Building Institutional Capacity in Digital Preservation

Matt Schultz
Educopia Institute
Atlanta, Georgia 30309
matt.schultz@
metaarchive.org

Mark Phillips
University of North Texas
Denton, Texas 76205
mark.phillips@unt.edu

Nick Krabbenhoeft
Educopia Institute
Atlanta, Georgia 30309
nick@metaarchive.org

Stephen Eisenhauer
University of North Texas
Denton, Texas 76205
stephen.eisenhauer@unt.edu

ABSTRACT

The *Chronicles in Preservation* project, being led by the Educopia Institute, is undertaking research to evaluate the degree to which several of the current digital preservation standards in use today (e.g., OAIS, TRAC, PREMIS, METS, etc.) can be applied to the diverse and at-risk content genre of digital newspapers. Institutions need guidance on incremental, skilled approaches and lightweight tools and resources if they are going to begin caring for such content in achievable and yet sustainable ways. The *Chronicles* project has researched, experimented, and begun advocating for a variety of skills, tools, and other resources that both embrace the current standards and seek to implement them in lightweight ways. They incorporate, apply and extend a number of existing as well as leading edge advancements such as BagIt, the DAITSS Description Service, UNT's PREMIS Event Service, and the NDSA Levels of Preservation.

Categories and Subject Descriptors

D.2.7 [**Distribution, Maintenance, and Enhancement**]: Extensibility; E.5 [**Files**]: Organization/structure; H.2.1 [**Logical Design**]: Data models; H.3.7 [**Digital Libraries**]: Collection, Standards

General Terms

Documentation, Design, Experimentation, Management, Performance, Standardization

Keywords

BagIt, DAITSS Description Service, Digital Newspapers, NDSA Levels of Preservation, PREMIS, PREMIS Event Service, Standards

1. INTRODUCTION

The *Chronicles in Preservation* project, being led by the Educopia Institute, has been contributing to the recent trend towards helping institutions take more manageable and incremental steps toward accomplishing their digital preservation. It has been doing so by researching institutional capacities for implementing existing standards (e.g., OAIS, TRAC, PREMIS, METS, etc.), and doing so in the context of one highly valued, yet at-risk set of digital assets—digital newspapers. The project has discovered that institutions need more lightweight approaches, less imposing data models, improved guidance, and non-sophisticated technologies in order to begin accomplishing their digital preservation and laying a foundation for more robust activities down the road. This paper will explain the trend toward incremental approaches; the research done in the *Chronicles* project that underscores the need for such approaches; and how the project is producing skills, technologies, and other resources to meet those needs.

2. BUILDING CAPACITY

Digital newspapers are a valuable, unique and at-risk set of scholarly assets. For more than a decade, stewards of historical newspaper holdings in the U.S. have been hard at work under the United States Newspaper Program (USNP) and the National Digital Newspaper Program (NDNP) to microfilm, catalog, digitize, archive and make accessible newspapers in the public domain. Under NDNP, the technical standards for digitizing this massive corpus of materials have achieved approval and uptake more broadly. Institutions seeking to digitize their newspaper holdings for long-term preservation now have a set of highly reputable and open standards to follow.

The NEH-funded *Chronicles in Preservation* project `http://www.metaarchive.org/neh`, is seeking to evaluate the degree to which the NDNP standards, and digital preservation standards more broadly (e.g., OAIS, TRAC, PREMIS, METS, etc.), can be applied to digital newspapers going forward, particularly in an environment where grant funds are becoming more scarce. The *Chronicles* partners all value the importance of following standards for achieving sound digital preservation but have first-hand knowledge that doing so can be costly. For that reason, the *Chronicles in Preservation* project is attempting to evaluate the current needs for preservation readiness of digital newspapers in all its wide diversity of forms (including born-digital and digitized), and identify the proper application of standards along a spectrum of achieving a minimum *essential* level of conformance up to a more robust *optimal* level of conformance. The hope being that stewards of digital newspaper collections can understand what they can achieve in the short-term with limited resources, and work their way up towards over the long-term with respect to existing standards.

