An OAIS-Compatible Data Repository for the National Digital Newspaper Program: Lessons Learned During Phase I

Ray L. Murray; Library of Congress; Washington, DC

Abstract

In 2004, an agreement between the Library of Congress and the National Endowment for the Humanities established the National Digital Newspaper Program (NDNP), with a goal to create a digital resource of historical American newspapers. The twenty-year initiative would build upon the success of the United States Newspaper Program (USNP) in two ways. It would provide wider access to USNP-created bibliographic data, which cataloged information on newspaper titles published in the United States from 1690 to the present. NDNP would also fund state projects to digitize, from USNP and other microfilm, their local newspapers, with an eventual aim to make tens of millions of digitized pages freely available.

With a mindset toward preservation, NDNP chose to follow the philosophy of the Open Archival Information System (OAIS) Model. An OAIS-compatible repository was envisioned that would interact with discrete groups of actors, including: producers, managers and consumers. Since each group would have different roles with respect to the repository, the repository would need to present different options to each.

During its first phase, NDNP created a working prototype, a proof-of-concept repository for implementation of ingestion, storage and dissemination processes. In order to accomplish this, the techniques used to produce, preserve and index the digital objects had to be reexamined and modified several times. Then in the move from a prototype scale to a scale that could encompass an increasing amount of pages, the initial repository assumptions were again reexamined.

As the development phase of NDNP neared completion, a wide public release of the web site was scheduled for March 2007. In the process of scaling operations toward that wide public release, aspects of the repository structure were modified again based on the experience gained in building it. This paper offers a case study of OAIS implementation, specifically how OAIS concepts were interpreted, applied and sometimes modified over the development phase of a national digitization program.

Historic American Newspapers

Newspapers chronicle American history in ways that other formats do not. Events can be seen unfolding on a day-by-day or weekly basis. Different newspapers cover events with different geographic perspectives, depending on whether their focus is regional, statewide, or on a particular county or city.

Genealogists and historians have long taken advantage of the rich material to be found in newspapers, tracing people and historical events at the most detailed level available. The Internet now holds out the potential to make historical newspapers available to an even wider audience, with access much greater than the limits imposed by microfilm surrogates or the brittle originals.

Newspaper archives become more useful as they grow in size. The odds of a researcher finding a particular event or person they seek increases proportionate to the number of pages available for them to search. As the size of an archive multiplies, search methodologies and preservation strategies are challenged by the increasing scale. This holds true whether the archive is analog or digital.

The National Digital Newspaper Program

In 2004, a joint agreement between the Library of Congress and the National Endowment for the Humanities (NEH) established the National Digital Newspaper Program (NDNP), with a goal to create a digital resource of historical American newspapers [1]. The twenty-year initiative would build upon the success of the United States Newspaper Program (USNP) in two ways. It would provide digital access to USNP-created bibliographic data, which cataloged information on newspaper titles published in the United States from 1690 to the present. NDNP would also fund state projects to digitize, from USNP and other microfilm, their local newspapers, with an eventual aim to make tens of millions of digitized pages freely available.

Six institutions were selected by NEH for two-year NDNP awards to run from 2005-2007. State programs in California, Florida, Kentucky, New York, Utah and Virginia were the first participants [2]. In the first grant year they designed workflows to create data fitting NDNP specifications and produced sample data. Then in the second year they scaled operations to produce data for the public release.

USNP had been primarily a cataloging and archival microfilming program. NDNP desired to take advantage of the information contained within USNP microfilm by migrating content to a digital environment, sustaining and providing access in a national system.

Prior to NDNP, there was no universally accepted data standard for digitized historical newspapers. Online historical newspapers produced by the public and private sectors often existed as discrete systems, their metadata structures not designed for interoperability. While the Library of Congress had a decade of digital experience with the National Digital Library and its product American Memory [3], this did not provide a definitive technical model. Much of American Memory's backend storage, access and management software reflected decisions made years before. NDNP offered an opportunity to examine repository and data structures anew.

During its first phase, NDNP created a working prototype system, a proof-of-concept repository implementing ingestion, storage and dissemination processes. In order to accomplish that, the techniques used to produce, preserve and index the digital objects had to be reexamined and modified several times. Then as the prototype neared completion, and with an eye toward incorporating larger amounts of data, the design and processes were again reexamined and modified.

NDNP Repository Design

In their previous collaborations, the Library of Congress and NEH shared a focus on long-term preservation. In the case of microfilm, the preservation timescale was measured in centuries. The goal of long-term preservation for a specific user community was a concept that carried over from USNP to NDNP [4].

That goal was gauged to align with the stated philosophy of the Open Archival Information System (OAIS) reference model [5]. An OAIS repository was envisioned for NDNP that would interact with discrete groups of actors, including: producers, managers and consumers. Each group would have different roles with respect to the repository, so the repository would need to present different options to each.

