Title:
Bulking up: how accepted standards and evolving technology advance research in Chronicling America

Authors:
Nathan Yarasavage, Deborah Thomas, Georgia Higley
Library of Congress, Washington, D.C., USA

Abstract:
The National Digital Newspaper Program (NDNP), a partnership between the National Endowment of the Humanities, the Library of Congress, and cultural heritage institutions across the United States, has been enhancing access to America's historic newspapers since 2005. As NDNP's open-access web site, Chronicling America, grows closer to reaching a corpus of 10 million digitized newspapers, it is worth reflecting on the decisions made nearly a decade ago that shaped this collaborative endeavor. While enhancing access to newspapers is its stated goal, at its core, NDNP is a preservation *and* access project, using widely-adopted, well-documented, existing standards for imaging, metadata, and web delivery. The use of accepted standards and availability of burgeoning content are now creating opportunities for scholars and technologists alike to imagine and experiment with data mining, patterning, and statistical analysis.

The technical approach for NDNP digital content creation, for the most part, has remained un-changed over the last eight years. Meanwhile, the technologies underpinning Chronicling America continue to evolve allowing for new and unexpected uses of the digitized newspaper images, metadata, and text. In addition to providing a reliable technical infrastructure to support traditional researcher use via a web browser, Chronicling America makes available a fully documented unrestricted Application Programming Interface (API), employs linked data concepts, and provides access to bulk OCR data downloads to assist in new and emerging digital humanities research on large or targeted portions of the collection.

While the NDNP technical specification for content delivery has been widely documented, this paper will discuss the rationale behind the technical strategies, practices, and infrastructure NDNP has established supporting sustainable access and the sometimes un-intended research and usage trends that result. It will also demonstrate how careful choices made early in the project have resulted in enhanced content access as well as increasingly sophisticated use by scholars.

Introduction
With its first awards issued in 2005, the National Digital Newspaper Program (NDNP) partnership continues to provide access to historical U.S. newspapers published between 1836 and 1922[1]. As of Jan 30, 2014, NDNP has awarded 36 institutions with 33

[1] http://www.loc.gov/ndnp/

having already submitted content. This amounts to approximately 280 terabytes of data received by LC. This number does not include an additional copy of archival data on tape storage, and an additional access copy (without TIFF images) on spinning disk. With over two dozen awardees actively sending content every month, LC receives approximately 135,000 pages worth of digital newspaper content representing approximately 7 Tb a month. Currently 6.7 million newspaper pages are available via Chronicling America: Historic American Newspapers, the web site hosted by LC that provides access to NDNP content.

Pillars of NDNP: Built for Sustainability

At the foundation of NDNP is the stated goal of the program, enhancing access to newspapers. Supporting this foundation are several core principles that shape the technical infrastructure of the program. These pillars ensure the program makes the most out of the finite public funding available and invested in NDNP while reaching the widest audience possible[2].

These pillars of sustainability include:

- a shared technical specification with widely accepted and adopted standards,
- institutional cooperation between all project partners,
- validation of data and verification of data integrity over time,
- a robust data management plan, and
- the expectation that change is inevitable.

NDNP Technical Guidelines 2005-present

During the planning phases of NDNP, it was clear that in order to sustain a program of scale, a shared technical specification with widely accepted and adopted standards was necessary. The NDNP Technical Guidelines for Applicants document (commonly referred to as the NDNP technical specification) has been issued every year to coincide with NEH's two-year award cycles. Since 2005, the technical specification followed by NDNP has been well-documented and widely promoted in national and international venues[3]. Over time, the requirements have remained stable without significant changes. Updates have been mostly limited to clarifications about existing practice and guidance to accommodate expanding content scope, such as the introduction of other languages or the inclusion of images scanned from paper. Below is a brief discussion of the rational

[2] For an in-depth account on the sustainability plan behind NDNP, see Thomas, Deborah and Mark Sweeney. "Sustainability in the United States National Digital Newspaper Program." (IFLA International Preservation News, No. 56, May 2012 p.12-20), http://www.ifla.org/files/pac/ipn/IPN%2056.indd.def.pdf.
[3] See a list of presentation and publications at http://www.loc.gov/ndnp/guidelines/pubs.html.

behind key pieces of the guidelines and how NDNP has adapted the requirements since the initial 2005 award phase[4].

