Building for the Future:
Preservation in the National Digital Newspaper Program

Deborah M. Thomas
Library of Congress
101 Independence Ave., SE
Washington, DC 20540-4760
deth@loc.gov

Shiow Huang
Library of Congress
101 Independence Ave., SE
Washington, DC 20540-1300
shua@loc.gov

ABSTRACT: This paper describes aspects of the technical development and policy decisions incorporated into building the National Digital Newspaper Program (NDNP), a partnership between the National Endowment for the Humanities (NEH) and the Library of Congress (LC). NDNP is a long-term effort to provide permanent access to a national digital collection of newspaper bibliographic information and selected historic newspapers, digitized by NEH awardees in all U.S. states and territories. This program, and managing the assets created, provides a rich testing ground for the development of large-scale digitization programs and predicting long-term needs for management and preservation of digital assets, both for NDNP and future projects. The current development phase focuses on ingesting and providing user-friendly 'access' to data produced according to experimental strategies for digital preservation. The lessons learned during this critical first phase will inform the design of future architecture and the implementation of additional management and preservation.

The National Digital Newspaper Program (NDNP), a partnership between the National Endowment for the Humanities (NEH) and the Library of Congress (LC) is a long-term effort to provide permanent access to a national digital collection of newspaper bibliographic information and selected historic newspapers, digitized by NEH awardees in all U.S. states and territories. This program builds on the legacy of the strategically-successful United States Newspaper Program (USNP), sponsored by the NEH and supported by the LC, for the past twenty years and concluding in 2007– an excellent example of successful collaboration at both the national level and within states to inventory, catalog, and preserve in microfilm the national corpus of at-risk newspaper materials. The new program not only extends the usefulness of USNP products, but also provides a rich testing ground for developing basic strategies for long-term digital content sustainability and management.

Historic newspapers are the primary record of events that shape our communities. They provide a venue for sharing the facts and opinions of moments in time, significant people, and local perspectives—a unique resource for recording and understanding the effects of both singular and united voices on ideas, events, and democratic identity, as well as defining the historic record. In recent decades, under USNP, the preservation of newspapers on microfilm and the establishment of imaging and bibliographic standards has been an important component of archival programs - however, even this critical aspect of newspaper librarianship does little to address the use and access needs of text-intensive newsprint. Utilizing this valuable resource, imaged on film or in original paper is a challenge for libraries and users alike, with its cumbersome physical aspects, discolored and brittle paper, and complex organization. Even with the best imaging standards and process, the intellectual content of the newspaper is contained in a complicated layout, with varying visual cues and small type faces, wearying to the eye and the mind. However, with the development of new technologies in digitization, text recognition, search engines, etc. the NDNP will provide enhanced access and discovery to this material, as well as the national leadership necessary to establish basic technical standards for the digitization and structure for historic newspaper materials. The primary goals of the program are long-term – provide enhanced access to select newspapers by creating and aggregating millions of digitized pages from geographically-diverse historic newspapers- expected to take 20 years - and a bibliographic and holdings directory of over 138,000 titles, created by USNP, in a freely-accessible and searchable repository.

Since 2004, the NEH and the LC have collaborated to develop a nationwide program that will enhance access to this material through the use of new technologies and information channels, scale to include representative content from all U.S. states and territories produced over several decades, and encourage interoperability between digital libraries through shared specifications and architecture. In 2005, NEH awarded $1.9 million among 6 institutions – University of California-Riverside, University of Florida, University of Kentucky, New York Public Library, University of Utah, and the Library of Virginia - to select and convert newspaper holdings representing their state collections. These awardees were selected for their experience with historic newspapers, digitizing collections, and digital library infrastructures. In the initial phase, currently underway, the program produced a developmental digital repository that stores hundreds of thousands of pages of historic newspapers converted, from both the collections of LC and NEH awardees, and in March 2007, the NDNP launched its Web service dissemination from that repository - *Chronicling America* at http://www.loc.gov/chroniclingamerica/. The site now includes a newspaper title directory with data created under USNP – 138,000 titles and 900,000 holdings – and over 225,000 pages of newspapers, published between 1900 and 1910. NEH will continue to hold annual award competitions to gradually increase the scope – both in geography and time – of the aggregated national collection and build expertise at the state level in large-scale newspaper digitization. The next NEH awards, for conversion of content published 1880-1910, will be announced in July 2007.[1]

In the development and overall management of the program, the Library of Congress provides technical support of the program's primary goal – creating open access to the nation's historic newspapers. The Library's role is three part: to establish technical digitization specifications that permit aggregation, access and preservation of content created by NEH awardees, to serve and unify this content through a publicly-available Web site, and to sustain the aggregated content permanently. As LC reviewed the means available to accomplish these tasks, it became clear the requirements of the final task – sustaining the content - would inform many decisions for the other tasks, including resource allocation, timelines, and program planning.