Producers – the state institutions receiving NDNP awards – were to create the data to be stored by the repository. A Submission Information Package was designed as a delivery vehicle. The challenge was to acquire and ingest the data, with verification throughout the process. Management functions were designed to ensure data stability over the long term. An Archival Information Package was created that would allow efficient internal storage, as well as all necessary access functions. Consumers - newspaper researchers - their needs were addressed with an access interface. Dissemination Information Packages were designed to distribute and display the repository objects in ways useful to researchers, taking advantage of the information within the data.

The overall repository was designed in a three-tiered way. The repository storage was seen as the foundation, with the business logic resting upon it, and the user interface on top of that. The system was intended to be modular, with the layers coupled only as closely as necessary. The user interface would not operate directly on the storage, but instead communicate through the business logic, as a mediating façade. This approach was to allow subsystems to be modified or replaced with the least risk to the rest of the system [6].

At its core, the repository used components of FEDORA software. This reflected the desire to use open-source software wherever practical. The FEDORA installation was done with minimal customization. Instead, the solutions to NDNP's specialized requirements were coded into the business logic, the layer of software interacting with the FEDORA parts.

The subsystems were made to interact in the simplest way that would achieve the desired functionality. With the choice of FEDORA, this pointed toward use of preexisting communications protocols, like WSDL, SOAP and REST [7]. XML data packages were employed as a standard way to ingest data, in a newly-designed METS data model for historical newspapers.

NDNP Data Object Model

Analog newspapers carry a relationship between titles, issues, pages and the words on those pages. When microfilmed, another informational layer is added. A newspaper can appear as a sequence of pages on a microfilm reel, certain sequences of pages represent original issues, and the sequence of issues may all share the same title. When digitizing historical newspapers, representing their characteristics requires a complex digital object.

To handle the complex links between these compound objects, NDNP developed a solution conforming to the Metadata Encoding and Transmission Standard (METS). METS is an XML document format designed to handle complex objects, and to facilitate management of objects within a repository, or between repositories [8]. The development team designed separate METS document templates for the following classes of objects: titles, issues and reels.

Metadata at the title level already exists for most NDNP titles. For over twenty years the United States Newspaper Program (USNP) sponsored creation of bibliographic records for newspapers published in the United States. This NEH-funded work included cataloging information and location of holdings, in standardized MARC format. Records for 137,000 titles along with their estimated 450,000 holdings records were incorporated into the NDNP system [9], allowing users to locate historical newspapers in all formats, digital, microfilm or the original paper. The title METS document brought together bibliographic and holdings data in a single title record, after having been transformed losslessly from MARC to MARC XML format. Titles that are digitized have additional data -- descriptive essays -- included in the title records. This new data takes the form of a Metadata Object Description Schema (MODS) object within the larger METS document [10].

The issue/edition information serves as an intermediary level of object, between page and title levels. It includes information about which pages belong to it, and to which title it belongs. The model allows for multiple editions on the same date, distinguishing one from another with an "edition order" data element. An issue present/missing indicator allows for records to be created for issues known to exist, but unavailable to digitize. This allows for retention of the collation work that often appears in the form of an "issue missing" frame on microfilm.

The page is the fundamental unit, the atom of the structural metadata. Metadata must exist down to the page level, to be able to associate and order files of pages within an issue. The page represents the smallest natural object to track with full structural metadata. The two dimensional layout of the page carries editorial information about the relative importance of the items on the page, and best replicates the way the page was perceived by its original readers. A sub-page-level metadata system is possible, analogous to a physical page carved up into clippings. However, each data item would lose the contextual information carried in the original two-dimensional order of presentation.

Page-level metadata was defined robustly enough to allow recording of information for missing pages, pages of the same issue digitized from different holdings and ability to keep original order on unnumbered pages and pages in multi-section newspapers. For simplicity, individual page information was rolled up into the parent issue/edition document.

Finally, the microfilm reels, from which most pages were scanned, were not ignored in the NDNP METS structure. Digitizing from microfilm is an efficient way to capture a high volume of data. Although in the end what is created is a digital image of the original page, the characteristics of the intermediary medium of the film should not be ignored. The content will have been transferred three times: once to the microfilm, once to the print negative and once when being digitized. Administrative metadata can help trace effects of that process on the final product. The reel document will capture metadata on whether the paper was filmed from loose leaves or bound volumes, the camera's effective reduction ratio, resolution quality of the film and photographic emulsion density. This will support future study of whether these characteristics impact the quality of the end product, especially OCR accuracy.