NDNP Awardee Roles

To permit a distributed digital content creation program at the scale of NDNP, it was decided that the awardee institutions would bear the cost of creating the entire NDNP digital content package (images and metadata), and that they would also be responsible for selecting content, evaluating microfilm, assigning metadata, and writing descriptive newspaper histories for each title. The NDNP specification, sometimes referred to as "richly detailed," places these efforts on the awardee institutions, where both the resources and expert knowledge of the content are most available - rather than the Library of Congress. Awardees deliver data produced to the NDNP specification monthly in a unit referred to as a batch. When a batch arrives at LC, it contains everything it needs to be "bagged" using the BagIt specification[5], copied, and ingested into the Chronicling America web site.

File Formats

NDNP requires its awardees to submit an 8-bit grayscale, 300-400 dpi TIFF images as the preservation or master image. Derived from this TIFF, awardees also provide a visually lossless 8:1 compressed JPEG 2000 image and a PDF. NDNP recognizes that some institutions and digital content creators decide to discard TIFF images or adopt policies assigning the JPEG 2000 as the preservation or master image. To satisfy the preservation nature of the program, the TIFF is considered the trusted preservation format for NDNP. Due to its proprietary nature and lack of widely-adopted tool support, the JPEG 2000 is not considered at this time to be an archival substitute for the TIFF for NDNP[6]. While the current Chronicling America web site does not use the TIFF image for display, it does require and use the JPEG 2000 to dynamically create image tiles that are sent to the browser for viewing. All three image file types required for NDNP have a published file format profile to provide guidance on embedded metadata. While developed before the Federal Agencies Digitization Guidelines Initiative (FADGI) recommendations, NDNP closely aligns its specification with those set forth during that initiative[7].

Metadata

NDNP's metadata specification includes descriptive, administrative, and structural metadata to support preservation and access of the program's digital assets. The

[4] The full NDNP technical specification is available at http://www.loc.gov/ndnp/guidelines/.
[5] BagIt is a hierarchical file packaging format for the exchange of generalized digital content. See https://wiki.ucop.edu/display/Curation/BagIt.
[6] For more discussion on JPEG 2000 as a preservation format, see http://blogs.loc.gov/digitalpreservation/2013/01/is-jpeg-2000-a-preservation-risk/.
[7] See http://www.digitizationguidelines.gov/.

metadata approach relies on pre-existing standards that have been widely adopted throughout the digital library community. The metadata primarily relies on the following:

- METS XML schema for structural metadata[8],
- ALTO XML schema for the Optical Character Recognition (OCR) information and corresponding coordinate mapping (hit highlights)[9],
- PREMIS[10] and MIX[11] for technical metadata, and
- MARC and MODS[12] for descriptive metadata.

METS ALTO

Producing OCR files to the ALTO specification for every newspaper page is a key requirement of NDNP awardees. As is common in many ALTO implementations, NDNP wraps ALTO information within METS. The NDNP ALTO specification requires, at minimum, column level text block zoning and appropriate coordinates to map or highlight text to image files. Though the granularity of the ALTO schema allows for finer detail of zoned content blocks including articles, illustrations, headlines, etc., NDNP supports page-level access. It was decided in the early stages of NDNP, that in order to achieve efficiency in a project of this scale, article level access that is common in many digital newspaper delivery platforms, would be more costly and result in less content being digitized. However, this would not prevent awardees from locally implementing article level access. Chronicling America's current and previous interface features page-level keyword access with visual representation of search results. When users conduct searches on the collection, search term highlights cluster on a page image to form a visual cue of where content of interest resides. With the ability to zoom, pan, and crop portions of a page image, this approach arguably achieves a satisfactory result close to or equivalent to that of an article segmented system but without the extra over-head involved.

