https://doi.org/10.2352/issn.2168-3204.2017.1.0.152

The Evolution of the US National Archives Catalog: From Access to Engagement

Michael Horsley, United States National Archives and Records Administration, College Park, Maryland, US

Abstract

Improvements to United States National Archives and Records Administration's (NARA) on-line catalog enable delivery of high-resolution images to the public. New transcription and tagging tools in the catalog allow citizen archivists to engage with digitized records and increase access to archival material. NARA has adopted a multitude of social media platforms that have connected millions of patrons to digital content. In an ever-expanding online world, the public expects to find everything on the Internet. With holdings in the billions of archival records of every conceivable format, NARA has had to embrace an adaptive and scalable digitization approach. NARA has a long history of developing digitization guidelines that have proved invaluable to ensure image quality, meet mandates for preservation and access, and create useful master and derivative digital objects. This paper analyzes the parallel evolution of NARA's digitization standards and how they are applied in online catalogs and social media platforms.

Beginning of Standards and On-line access

In 1994 NARA began a pilot project to deliver information about the agency through "CLIO" a "gopher protocol" that predated the World Wide Web. With the advent of the Web, NARA embarked on the 1998 Electronic Access Project (EAP) that digitized over 100,000 core holdings and linked them to archival descriptions from the "NARA Archival Information Locator Database" (NAIL). [1] The project published the

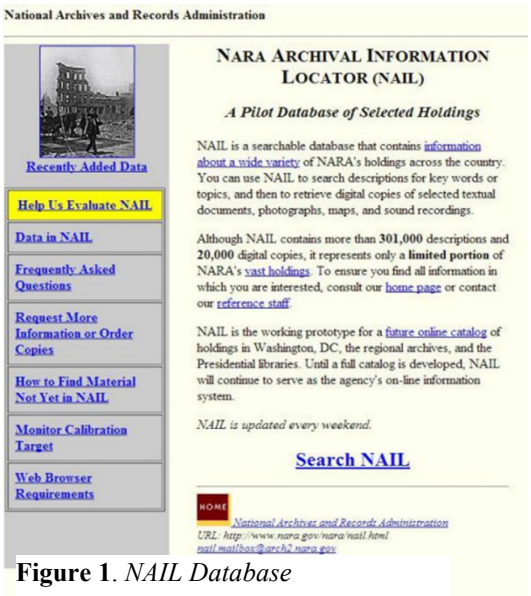

Figure 1. *NAIL Database*

"*Guidelines for Digitizing Archival Materials for Electronic Access"* that defined specifications for Master, Access, and Thumbnail files, and provided Quality Assurance (QA) criteria to judge image quality. [2]

For example the specifications for textual records selected 300 dpi as the scanning resolution "...to be compatible with OCR software", and 3000 pixels long for photographic materials that would be suitable for reproduction quality. The Guidelines contain capture specifications that delineate resolution, file format, scanning aim point values, tonal range, tonal reproduction control, and scanner and monitor calibration. The intent was to provide guidance to ensure proper capture and display. Included in the Guidelines was a user-friendly matrix that summarized digital file specifications broken down by media source for all the deliverables to the project. [3]