NDNP Data Validation

Since the data was intended to be stored permanently, it was considered important to know the technical characteristics of the data, and to identify and correct data errors prior to ingestion [11]. Many of the objective criteria in the data requirements can be measured in an automated way. NDNP sought to validate data with respect to the specifications as much as was practical prior to ingestion. Programmers built upon the work that had been done in the JHOVE (JSTOR/Harvard Object Validation Environment) project, the modular open-source software written to identify, characterize and validate files in various formats [12]. It can, for example identify that a file is a TIFF file, and that it conforms to the TIFF 5.0 specification. Several file formats with existing JHOVE modules were of interest to NDNP, the TIFF, JPEG 2000 and PDF modules. Code was written to bundle these modules and to extend the JHOVE code, adding specific validation rules. For example, the TIFF specification was extended to check not only the well-formedness of the TIFF file, but also that it is uncompressed, that it is 8-bit grayscale, and that it contains the microfilm reel number in tag 269.

Additional code checked characteristics of the METS files for newspaper issues, and microfilm reels, as well as the overall batch METS file that served as the manifest for the batch.

In all, over a hundred technical characteristics of the data SIP were incorporated into the NDNP validation process.

Changes upon Implementation

While the repository, data object model and validation strategy were all built using standards and best practices, each had to change as they were implemented. A single example of typical conflict and resultant change for each of these three areas may give some sense of the technical evolution within the NDNP program.

Validation Change

Even after success in ingesting data according to the NDNP model, unusual cases revealed flaws in validation that required correction. One example concerned the dates in the metadata associated with the original publication of the newspaper.

The metadata design from the outset allowed for the fact that the date printed on a newspaper can be wrong. The typesetter may not have correctly advanced the date on the masthead, or advanced it to February 29 in a non-leap year, to give some common cases. A researcher looking at a page labeled January 1, 1904 (actually published on January 1, 1905) might be confused as to why the paper reported news from a year in the future. It was seen as important to allow logical consistency within the system while making allowance for this common typographic error.

The NDNP information object was designed with two dates associated with a given newspaper: the date of issue, and the date as printed upon it. This model would allow a researcher to find information about the page with or without a date discrepancy, whether or not the researcher knew about the date discrepancy. The NDNP data object was also correctly designed to allow this case, with two separate and distinct XML elements for data incorporated into the issue XML.

Further, the NDNP validation code correctly approved newspaper issues where the date of issue and date printed on it were different. However, the validation failed those issues when they were presented in the SIP structure. The reason was that the validation software compared component parts described in the overall manifest file of the SIP, to the component files delivered as part of the SIP. In that step, the code was comparing the true date of publication (in the SIP's manifest XML file) with the date as printed (in the issue XML file). When they differed, the validation rejected the SIP, even though it had been correctly encoded.

Once known, it was simple to fix in the validation code. Special work-arounds for data already in the pipeline required some effort. Testing and distributing the new validation code was a non-trivial effort.

A dozen discoveries analogous to this one created the need for new versions of the validation software over the course of the development phase. Those instances demonstrated the gaps between theoretical ideas of how data should appear and the actual data delivered, particularly when producers encounter unusual cases in the source material.

Data Object Change

USNP standards encouraged collation before filming in order to create the best, most complete run of a newspaper for its filming. Despite best efforts, not every copy of every paper was available or could be located at the time of microfilming. This was seen as a case where NDNP could include corrective measures. When gaps were found in the microfilm coverage, paper copies could be scanned to fill in.

The METS page objects did not initially contain a way to distinguish between pages that had been scanned from paper and those scanned from microfilm. The first set of NDNP data was created using newspapers from the District of Columbia, which had only been microfilmed by the Library of Congress. For these pages, all scanning was done from microfilm, none from originals. Because of this, the lack of ability to distinguish source types went unnoticed longer than if the initial test data set was more heterogeneous in all its characteristics.

The METS template for issues [13], containing the page objects, was modified to add an element describing the source format for the scanned materials, paper or microfilm. The change was conveyed to all users of the templates, the NDNP awardee institutions and through them to their vendors.

There was a cascade effect to this seemingly small change; the change to the object model necessitated a change in validation. The issue METS object also contains a microfilm reel number (an unique number assigned by the Library of Congress as part of the NDNP scanning process) and reel sequence number. The validation software required both a reel number and reel sequence number to be associated with each page. As NDNP awardee institutions scanned pages from paper and attempted to create metadata for them, they realized the reel data elements made no sense in that case. They considered whether to use the microfilm reel number the page might have been on, if it had been filmed with the pages its parent title. However, a speculative approach would not add substantively to the preservation goals of the metadata.

The validation code was modified to allow a reasonable encoding of scans from paper. The new code required no reel number or reel sequence number for paper scans, but continued to require both if the image was scanned from microfilm. This required a new release of the validation software, and specific reeducation of interested awardees.