Descriptive Metadata

Descriptive metadata for each NDNP object is based largely on title-level MARC records, many created during the United States Newspaper Program (USNP)[13] to the CONSER-level standard for cataloging newspapers[14]. By using a pre-existing cataloging standard, participants in NDNP have an authoritative source for consistently describing titles selected for digitization. NDNP has developed a detailed metadata dictionary that

[8] Metadata Encoding and Transmission Schema: http://www.loc.gov/standards/mets/
[9] Analyzed Layout and Text Object (ALTO): http://www.loc.gov/standards/alto/
[10] Preservation Metadata Maintenance Activity: http://www.loc.gov/standards/premis/
[11] Metadata for Images in XML: http://www.loc.gov/standards/mix/
[12] Metadata Object Description Schema: http://www.loc.gov/standards/mods/
[13] United States Newspaper Program http://www.neh.gov/us-newspaper-program
[14] Cooperative Online Serials Program of the Program for Cooperative Cataloging: http://www.loc.gov/aba/pcc/conser/

directly maps specific existing MARC fields to required and optional metadata components in the specification. Library of Congress Control Numbers (LCCNs) are required for NDNP and serve as the unique identifier for titles chosen for digitization as well as for the over 150,000 records in the United States Newspaper Directory (described below). ISSNs do not exist for many historical newspaper titles, and thus are not used as a unique identifier for NDNP. When titles are digitized and made available in Chronicling America, ISSN numbers, relevant notes, and hyperlinks to the digital format are added to the MARC records in OCLC WorldCat. The adoption of the Resource Description and Access (RDA) bibliographic rules that supersede the AACR2 rules by LC in 2013 has little impact on NDNP's use of MARC records at this time. Specific rules in the 2014-2016 NDNP technical guidelines have been updated to provide guidance on how to interpret RDA MARC records as they relate to NDNP.

Other page and issue-level descriptive metadata is captured by awardees when collating or evaluating microfilm for digitization, as needed to enhance content navigation. Awardees (as curators) evaluate each title as they prepare material for digitization to determine when and if such additional metadata should be captured. This includes page numbers, section labels (Sports Section, Culture, Local, etc.) and edition labels (Daily, Final Edition, Weekend, etc.). Over time, as a result of awardee communication, it was determined that these descriptive section and edition labels, when present on newspaper pages were often inconsistent and required additional labor to identify and encode in an efficient manner. Rules about capturing this information were also clarified in the 2011-2013 guidelines. Modifications such as these enable a wider variety of state partners to reach success in NDNP within their existing digital project infrastructures without sacrificing core preservation elements.

Administrative Metadata

Administrative metadata provides information necessary to manage the NDNP digital objects over time and encompasses both technical and preservation metadata. Technical metadata relating to the microfilm used for digitization (e.g. reduction ratio, density, etc.) is captured by most awardees during collation and film evaluation activities. While originally this technical metadata was required for titles scanned from microfilm originals, in the 2011-2013 guidelines, a slight change to the specification changed this from required to optional. Because the level of detail and effort involved in capturing this metadata varies across the program, technical metadata for microfilm is now encouraged, as a matter of good practice to get the best digitization possible, but not required.

Validation and Verification

A critical piece of administrative metadata required for NDNP is known as a digital signature. Digital signatures are an important piece of the NDNP sustainability plan, validation and verification of data. Having multiple content creators using different

commercial vendors, it was necessary to incorporate a uniform method of ensuring data integrity not only within a given batch, but across cycles of the program. Data created in 2005 must be compatible with data produced at the end of the program. Created and inserted into the METS metadata files by the NDNP Digital Viewer and Validator tool, digital signatures are one way for LC to ensure data integrity over time. Once created, vendors, awardees, and LC can check and recheck the digital signatures for accuracy. Though the Digital Viewer and Validator software has evolved over time since 2005, its core function remains the same today and is incorporated within the LC repository services workflow for regularly scheduled verification of NDNP data. Batches created in the first award of NDNP in 2005 can be verified alongside batches arriving today in the most recent NDNP award cycle[15].