The evolving NDNP preservation environment is based on requirements to support four major workflows as identified in the Open Archival Information Systems (OAIS) Reference Model: ingest, archiving, dissemination and preservation/curation management. In addition, the system may eventually need to support other librarian or archivist business – such as providing analytic data for research (e.g. OCR accuracy, microfilm density) or data mining requirements. From the outset, LC recognized the scope of the planned program – millions of newspaper pages produced by many different organizations over approximately 20 years (equaling, at least, hundreds of terabytes) – and the commitment between Congressionally-funded agencies to manage these assets long-term required emphasis on the creation of digital assets according to standards and uniform practices and establishment of infrastructure, both mechanisms and capacity, to ensure cost-effective management of the content over time.

The Library's first steps included determining high-level operating principles and functional requirements for the digital asset system and the associated dissemination workflow. In a climate of emerging (and evolving) best practices for digital preservation, LC initiated an explicit development phase to allow for research and assessment of long-term workflow and curation needs, as well as incremental progress toward NDNP goals. The principles applied in making technical choices were intended to support the development of a system that is sustainable in today's best estimation – open, modular, certain to change, and able to evolve to meet future uses.

In addition, the decisions made were informed by realities of the overall program structure:

- The content in question – analog versions of historic newspapers - resides primarily in state repositories, rather than the national library, therefore the program requires distributed production of the digital assets;
- The funding to apply new technologies to enhance access to this material is finite, therefore,
    - given the sheer quantity of available material, content included in the program will be selective, rather than the entire corpus available;

  - technical requirements for converted materials should account for potential re-use and reprocessing over time (scan once, use many times)
  - should provide a model for similar distributed efforts that may eventually interoperate – sharing best practices, conversion specifications, and standardizing basic access for historic newspapers;
- Demonstration of good use of federal funds by providing open and perpetual access;
- In expectation of change, avoid closing off options, by developing a preservation environment that would be open, expandable, and modular.

Aggregating the Content

In order to build an extendable and scalable activity, NDNP considered various requirements for production and management of the digital information created by NEH awardees. First, in order to fulfill LC's role in aggregating and managing the digitized over the long-term, LC needed to consider five main requirements:

- convert the content to achieve the highest quality information for discovery and re-use,
- be able to ensure technical consistency across content created by multiple producers over time,
- use open and sustainable formats to encourage long-term preservation,
- develop a data architecture that would allow for both manageability and scalability over time, and
- develop scalable workflows and processes that support the large-scale ingestion of content from multiple producers.

Building on its lengthy experience with large-scale digitization of historic materials, LC developed a rich set of technical specifications for content created in NDNP. The image specifications – TIFF, JPEG2000, PDF – are intended to play specific roles in the NDNP repository (TIFF for archiving, JPEG2000 for production and PDF for portability) and conform to current best practices for digital file format sustainability.[2] These practices include wide-ranging adoption in the cultural heritage community, transparency of the digital information itself, and self-documentation within the file format. The specifications for NDNP – primarily 8-bit grayscale at 400 dpi – attempt to capture the most data possible from microfilm imaging in this content type, in order to provide for reprocessing and reuse at a later date with improved technology. In addition, LC chose a standard XML metadata scheme (Metadata Encoding and Transmission Standard[3]) for description of the digital objects at the newspaper issue and page level and the ALTO (Analyzed Layout and Text Object) schema extension[4] chosen for structuring the automatically-recognized machine readable page text. Metadata requirements were intended to provide a basic level of access to newspaper pages, capturing as much structural and technical information as possible from both film and intellectual content at the point of digital creation.

NDNP recognized the distributed production model would require improved mechanisms for quality assurance of the content as it was aggregated, as well as explicit incorporation of metadata intended to assist in long-term management and sustainability of the digital objects. These requirements led to the development of NDNP-specific validation tools and workflow that can be distributed to awardees and associated production staff to ensure conformance with stated NDNP technical specifications.[5] The validation toolkit, an extension of the