3. IDENTIFYING SOLUTIONS

To make better sense of the current state of digital newspaper holdings, and the degree to which standards, and preservation oriented technologies have been applied toward their maintenance, the *Chronicles in Preservation* project has carried out a number of assessments, including:

1. a collections readiness assessment survey;
2. a sample data analysis; and
3. a focused set of interviews with digital newspaper stewards and curators (including the project partners, commercial publishers, state libraries, NDNP participants, as well as non-NDNP participants).

Each of these assessments have helped the project staff and partners gauge the gaps in resource availability for achieving various levels of conformance toward standards, and more importantly how best to improve and develop new skills, tools, and other resources that can help digital newspaper stewards to begin meeting various tiers toward preserving these valuable, unique and at-risk set of assets.

To begin with, a survey was formed that queried the project partners in four major areas, namely:

1. Collection & Repository Information: Partners were asked about the size and scope of their collections, formats, repository systems and other storage media in use, and whether they had ever been required to restore their collections under any scenarios of loss.
2. Collection Data Management: Partners were asked about their data management practices, including what sorts of object identifier schemes and file naming conventions were being used, how their newspaper data was structured, and the nature and extent of any metadata (particularly preservation metadata) that had been defined.
3. Preservation Assessment: Partners were asked about incidences of obsolescence or format migration or conversions for their digital newspaper files, if any, and what sorts of tools may have been used to manage such activities, as well as their perceived capacity for beginning to manage their digital newspapers from a more robust preservation perspective.
4. Ingest & Recovery: Partners were also asked about rates of collection growth, nature of changes and remediation, and any policies and practices that would have an impact on their ability to package and exchange their digital newspapers for a separate preservation system and what the parameters might be for recovering and rebuilding any preserved collections in the event of local loss.

It was found that institutions had a wide variety of local repository implementations and data management practices for their digital newspapers, had begun to do little more than routine backup for their content, and were not very far down the road toward applying preservation standards or technologies. It became very clear from the survey that digital newspaper stewards would require lightweight approaches for beginning to advance toward more standards-oriented practices for managing their digital newspaper collections.

Secondly, to observe first-hand the state of these digital newspapers, we proceeded to request sample newspaper data from each of our project partners Partners were asked to provide at least one full issue (up to 8 GB) worth of newspaper data for analysis. What emerged was that institutions had a range of different title/issue and sub-foldering schemes for their data, a variety of file-naming and object identifier schemes (often imposed by their repository/access systems), and varying amounts of descriptive, technical, administrative, and structural metadata. This made it clear that much work might be needed to apply some consistency across their collections for the purposes of packaging them for long-term preservation, and that this could prove to be a barrier for taking action in the short-term. Less imposing data models for preservation packaging were clearly in need.

Finally, effort was taken to reach out to stewards and curators of digital newspapers outside of the project to gain a broader perspective on the vast array of preservation challenges that may be facing such institutions. Interviews were arranged with a social media reporter for the Calgary Herald, a newsroom librarian at the Dallas Morning News, the State Librarian of the Wyoming State Library (Wyoming Newspaper Project), the University Librarian at UC Berkeley (California Newspaper Program), and the State Archivist at the Minnesota Historical Society to better understand how both born-digital and digitized news is being created, acquired, and managed in those contexts. In these interviews the urgency to get digital news under preservation quickly in the face of numerous institutional obstacles and barriers to partnerships was underscored. Institutions need help navigating existing standards, applying them in reasonable ways that respect their current capacities, and doing so with non-sophisticated technologies.

The *Chronicles in Preservation* project is working towards meeting this need by proposing, testing, and validating a combination of lightweight skilled approaches, technologies, and other resources to demonstrate how a diverse and complex set of digital assets like digital newspapers can be better curated in line with a tiered-spectrum of standards adoption and conformance. Below we talk about each of these skills, tools, and other resources. They include:

1. BagIt: Institutions need a beginning preservation data model in the absence of a consistent existing model;
2. Preservation Readiness Plans: Institutions need an incremental roadmap for improving curation and preservation packaging over time;
3. Simplified Preservation Metadata: Institutions need simpler (PREMIS) creation and management tools that can build off of data models like BagIt; and
4. Levels of Preservation Metadata Guidelines: Institutions need guidance on enhancing their data model and preservation metadata in incremental ways over

time. The NDSA Levels of Preservation are proving helpful in this area.