In that vein, LC has successfully migrated the entire NDNP data set twice across storage systems since the program's beginnings. Each time the need to migrate was driven by changes or upgrades to the hardware infrastructure supporting the Library's collections. Running verification (checking the digital signatures) after every batch is copied to a new storage location is a way of verifying that the data has successfully migrated each time.

Chronam

Keeping with the goal of the National Digital Newspaper Program of improving access to U.S. newspapers, the Chronicling America web site has been architected to promote free and open access. Along these lines, the open source, Django-based software application running Chronicling America, coined chronam, is now developed and made available in Github, a popular code repository site focused on open source software development projects[16]. By providing access to the chronam software in this way, LC enables NDNP awardees and others interested in the application to install and have available a core set of functionality for loading, modeling and indexing NDNP data, while also allowing users to customize web templates for presentation as needed. An unrestricted listserv, CHRONAM-USERS, was launched by LC in 2013 in hopes of encouraging dialog related to chronam development.

Data available in Chronicling America is copyright-free being held in the public domain and is therefore not restricted in any way. But beyond that, using common web protocols, several views of the data are made available through an extensive and well-documented Application Programming Interface (API)[17] to support use in alternative interfaces or data harvesting. Furthermore, in 2013, NDNP introduced bulk access to its OCR files. Upon ingest, every batch in Chronicling America runs through a process to extract all of its OCR in both text and ALTO xml formats. This is then compressed

[15] See http://www.dlib.org/dlib/may06/littman/05littman.html for more information on NDNP validation of digital objects.
[16] Chronam software is available at https://github.com/LibraryofCongress/chronam.
[17] Chronicling America's API is documented at http://www.chroniclingamerica.loc.gov/about/api/.

into .tar files for manual or programmatic download[18]. Adding the new bulk downloads of OCR was largely in response to the increasing amount of requests received at LC from researchers and commercial service providers who would benefit from direct access to the entire corpus of text rather than using the Chronicling America API as an external dependency[19].

United States Newspaper Directory

In addition to providing access to the digitized content created by NDNP, Chronicling America provides access to a database of bibliographic titles for newspapers published in the United States 1690-present based on data held in the OCLC WorldCat database, supported by the world's libraries. Referred to as the United States Newspaper Directory, these records form the structural backbone to newspaper access in Chronicling America. Per the program's guidelines, all newspapers digitized through NDNP must be represented in the Directory prior to ingest into Chronicling America. This approach ensures the content included in the collection adheres to formal rules of identification and description that allows for a shared understanding (beyond the site itself) of complex newspapers histories and bibliographic representation. Utilizing this authoritative and communal source for serial documentation also provides standardized set of metadata that can be neatly applied to all possible Chronicling America content and used for reliable search and browsing functions (e.g., subject terms, geographic designations and chronological information).

Though the concept and structure of the database has not changed much throughout the life of NDNP, the methods of updating this core data has evolved in the last few years, improving accuracy and currency. LC utilizes the OCLC WorldCat OpenSearch API to pull bibliographic records of newspapers published in the United States and its territories. Using this API, LC pulls new records at regular intervals for inclusion in the directory. Searching the directory is not only available via the web interface through a search form, but as documented through the Chronicling America API, searching is also possible using the OpenSearch protocol, accommodating both Atom and JSON (JavaScript Object Notation) response formats[20]. A comma separated value (.csv) format is also now available for queried search results of the newspaper directory. This .csv format enhancement assists researchers who may prefer to use a common spreadsheet for analyzing data retrieved en masse from the newspaper directory.

[18] In addition to HTML reports of the OCR bulk download files, alternate versions are available in Atom and JSON formats.
[19] See http://chroniclingamerica.loc.gov/ocr/.
[20] See http://chroniclingamerica.loc.gov/about/api/.