Repository Change

Even the data structure internal to the repository, the archival information package, was not immune to change. Through most of the development phase, the ingestion of the data pulled all parts of the data – image files, OCR and metadata – into FEDORA. This worked as a way to store, index and access pages. A prototype was built that successfully accessed and displayed this data. However, in that approach, the integrity of the data once in the repository was not as well known as it was before it went in. Having been broken up into different data objects and data streams, the data had been transformed from its previous arrangement in the SIP.

The validation software had been written to evaluate a SIP, its components in relation to each other, before ingestion. It was not designed to run on the transformed data inside FEDORA. Additionally, moving the data to the repository required a rewrite of each of the image files, with every rewrite an opportunity for file corruption. Even if the validation code had been modified to operate on individual files, validation inside and outside of the repository would always be akin to comparing apples and oranges.

The ingestion was also slow, taking about a day to process a typical SIP, which represented 5000 newspaper pages. When the amount of data to be ingested in phase one was calculated, greater speed would be advantageous.

Programmers experimented with a different approach to storing the data internally. The new model ingested the metadata for a newspaper issue into FEDORA, as before, but not the OCR text or images. Instead, the files would be represented in FEDORA by pointers to the files. The files themselves would exist on a Library of Congress server outside the control of FEDORA. A byproduct of the new concept that was seen as advantageous was that the files could remain on the server in their SIP format. It would then be a straightforward matter to revalidate them in the future, to ensure no loss of data over time.

Testing proved the new method of data storage could work. Ingestion was accelerated by an order of magnitude. The decision was made to shift to the new approach, though it was not without cost. New ingestion code had to be written. All the data previously ingested had to be purged and reingested. However, with the faster ingestion, this was a one-time tradeoff to provide greater speed in the future.

The new approach to ingestion also provoked discussion about an important aspect of handling data over the long term. With the data outside of FEDORA, its care, maintenance and migration would have to be explicitly defined. It could not be assumed that FEDORA would automatically take care of these stewardship issues. In reality that was always the case, but the focused attention on this particular change proved valuable in making that apparent.

Conclusion

The NDNP experience creating a digital repository for historical newspapers has underscored the importance of upfront planning and regular replanning based on accumulated experience. While the OAIS reference model has proved useful in clarifying goals and requirements for the NDNP repository, it has been no simple matter implementing the model. No off-the-shelf solutions or combination of off-the-shelf solutions exist that do not also require sharp attention to data stewardship. While developing a repository system may be thought of as a project with start and end dates, maintaining a trusted repository into the future must be seen as an ongoing evolutionary process.

References

[1] B. Cole, "The National Digital Newspaper Program." Organization of American Historians Newsletter 32 (2004).

[2] H. Aguera, "The National Digital Newspaper Program: Thinking Ahead, Designing Now." Proc. IFLA 122, p. 80. (2006).

[3] "American Memory, Mission and History." http://memory.loc.gov/ammem/about/index.html

[4] M. Sweeney, "The National Digital Newspaper Program: Thinking Ahead, Designing Now." Proc. IFLA 122, p. 83. (2006).

[5] RLG, "Trusted Digital Repositories: Attributes and Responsibilities." http://www.rlg.org/en/pdfs/repositories.pdf

[6] G. Schlukbier, "The National Digital Newspaper Program: Thinking Ahead, Designing Now." Proc. IFLA 122, p. 95. (2006).

[7] Fedora Project, "Overview: The Fedora Digital Object Model." http://www.fedora.info/download/2.2/userdocs/digitalobjects/objectModel.html

[8] Metadata Encoding and Transmission Standard, Official Web Site. http://www.loc.gov/standards/mets/.

[9] "Chronicling America: Historic American Newspapers." http://www.loc.gov/chroniclingamerica/.

[10] Metadata Object Description Schema, Official Web Site. http://www.loc.gov/standards/mods/.

[11] J. Littman, "A Technical Approach and Distributed Model for Validation of Digital Objects," D-Lib Magazine, 12, 5, (2006).

[12] JSTOR/Harvard Object Validation Environment, http://hul.harvard.edu/jhove.

[13] NDNP Technical Specifications. http://www.loc.gov/ndnp/techspecs.xml.

Author Biography

While on staff at the University of Arizona, Ray Murray worked to create optics for giant telescopes, including casting the second 8.4-meter diameter primary mirror for the Large Binocular Telescope on Mt. Graham near Safford, AZ. After coming to the Library of Congress, he led the first newspaper digitization project for the American Memory web site. He is currently involved in the technical development and digital conversion activities of NDNP, and its web site, "Chronicling America: Historic American Newspapers."