JSTOR/Harvard Object Validation Environment (JHOVE)[6] and known as the NDNP Digital Viewer and Validator (DVV), provides both a tool for automatically checking the Submission Information Package (SIP) for technical conformance (e.g. whether a field is populated with the appropriate data type), as well as a viewer for subjective checking (e.g., whether the field data is correct) and to visually inspect content. Not only does the DVV provide for quality assurance of all deliverables, but at the time of validation, it also extracts header data from the various self-documenting filetypes for transformation into PREMIS and MIX schemas within the associated METS object. In addition, the DVV adds a digital signature to the METS object for each associated file, enabling a fixity check at any point in the data ingestion workflow. Figure 1 describes the high-level workflow currently in use for the initial creation and validation of the digital objects using these tools. In the next phase, NDNP plans to automate and integrate current processes as much as possible to minimize the steps involved in quality assurance and transfer of the data to a repository environment. . This will also include development of an Ingest API (Application Programming Interface) to allow a variety of submission workflow/applications to integrate with the NDNP repository.

This ingestion process, at minimum, should allow non-technical staff to batch load, verify, validate, ingest and index SIP data and integrate with the curation modules described below, for the purposes of acquiring and managing the data efficiently over time.

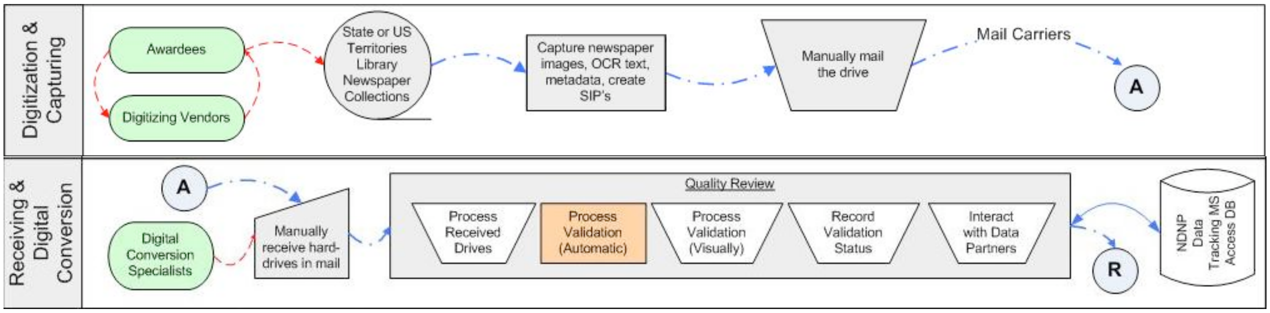
Digitization & Capturing
Awardees
Digitizing Vendors
State or US Territories Library Newspaper Collections
Capture newspaper images, OCR text, metadata, create SIP's
Manually mail the drive
Mail Carriers
A
Receiving & Digital Conversion
A
Digital Conversion Specialists
Manually receive hard-drives in mail
Quality Review
Process Received Drives
Process Validation (Automatic)
Process Validation (Visually)
Record Validation Status
Interact with Data Partners
R
NDNP Data Tracking MS Access DB

Figure 1. *NDNP data acquisition includes the tasks completed by awardees working with their respective contracted digitization vendors and the tasks completed by the Library's digital conversion specialists. The steps included represent the high-level activities completed by each party.* ***NOTE:*** *The circled letter* ***A*** *links from the step, "Manually mail the drive", in the top swim-lane to the step, "Manually receive hard-drives in mail," in the bottom swim-lane. The circled letter* ***R*** *links from the step "Quality Review" to the step "Repository" in Figure 3 (see below).*

Providing Access

After a period of exploration, LC developed requirements for an access system using user-based scenario planning and use cases to determine likely user behavior and conducted formal usability testing on an early prototype. The Web-based Graphical User Interface provides a portal to most of the information available in the digitized newspaper archive through the NDNP Service Components Architecture. It can be roughly divided into four layers:

- A client application in the Web Browser container
- A Web application server on Apache Cocoon web development framework
- Business Logic object servers in Apache Excalibur (Avalon) container
- A Repository server on various platforms:

- MySQL for object relationships and navigation
- Fedora for Archive Information Package (AIP) – metadata only
- Apache Lucene for full-text search indices.

At the highest level, users have direct access to the information of interest such as 1) searching digitized newspaper pages; 2) searching and/or alphabetically browsing newspaper title directory records. The current implementation tests the viability of the program model and will help justify the usefulness of the project, while the next phase of development will emphasize production-friendly content acquisition, stewardship curation workflows, and expansion of production capacity.

Architecturally, the public user interface functionality is built through a web-application framework accessing back-end business objects using APIs. This architecture supports additional applications and/or user interfaces through the same set of APIs that access the managed digital assets in the archive. The NDNP repository Search API is a Search and Retrieve Web Service (SRW) interface exposed over a SOAP connection. The Access API interface module provides a simple, REST-like syntax using simple-to-construct URLs to obtain content disseminations from the NDNP digital asset repository.