Advancing Research with Chronam

The overall reliance on existing well-adopted standards approach followed in the NDNP technical specification for content delivery coupled with Chronicling America's well-documented specifications and unrestricted API and bulk data has led to new and interesting research endeavors in the past several years. Following is a summary of some of these projects.

In 2011, the Rural West initiative out of the Bill Lane Center for the American West at Stanford University created a remarkable data visualization, The Growth of Newspapers Across the U.S.: 1690-2011. This visualization was included along side the center's report on Rural Newspapers in the West[21]. So how did they do it? They harvested the title records from the U.S. Newspaper Directory of Chronicling America and plotted them by time and location, standardized values that had been developed according to rigorous cataloging requirements. What results is a beautiful dynamic map showing the spread of newspaper titles from east to west across the country and the origins of immigrant press.

The Bill Lane Center for the American West also collaborated with the University of North Texas on a project called Mapping Texts which developed two other interactive visualizations using data from Chronicling America[22]. These visualizations allowed users to dynamically asses the quality and language patterns in the digitized text of 232,500 pages of Texas newspapers. Combining text-mining with geospatial mapping approaches to research, this project opens new doors to what can be asked of and done with a substantial collection of digitized historical text.

Similarly, as part of the Digging into Data Challenge[23], "An Epidemiology of Information: Data Mining the 1918 Influenza Pandemic[24]" used text mining methods, specifically topic modeling with segmentation and tone classification, on data in Chronicling America and other sources "to understand how newspapers shaped public opinion during the 1918 influenza pandemic." The project used the raw OCR data from select weekly newspapers in Chronicling America. As reported by the project team, the methods used in their project have potential to reshape historical analysis, while recognizing that additional effort is needed to increase accuracy of the text[25]. The substantial corpus of newspapers available through Chronicling America, diverse in both geographic coverage and political orientation, for this time period allows for extensive analysis of this singular event.

---

[21] http://www.stanford.edu/group/ruralwest/cgi-bin/drupal/visualizations/us_newspapers
[22] http://mappingtexts.org/
[23] http://www.diggingintodata.org/
[24] http://www.flu1918.lib.vt.edu/
[25] http://www.historians.org/publications-and-directories/perspectives-on-history/january-2014/mining-coverage-of-the-flu-big-data%E2%80%99s-insights-into-an-epidemic

Another project, "Infectious Texts, Viral Networks in 19th-Century Newspapers" uses advanced computational methods to analyze large amounts of OCR text from Chronicling America[26]. This project "seeks to develop theoretical models that will help scholars better understand what qualities—both textual and thematic—helped particular news stories, short fiction, and poetry 'go viral' in nineteenth-century newspapers and magazines." Making use of the bulk OCR files made available on Chronicling America opens up the possibility for researchers to ask these new types of questions.

Most of these projects are ongoing and are at the tip of the iceberg in revealing to us as curators and collection custodians what is possible when historical documents take on new life in the form of open "big data," and made widely available through common and standard protocols. Each of these research topics is uniquely served by the key aspects of the program – curatorially-selected content (representing geographic, chronological and audience diversity), standardized digital objects, and lots and lots of potential research targets (i.e. pages.) LC has encouraged researchers to communicate directly with its software developers and collection managers on their endeavors which has proven very fruitful for both LC and the researchers themselves.

Use of bulk data from Chronicling America and other newspaper repositories is in its infancy, and the growing pains of new research methodologies highlight both the research potential and problems of large data sets. Although text mining and data visualization programs have existed for years, only recently have they been applied to large runs of newspapers. Repositories like Chronicling America provide a solution to the earlier problem of accessing historical newspaper content; now attention can turn to developing new ways to take advantage of the raw data behind it. Analysis of nineteenth century newspapers, rich in poetry, text and images of daily life, and replete with the language of local communities, has the potential to rewrite social and literary history. As Franco Moretti said, "What would happen if literary historians, too, decided to 'shift their gaze'…'from the extraordinary to the everyday, from exceptional events to the large mass of facts'? What literature would we find in 'the large mass of facts'?". [27]