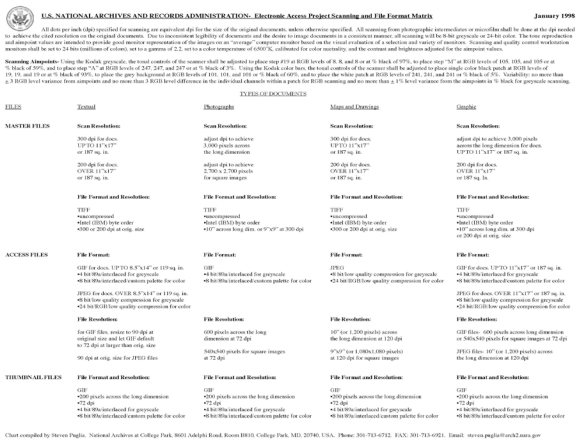

FIGURE 2. *EAP PROJECT SCANNING AND FILE FORMAT MATRIX*

The resulting master files from these specifications reproduced at 8 ½ x 11" at 300 dpi, or the equivalent of a good 8" x 10" analogue print. An Access GIF and a Thumbnail GIF file were created from the Master file for display in the NAIL database. Existing Web browser and bandwidth capacity could not handle large files efficiently, thus the actual file displayed in the catalog were much smaller derivatives that unfortunately did not display with sufficient quality for online discovery and research. The NAIL database and EAP derivatives served as visual markers to the actual hard copy analogue records. Researchers would use these resources as a precursor to performing on site research. However to this day many of these legacy files still reside in the current catalog.

Improvements to NARA's web presence

NARA's website received a major redesign in 1999 that included links to the Presidential Libraries, the Federal Register, and a new search engine to help researchers locate content. Another redesign occurred in 2002 when the Archival Research Catalog (ARC) replaced the NAIL Database, and allowed for more sophisticated search capabilities, better display interfaces and

ability to download images. Also, the 1930 U.S. census was hosted online, and included the addition of finding aids to make searching data much easier. Prior to this development census records were only available by purchasing microfilm copies.

A major renovation to the Archive's Rotunda closed public access to the main exhibit halls that house the United States' founding documents. As a result, new online exhibits were developed and NARA's digitization lab was tasked with providing high quality digital images of archival documents at greater resolution than the 1998 EAP project. Exhibits such as the Public Vaults relied upon high-resolution images for interactive displays that combined electronic, original, and facsimile records. The use of online exhibits developed by the education and exhibit departments increased NARA's online presence and provided additional contextual interactions with archival content. These efforts were the precursors of today's social media interconnected world.

Need for Better Guidelines

During this period NARA's photo labs were getting a greater amount of digitization requests than for analogue photography. The EAP project specifications were not suited to serve the digital products and services of the agency. NARA's 2004 *Technical Guidelines for Digitizing Archival Materials for Electronic Access* were developed out of a need to codify best practices as digitization emerged as the primary imaging tool in the archival field.

The publication of the 2004 Guidelines was a major leap forward to describe best practices for the archival imaging community. They define NARA's approach to create digital surrogates for access and reproduction purposes. The challenge was to develop an approach that met many competing uses that have very different requirements such as hard-copy reproduction, online access, and exhibit needs. What the Guidelines do is define the approach to create use-neutral reasonable reproductions of original documents.

Defining Production Master and Derivative file types

By defining the image quality parameters for "master" files, derivatives share the characteristics of the parent file. The digital lab workflow created several types of files for these purposes: an unadjusted scan of photographic negatives (a high bit TIFF that has no tonal adjustments and retains the same reversed polarity as the negative), a master file that is a tonally corrected TIFF that included color and dimension targets while retaining the original document dimensions, an access master tiff or jpeg at original size that has all reference targets cropped out, and a print file resized to create an image within a 2400 x 3150 pixel array suitable for reference printing and web and social media uses.

The following are the standard file types and attributes for digital scans of original archival records. The file naming protocol for digital objects contain the suffix _us, _ma, and _ac (123456_2014_001_us.tif) to denote what kind of file it is. A standard filename contains a local identifier (negative number or other archival identifier), the digital file creation date (useful to avoid duplicate numbers and manage rescans), an iteration number (an aid to managing sequentiality of a multi-part record), and a digital file type identifier. Examples below:

_us Unadjusted Scan of color or black and white negative
- Preservation or reproduction master file
- Saved as a positive (negative appears as a negative)
- Edge of film is visible
- Image de-skewed
- 16-bit uncompressed/LZW grayscale TIFF, 48-bit RGB color TIFF
- Image is "right read" (image is not backwards text in image reads correctly)