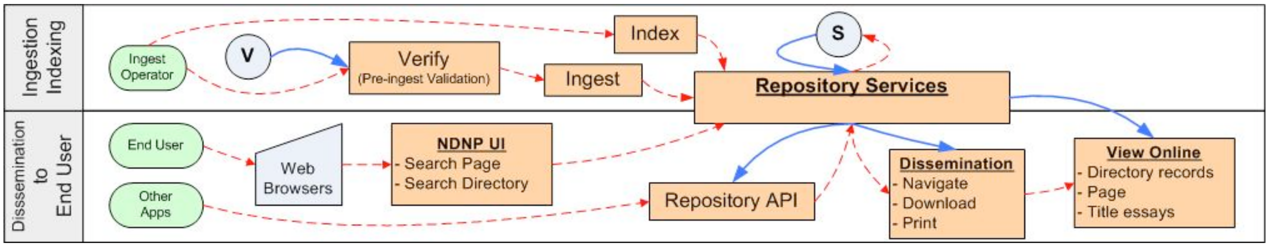

Fig.2. *The steps required to put data in the Repository and to access data from the Repository are depicted in this figure. The ingestion and indexing steps are performed by the Ingest Operator who verifies each SIP package in the Staging area. The Operator then ingests successfully-verified SIPs into the Fedora repository and creates Lucene indices. On the access side, an End User uses a Web browser to access the NDNP User Interface that utilizes NDNP's Repository API to search and retrieve metadata and digital content through the internal Repository Services.* ***NOTE:*** *The circled letter* ***V*** *links from the step "Repository" in Figure 3 (see below) to the step "Verify" in Figure 2 meaning the "Verify" step is acting on content from the storage area (i.e. the Staging area). The circled letter* ***S*** *links from the step "Repository" in Figure 3 to the step "Repository Service" in Figure 2. This indicates that data in the Repository are stored and/or accessed through "Repository Services" which consist of a set of functions managing data in and out of the data stores that make up the Repository.*

In the current state of the NDNP program, the preserved digital asset lifecycle is achieved through employing discrete system level tools and technical staff's system skills for the processes of ingesting, indexing and dissemination through repository services to the Browser Application (Fig. 2). Future development in the NDNP program may consider supporting more automated data acquisition and ingestion workflow and more stewardship (or curation)-friendly preservation management features & functions.

Sustaining the Content

An important component in the fulfillment of LC's role in this program is the development of a repository - a system environment that ensures the digital assets acquired for preservation will be preserved forever outliving people, processes, and technologies. A repository is an essential component in determining if a digital preservation environment is successful. The environment must guarantee that when people, process, and technologies change, the digital asset can be (transparently and automatically if possible) migrated from old generations to new.

The two major architecture layers in the repository are the preservation (or archive) layer and the data management layer. The lifecycle functionality implemented in this workflow addresses the needs of these two architectural layers. The key difference between the two layers is the focus on the performance. The preservation layer emphasizes more the durability or longevity of the preserved digital asset and the data management layer emphasizes more the input/output (I/O) speed, richness of functionality, and flexibility of data management. The typical three-tiered Model-View-Control architecture is used to ensure the separation of concerns for design and implementation. There are specific relationship requirements, constraints, and interfaces between these two layers of architecture to ensure the repository meets the overall digital preservation requirements.

The current implementation only allows technical staff to find, retrieve, and manage these archived assets using system level tools. But the NDNP program plan is to implement a curation manager that can provide more user-friendly functions/features to create, read, update, delete, navigate, monitor, and report the permanently preserved digital newspaper content. These capabilities may or may not be implemented as one single integrated application (see Fig. 3).