Perhaps the greatest deterrents to facile use of NDNP data (and digital newspapers in general) are twofold: the computing resources necessary for the analysis of large buckets of data (such as the entire corpus of Chronicling America text) and the challenges of historic newspaper format, available tools, and uncorrected text recognition that results in lower than optimal character accuracy[28]. Projects highlighted in this paper prove that access to bulk data has significant research value, but resources to do so are often scarce and may require additional development to make it a routine

[26] http://www.viraltexts.org/

[27] Franco Moretti, *Graphs, Maps, Trees: Abstracts Models for a Literary History* (London: Verso, 2005), 3.

[28] Elizabeth Lorang and Brian Pytlik Zillig, "Electronic Text Analysis and Nineteenth-Century Newspapers: TokenX and the Richmond *Daily Dispatch*," *Texas Studies in Literature and Language*, 54(3), Fall 2012, 307.

research technique. However, the necessary tools have progressed quickly and over time the costs associated with this type of research will likely improve.

Weathering the Storm: Peaks in Web Traffic

In November of 2013, a truly devastating typhoon hit the islands of the Philippines leaving mass destruction in its wake. Days later Chronicling America saw its highest web traffic to date, averaging six times the normal usage. Given the historic nature of content in Chronicling America, a connection was not immediately apparent, but after some analysis, it was clear that an article with the headline "15,000 Die in Philippine Storm" published in Washington DC on November 30, 1912 was going viral[29]. Traffic sources were not the usual, however. Traffic was higher than average for several days, but on the busiest day of the traffic surge, over 20% of the entire site traffic was originating from Facebook links (this percentage is normally in the single digits), and approximately 40% of the entire site visits came from users in the Philippines. It is apparent that several news outlets in the Philippines picked up the story and as they say, the rest is history[30].

Over the past year, LC has also observed increased exposure to Chronicling America content in social media, blogs, and popular news outlets. The Philippines typhoon story, a comedic huffingtonpost.com blog post[31] on "vintage slang terms", and a fascinating feature article in the *New York Times*[32] on the etymology of the phrase, "the whole nine yards", are just a few of the more recent causes of high use spikes.

What is noteworthy is that due to the Library of Congress' recent bulking up of the technical infrastructure running Chronicling America, the site survived these sudden peaks of traffic with no noticeable effects on individual user experience or harvesting/API activities. Specifically, implementation of a web application accelerator, known as Varnish Cache[33] has ensured that web site performance remained stable throughout and suffered no negative effects of these surges. Through this approach, recently accessed images and web pages get added to a cache, and this drastically improves Chronicling America performance.

---

[29] http://chroniclingamerica.loc.gov/lccn/sn83045433/1912-11-30/ed-1/seq-1/

[30] More on this viral story can be found at: http://blogs.loc.gov/loc/2013/11/youve-heard-the-phrase-100-year-storm/.

[31] http://www.huffingtonpost.com/2013/11/13/vintage-slang-terms-drunk_n_4268480.html?utm_hp_ref=college&ir=College)

[32] http://www.nytimes.com/2012/12/27/books/the-whole-nine-yards-seeking-a-phrases-origin.html?_r=1&

[33] https://www.varnish-cache.org/

Conclusion

NEH and LC continue to support NDNP with the aim of expanding content to include new titles from new states and territories, new languages, and additional content from existing titles. The standardization of content (selection criteria, formats, metadata elements) has created a curated and uniform dataset that can be sustained and utilized well into the future. Even so, the NDNP technical guidelines for applicants, while stable in their current form, are evaluated every year and updated as needed. As the NDNP collection grows, Chronicling America's underpinning technology will also continue to evolve, not only with the goal of enhancing the traditional web site user's access to historical U.S. newspapers but also with the objective of supporting existing and new research methodologies that can be applied to millions of pages of text.