata from repositories, and manage the storage of metadata
4. Checksum Management for Digital Newspapers – How to generate and monitor fixity information
5. Organizing Digital Newspapers for Preservation – How to structure folder hierarchies and names
6. Packaging Digital Newspapers for Preservation – How to organize a collection for ingest into a digital preservation system

Each module can be read and used independently of the other modules. Additionally, there are six subtopics with references to deeper studies on topics such as collection acquisition and change management. The structure of Guidelines allows curators to reference and adapt only the portions necessary to address their particular concerns.

The Guidelines recognize an important fact about the digital preservation community, namely that every organization has different resources in terms of finances, skills, and tools, among other things. Different organizations cannot all perform the same preservation actions. Therefore, the Guidelines are organized as a spectrum from essential to optimal.

- Essential actions are the minimum actions required for an object to be preserved. These can be completed with limited resources.
- Optimal actions are the actions that will best preserve an object. These are only possible with more resources.

To explain this approach, a sample of the Inventorying Digital Newspapers for Preservation module follows.

2.1 An Example Module from the Guidelines

Understanding the content of a digital newspaper collection is the first step to increase its preservation readiness. Collections are created over time. Even within that time, a collection's curators, acquisition strategies, storage media, and file formats can change. By recording this information in an inventory, curators can quickly understand the current condition and preservation risks of the collection. This is essential in planning future preservation actions.

Because the inventory is for curators, it must be human-readable. As an example of a human-readable inventory, the California Digital Library distributes an inventory template in Rich Text Format (RTF)[11]. However, digital newspaper collections can include tens of thousands of files. To better access, manage, and interact with it, a better inventory would also be machine readable, such as a standardized text document, a spreadsheet, or a database. Information recorded in the inventory can include file counts, file size, lists of files and folders, file paths, formats, checksums, and object identifiers.

The tools available to create an inventory depend on the resources available to an organization. The file managers included in every operating system, such as Nautilus for

[11] http://www.cdlib.org/services/dsc/contribute/docs/submission.inventory.rtf

Linux, Finder for Mac, and Windows Explorer can be used to estimate file counts, the size of the collection, and included formats. Curators that are comfortable using command-line functions in UNIX such as ls, find, locate, and file can export the directory structure of a collection. Tools built for the digital preservation community like the popular BagIt[12] utilities can require more resources for training, documentation, and support, but they also provide well-formatted information. For instance, BagIt utilities record the file path and checksum for each object in the bag.

The essential level of an inventory only requires familiarity with a file manager. It includes an identification of the collection, a list of included newspaper titles, the number of files, the physical location of the files, a list of file names, and the date of the inventory. This information is descriptive enough that a curator can track a collection and plan future preservation actions.

In an optimal situation, inventorying would take place on a single computer to control for consistency. In addition to capturing the information for an essential inventory, the computer would have programs to monitor files for changes like format migration, to identify file formats and the associated applications, to create checksums, and to record object identifiers. To fulfil the optimal level, an organization needs the sufficient computing resources, the skills of a curator familiar with selection of tools such as BagIt, PRONOM, and JHOVE and checksum utilities, and the time necessary to compile in depth inventories of every collection. Investing these resources early in the curation process simplifies later stages such as packaging metadata and managing checksums.

3 TESTING THE GUIDELINES

Before writing the Guidelines, the Chronicles team prepared a Preservation Readiness Plan for each library based on the collections they planned to exchange. After completing the actions outlined in the plans, the libraries exported their collections for the DDP networks to ingest. The process gave valuable feedback about the modules needed for the Guidelines and the strengths and weaknesses of each DDP network.

The collection preparation process began with a staging period. Each partner installed the software necessary to process collections, namely BagIt, the project's chosen packaging tool. The partners then staged the collections to a server for the purposes of preparation. Then, representatives from the library and DDP networks met to coordinate their communications, agree upon process, and grant necessary server permissions. After the meeting, the library began the data wrangling process: creating an inventory, evaluating the sustainability of file formats in a collection, generating checksum manifests, ensuring the collection structure and file names were stable, validating metadata, and packaging the collection in a bag. One of the most valuable results of this process was the discussion generated from evaluating choices such as what to consider as a preservation format, how to package collections on local storage, and how to allocate resources for resource-intensive tasks like generating checksums.

[12] California Digital Library, "BagIt." UC Curation Center: Curation Wiki. https://wiki.ucop.edu/display/Curation/BagIt

4 COMPARATIVE ANALYSIS OF DISTRIBUTED DIGITAL PRESERVATION FRAMEWORKS

The collection exchange test is also offering the DDP networks an opportunity to evaluate their technical frameworks. The result of this evaluation will be published as the Comparative Analysis of Distributed Digital Preservation Frameworks. The structure of that document will be discussed here.