4. BAGIT DATA MODEL

In light of the information gathered in the previously mentioned survey and analysis of contributed test data, it became clear that there were a number of data models in use at the various partner institutions. Much digital newspaper content is being organized without a unifying data model that would allow institutions to make assertions and discuss characteristics of their underlying data in a consistent way. In order to resolve this challenge a decision to implement a data model utilizing the BagIt packaging specification was made [2]. The BagIt packaging specification has been used by a number of collaborative projects to package and share data between different technology and organizational platforms and was seen as an easy step towards a simple data model for the *Chronicles* project.

In order to implement the BagIt specification in the project the project team compiled a list of commonly used and maintained open-source BagIt tools, and documented them for project participants. In addition to the identification of these tools, we prepared a set of simple instructions outlining the tools and their use in the project. This documentation included installation information as well as a guide to the metadata fields (bag-info.txt). All of the partners reported success in making use of the simplified instructions. Making use of BagIt for their collections provided the partners with an opportunity to revisit their data structures, apply a simple packaging scheme for that data, and in many cases provided them with a previously non-existent layer of information (inventory and checksums) that could be elaborated on further (as will be discussed below). These simplified BagIt usage instructions will be made available for other institutions to use along with all of the project's code products at the conclusion of the project in April 2014.

5. PRESERVATION READINESS PLANS

In collaboration with the partner institutions, the project team also created a set of preservation readiness plans that established a number of lightweight preservation steps that could be applied to the partners' digital newspaper collections. These plans included contact information, roles and responsibilities for the collection, as well as scope of the collection in relation to the Chronicle in Preservation project. Additionally a series of goal statements followed by action plans for completing these goals were established for each partner.

These preservation readiness plans served as a starting point for conversation with collection owners to identify possible gaps in infrastructure, training, or tools at their institutions. A template example of these preservation readiness plans that other institutions can make use of will be made available along with all of the project's code products at the conclusion of the project in April 2014.

- Inventorying
- Checksums
- Format Identification

Inventorying files for the *Chronicles in Preservation* project involved partner institutions making explicit file-level inventories for content being used by the project. This inventory process aligns with the usage of the BagIt specification because the specification requires the creation of a manifest that defines the content of the valid bag.

Check summing of fixity information is another area of interest for the project participants. In sharing the readiness plans it became apparent that partner institutions varied widely in the tools used to generate fixity information and the use of that information for managing their digital newspaper collections. Identifying fixity as another area to focus was again complementary with the usage of the BagIt specification and data model for the project as the specification requires the inclusion of fixity information in the manifest in order to validate bags.

Finally format identification was identified as a goal in the preservation readiness plans. The readiness plans identified this as an area that would be more challenging for partners to implement locally than the previous two areas. A decision was made to build a set of identification services around the BagIt model that could be executed with limited overhead for the partner institutions. These services are described below.

6. SIMPLE PREMIS CREATION & USAGE

To simplify the process of performing format identification analysis over bagged collections of files, we leveraged the powerful DAITSS Format Description Service originally developed by FCLA (now FLVC) [1]. The service exists as a Ruby web application that can be run on a local machine, making it ideal for batch usage. We have developed a lightweight script, which when paired with the Format Description Service, can be used to analyze the entire contents of a bagged set of files and produce PREMIS records as output (stored within the bag itself). The script uses basic Unix commands to loop through the contents of a bag. Each file is sent to the Format Description Service (running on the local machine), and the resulting output is saved in a corresponding file inside a "premis" directory placed at the root of the bag. The output files are named and organized identically to the input files, with an ".xml" extension added at the end.

For more robust management and tracking of PREMIS data, we have also prepared the PREMIS Event Service software for general release in the near future. The PREMIS Event Service is a Django-based web application designed to manage and relate PREMIS records and related metadata in a database-driven system, originally built for internal use at UNT. The service provides a web-based user interface and REST API through which records can be fetched, queried, and stored in a way that allows for consistency and centrality throughout preservation systems [4]. With some light modifications, the scripting described above could be adapted so that the results of the Format Description Service is sent to the PREMIS Event service for storage. The project's code products will be made available and appropriately licensed for other institutions to make use of at the close of the project in April 2014.