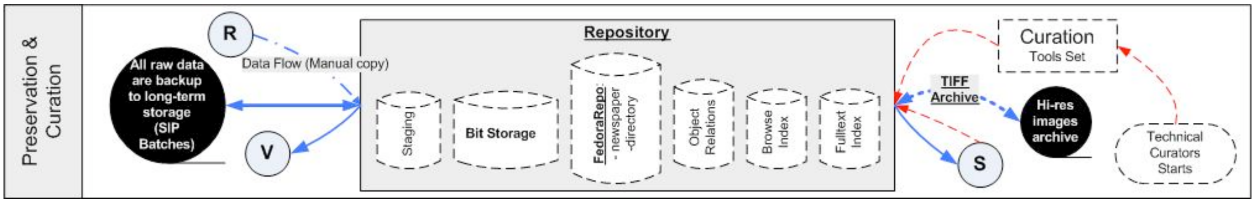

***Fig. 3**. There are basically three major types of storage area: (1) long-term preservation is stored in tapes (2) SIPs and Lucene indices supporting online access and search are stored in file systems (3) Fedora-managed metadata and object relationships are stored in MySql relational databases. Curation of all preserved content and metadata is achieved by technical staff using various system tools. **NOTE:** The circled letter **R** links from the step, "Quality Review" in Figure 1, to the step, "Repository", in Figure 3 meaning the successfully-reviewed SIPs are manually copied to the Repository. The circled letter **V** links from the step, "Repository" in Figure 3, to the step, "Verify", in Figure 2. The circled letter **S** links from the step "Repository" in Figure 3 to the step, "Repository Service" in Figure 2.*

Supporting Infrastructure for Sustainability

The NEH and LC have made a long-term commitment to the development of this program and its digital assets, including a formal agreement regarding goals of the program, cost-sharing for development and management of the program products and cooperatively guiding the program's development. In order to fulfill its role in providing permanent access to this high-value historic content, LC initiated the development of a supporting infrastructure – both programmatic and technical – to enable the long-term sustainability of the collection.

The infrastructure established at LC included a program management team, made up of stakeholders representing collections interests, digital production (conversion and acquisition), and digital preservation. These stakeholders had hands-on experience in a broad range of LC programs, including newspaper collection development, the American Memory digital historic collections, Ameritech-funded partnerships, information technology and the National Digital Information Infrastructure Preservation Program. Together, these committee members represented various management groups in the Library and successfully scoped the LC roles and deliverables that would fulfill the first phase of program development – prove the viability of the program mission by administering a successful distributed production model, build a Web interface to acquired data, and establish a preservation system to maintain and sustain the assets created.

To accomplish any of these goals it was essential that LC also establish a dedicated technical development team, representing various specialties - including preservation architecture and repository development, data modeling, software development, search analysis and UI development - and who were willing to experiment and contribute to the advancement of best practices in digital preservation. This team shared expertise (and in some cases, staff) with other LC repository efforts – electronic journals, , using and generalizing the lessons learned in initial NDNP development to extend the repository efforts to other content types. For the past 2 years, the team included 5-7 developers at any one time, as well as a technical coordinator, a Web interface developer, a systems analyst, and several quality assurance specialists. In addition, an operations coordinator and a team of 3 digital conversion specialists handled the acquisition, verification, and quality assurance of the content produced by awardees, as well as from LC's own collections. The components of these teams are flexible and will be modified over time to keep pace with production and management needs.

The development group established for NDNP is involved in not only the creation of the preservation environment to meet NDNP goals, but also the establishment of a repository development center (hardware, software, and systems) within LC for on-going research into the challenges of preserving all types of digital information. In the next phase, the technical team will address the need to add generalizable content administration and curation tools to the repository system. As mentioned earlier, these tools will allow for efficient management of the dataset without system-level manipulation, including the ability to define metadata, register content formats, ingest bulk and individual content to bit storage, support high-speed transfer options, and provide other processing support.

In conclusion, the overall goals of the NDNP provide an opportunity for testing strategies, developing expectations, and establishing mechanisms for managing and curating large quantities of digital assets over the long-term. At the same time, the immediate need to develop a working program meant up-front decisions on the best practices and strategies available that would lead to a successful activity. As the program continues to develop and expand, LC will adapt and evolve the tools and systems available for this program. Facing the challenges of building a national digital collection will inform universal understanding of needs and capabilities for the preservation of all digital information.

[1] "National Digital Newspaper Program Guidelines", http://www.neh.gov/grants/guidelines/ndnp.html, accessed 16 March 2007.

[2] "Sustainability of Digital Formats – Planning for Library of Congress Collections". http://www.digitalpreservation.gov/formats/ , accessed 12 March 2007.

[3] Metadata Encoding and Transmission Standard, http://www.loc.gov/standards/mets/, accessed 16 March 2007.

[4] Analyzed Layout and Text Object Schema, http://www.ccs-gmbh.com/alto/, accessed 16 March 2007.

[5] Justin Littman, "A Technical Approach and Distributed Model for Validation," *D-Lib Magazine*, 12:5. (May 2006): http://www.dlib.org/dlib/may06/littman/05littman.html, accessed 12 March 2007.

[6] JSTOR/Harvard Object Validation Environment. http://hul.harvard.edu/jhove/ , accessed 16 March 2007.