_ma Master File: A color and tonal corrected master scan
- Preservation or reproduction master file
- Contains dimensional and/or color reference targets
- 3-point color neutralization (+/- 3 levels between channels) of Kodak Q-13 grayscale. Grayscale: highlight below 97% and shadow above 3% black
- No cropping of image
- 8-bit uncompressed grayscale TIFF, 24-bit RGB color TIFF
- Positive image of a negative

_ac Access master: reproduction master file derived from the MA file for online access
- color targets and dimensional scales cropped out
- JPEG quality level 10 and above
- sRGB 24 bit color or 8 bit grayscale

The Guidelines provide sections such as metadata, quantifying scanner/digital camera performance, image adjustments, and an expanded section on digitization specifications for a large group of archival record types. The metadata recommendations conform to NARA Directive 1301 that define Lifecycle Data Standards that form the data authority structure for the ARC Catalog. Eventually the 2004 Guidelines were transformed into the Federal Agency Digitization Guidelines Initiative (FADGI) publication "Technical Guidelines for Digitizing Cultural Heritage Materials". [4]

Innovation Revolution

A major evolution took place with the inception of NARA's Office of Innovation whereby the major digitization functions, management of NARA's online catalog, and Social Media efforts were brought together for the first time. A crucial change occurred with the 2007 launch of the Online public Access Catalog (OPA) that allowed for the digital object to be a central focus of the user experience. Prior to OPA, traditional archival description was the primary access point to archival holdings and the digital version was an afterthought. The introduction of OPA allowed researchers to visually zoom in and see details of digital images. These redesigns were introduced as the result of user studies that researchers wanted better access to the digital object.

On the other hand these changes required a skillful paradigm shift within institutional culture that was anxious about the emphasis on the digital object. A pivotal point was the change management initiated by the Office in order for the agency at large to accept this new emphasis. After the initial period of adjustment,

users were not only delighted with the improvements, they asked for more enhancements. This change in the catalog helped to spur the digitization strategy and the strategic objective of digitizing everything. We realized that for many users, if a record wasn't available online, it simply did not exist for them. OPA was also the first iteration of the catalog that allowed the public to add tags to the digital copies, making the copies easier to find by others. This was a big step toward developing the social catalog we have today.

The impact of Social Media and user experience

While the Catalog remains the core online platform for digital discovery, NARA's Office of Innovation sponsored pioneering Social Media projects to connect archival content to audiences beyond the traditional researcher community. In 2009 NARA launched its first blog, and soon afterwards added content to Flickr, YouTube, and Facebook. These efforts were part of NARA's 2010

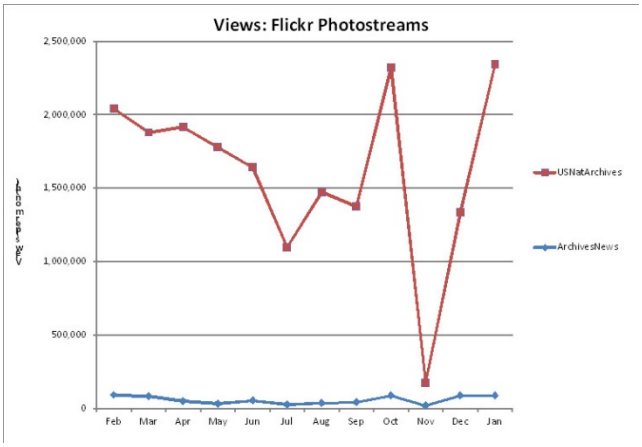

Figure 3: *Number Flickr views per month fiscal year 2015*

Social Media Strategy to increase government transparency, empower staff to use social media, and build communities of Citizen Archivists. [5] The impact of this strategy has resulted in a digital presence that reaches hundreds of millions of people, involves over 200 staff that create content on 14 different platforms.