MetaArchive, Chronopolis, and UNT Coda employ distinct technologies to accomplish distributed digital preservation. MetaArchive stores copies of collections on seven international geographically distributed servers using LOCKSS (Lots of Copies Keeps Stuff Safe). Collection integrity is monitored through the Conspectus, and collections are frequently revisited to capture new or changed objects. The servers regularly compare their copies to monitor for data corruption. Chronopolis uses iRODS (Integrated Rule-Order Data System) to maintain data in three locations spread across the United States. Data validity is monitored by ACE (Audit Control Environment). UNT Coda uses a suite of microservices to replicate data on a local level, check fixity, and log preservation actions.

Each DDP system is documenting the workflow necessary to preserve a collection in their network. For instance, because the process of comparing copies can require significant server resources, the MetaArchive limits the size of Archival Units (AUs) in its system to 30 GBs. However, libraries provided digital newspapers in bags of more than 200 GBs. As a result, MetaArchive created scripts to split large bags into a series of 30 GB bags and to later reproduce the original bag. During the test exchanges, each library successfully validated their recombined bags. The Chronicles project plans to publish the bag handling scripts with an open-source license for community use.

5 INTEROPERABILITY TOOLS

The third Chronicles deliverable is a set of lightweight services that organizations can use to achieve actions outlined in the Guidelines. Although a number of digital preservation tools exist that could fulfil recommendations from the Guidelines, they can lack documentation or generalizations that make them useful for the wider digital preservation community. The goals of the Chronicles project with the set of interoperability tools is to build upon the substantial work that created these tools to increase their usability. The tools include:

- DAITSS Description Service[13] – The Florida Center for Library Automation (FCLA)[14] released this web app in 2009 as a service that uses DROID (Digital Record Object Identification)[15] and JHOVE[16] to identify and validate formats. Results for each file are recorded as a PREMIS (Preservation Metadata: Implementation Strategies)[17] xml object.
- Bagger[18] – The Library of Congress released the Bagger tool in 2012. Because it is a Java-based BagIt utility with a graphical user interface (GUI), it can run on nearly any computer and has a much lower learning curve than command-line BagIt tools.

[13] http://description.fcla.edu/
[14] http://fclaweb.fcla.edu/
[15] http://sourceforge.net/projects/droid/
[16] http://sourceforge.net/projects/jhove/
[17] http://www.loc.gov/standards/premis/
[18] http://sourceforge.net/projects/loc-xferutils/files/loc-bagger/

- bagit.py[19] – Ed Summers from the Library of Congress has maintained bagit.py since 2010. As a command-line tool, it can be more easily incorporated into scripts such as the MetaArchive bag splitter. Additionally, it might be preferable for curators more comfortable with command-line interfaces.
- UNT PREMIS Event Service – Written as part of the UNT Coda suite of microservices, the PREMIS Event Service monitors the actions performed on files stored on a server. When the server duplicates a file, checks its checksum, or performs other actions on it, the Event Service logs the action in a PREMIS object linked to the file.

The DAITSS Description Service, Bagger, and BagIt were tested as part of the Preservation Readiness Plans. The tools assisted with completing the following Guidelines recommendations.

- Inventorying – Essential – Document file names and locations
    - The BagIt manifest includes the full path and filename of every object in the bag.
- Format Management – Essential – Identify file formats
    - The DAITSS Format Identification Service records file formats in a PREMIS record. Using this list, curators can plan policy about format normalization and migration.
- Checksum Management – Essential – Create checksums
    - All BagIt tools create checksums for each object in its data folder.

Many essential actions can be completed manually or with simple tools, such as file counts and organizing directories. However, some require repeated, complex actions such as deriving checksums. With lightweight, automated services, curators can gather the information to decide on larger collections issues, for instance, format policies.

It is important to note the use of open-source software in the set of operability tools. Open-source licenses allowed for the FCLA to build a web app for DROID and PREMIS and for the Chronicles project to adapt that work further. The Chronicles project has also suggested improvements for bagit.py that have been incorporated into the source code. Currently, the project is helping to test the UNT PREMIS Event Service for release under an open-source license. It is hoped that other repositories will use and extend this tool even further.

6 CONCLUSION

The Chronicles project has developed a collection of resources to help curators manage digital newspapers. The resources are written to be useful to all organizations. Future digital newspaper programs will hopefully find guidance to incorporate into their workflows from the beginning. Established digital newspaper programs will find useful tools to improve their curation workflows. The Chronicles project has been a community project from the start. We welcome suggestions for improvement from the community.

[19] https://github.com/edsu/bagit