7. NDSA LEVELS OF PRESERVATION

Finally, digital preservation standards such as OAIS and TRAC advocate for the application of a robust set of tools and practices to better accomplish long-term digital preservation, but these standards do not offer much practical guidance. The NDSA Levels of Preservation were formulated in response to this problem [3]. While they have been published only recently, the Levels have come to serve as a useful starter assessment resource for institutions. The *Chronicles in Preservation* project found the Levels especially useful for providing guidance on enhancing the proposed BagIt data model and preservation metadata in incremental ways over time.

The NDSA Innovation Working Group that is developing the Levels has suggested several methods for assessment including establishing a threshold level and providing an analysis for each stage and row of methods in the Levels. Because requirements for metadata can be found in all levels (even those outside of the metadata row), this assessment began by identifying all the suggested metadata requirements in the five categories:

1. Storage and Geographic Location: Important to retain metadata on accessible systems even in emergencies (Level 4)
2. File Fixity: Important to check or create fixity information for all objects on ingest (Level 1); virus check high-risk content (Level 2); check fixity at fixed intervals with logs, virus check all content (Level 3); check fixity in response to events (Level 4)
3. Information Security: Important to maintain logs of who performed what actions on objects (Level 3); audit logs (Level 4)
4. Metadata: Important to store object manifest separately (Level 1); store administrative and transformation metadata (Level 2); store standard administrative and technical metadata (Level 3); store preservation metadata
5. File Formats: Create an inventory of file formats (Level 2)

In the *Chronicles in Preservation* project, BagIt is a foundational tool (as described above). As mentioned, institutions often overlook creation of an object manifest. While the NDNP METS standards describe newspapers on an issue level, there is no requirement for a collection-level manifest. The BagIt specification includes a manifest of all objects in the bag with checksums. Creating and backing-up the manifest fulfills Level 1 Metadata. Bag validation utilities allow organizations to transfer bags and check fixity on ingest, in accordance with Level 1 File Fixity. The *Chronicles in Preservation* project also required the use of a bag profile to record administrative metadata for each collection such as the owning organization, contact information for the content's steward, the size of the bag, and a short description of the bag's contents, partially fulfilling Level 2 Metadata.

The Format Description Service mentioned above identifies file formats and creates corresponding PREMIS records once a collection or set of content has been "bagged". This process primarily accomplishes Level 2 File Formats requirements to inventory file formats, but wrapping the metadata in PREMIS also complements the administrative bag metadata in fulfilling Level 2 Metadata.

The final component of the Level 2 Metadata Requirements is logging transformative events that the organization performs on the objects over time. The creation of new derivative copies or the migration of master objects to new formats includes updating the metadata of the object. The PREMIS Event Service can provide ongoing monitoring of stored digital objects, allowing the organization to query changes in this metadata over time.

By utilizing the three tools above—BagIt, the Format Description Service, and the PREMIS Event Service—the *Chronicles in Preservation* project is able to automate the creation of nearly all metadata required below level 3.

8. CONCLUSION

Institutions engage digital preservation standards and methodologies with certain degrees of current capacity that determine what they can realistically accomplish in the short-term. There are legitimate trends in the community that are embracing incremental approaches. The *Chronicles in Preservation* project has underscored the need for such approaches and has sought to produce skills, tools, and other resources that embrace the current standards yet seek to implement them in lightweight ways—laying a foundation for more robust implementations over the long-term.

9. REFERENCES

[1] Florida Center for Library Automation. Daitss format description service, Apr. 2013.

[2] J. Kunze, C. D. Library, J. Littman, L. Madden, L. of Congress, and B. Vargas. The bagit file packaging format (v0.97). Internet-Draft draft-kunze-bagit-09, Internet Engineering Task Force, Apr. 2013.

[3] Library of Congress. Ndsa levels of digital preservation: Release candidate one. `http://blogs.loc.gov/digitalpreservation/2012/11/ndsa-levels-of-digital-preservation-release-candidate-one/`, Jan. 2013.

[4] M. Phillips, M. Schultz, and K. Nordstrom. Premis event service. In *Open Repositories 2011*, June 2011.