Catalog Stats as of 2/2017

Page views:	3,846,804
Total number of Descriptions:	13,547,466
Total Number Digital Objects:	24,098,261
Digital partnerships:	15,593,880
Citizen Archivist Tags:	273,453
Citizen Archivist Transcriptions:	226,958

Social Media stats: end of FY 17 [6]

Total cumulative views:	730,891789
Flickr Views:	183,372,325
Twitter Followers:	590,649
YouTube views:	9,220,259 (2,413 videos)

The Catalog in 2016 experienced major improvements that allowed for higher resolution images to be hosted, new citizen archivist tools, and advanced search functions that use optical character recognition of text in API data. The ability to display and download high resolution images from a web based catalog is a major step towards fulfilling some of the promises of digitization to help preserve originals by decreasing handling and increasing discovery through higher quality representations. With these enhancements in place the Catalog staff embarked on a pilot project to replace low quality images in the catalog with preservation master unadjusted scans, production and reproduction master scans, as well as reasonable quality print master images.

A test case to bridge the gap between access and discovery

For example, a well-known image of Abraham Lincoln at Gettysburg was chosen as a test case to present online the master scan of the negative, a master production scan as well as master

Figure 4: *Cropped image of "Crowd of citizens, soldiers, and etc. with Abraham Lincoln at Gettysburg" the person depicted in the red box is Abraham Lincoln.*

access copies. [7] In 1952 NARA archivist Josephine Cobb discovered in a glass plate negative taken by Mathew Brady an image of a person suspected to be Lincoln at the speaker stand at Gettysburg where he gave the Emancipation Proclamation. [8] Up until the time of the discovery no depiction of Lincoln at Gettysburg was known to exist. Because of its rarity, historical significance, and fragility access to the original is restricted. In the 1990's a direct duplicate 8"x 10" contact negative was created for preservation access purposes.

In 2008 NARA Digital Imaging Lab staff made a high-resolution master scan of the negative and a set of derivative files. NARA Catalog staff placed the Unadjusted Scan (_us) TIFF and derivatives into the catalog record so that researchers could download for themselves the same high quality digital file previously only accessible to a small group of specialists.

Figure 5: *Unadjusted Scan, "Crowd of citizens, soldiers, and etc. with Abraham Lincoln at Gettysburg*

Figure 6: *Access Scan, tone reversal and scales cropped*

The test case to place high-resolution master files was an experiment to give researchers access to as close as possible to the original archival records. Limitations in catalog platforms, web display, and governmental accessibility regulations all contribute to a somewhat fragmented user experience. However for high demand or historically relevant items it is assumed the public and researchers want access to unadjusted and master scans of important archival records. Uploading all high-resolution files to the catalog is not practical for all records. A scalable approach is necessary to balance quality and quantity.

The emergence of the Citizen Archivist

One of the most important innovations to the Catalog is the addition of the Citizen Archivist Dashboard that provides tools to crowdsource tagging and transcription of online records. [9] Most of the founding documents and a majority of archival records consist of handwritten 18th and 19th century manuscript. NARA has worked diligently to engage with citizens to help with transcription of these records. NARA has tapped into enthusiastic volunteer communities who have transcribed over 250,000 pages of text. These efforts allow for easier reading, searching, and use of these records.

A well thought out strategy of creating "Transcription Missions" and "Tag It" campaigns improve the search features of the catalog. NARA continues to expand its engagement by exploring social media efforts such as "Today's Document" on Tumblr and working with App developers such as History Pin's "Remembering WWI collection" WWI App. Also with our Open Government efforts, we wanted the users to not only be able to find and use digital copies, but actually help us to get more digital copies into the catalog. Our public scanning stations in the Innovation Hub have supported over 100,000 digital copies getting into the catalog since 2011. Today we have community managers devoted to developing public input into the catalog. We see the catalog as a platform for user interaction on a number of levels, rather than simply a listing of our holdings. [10]

Need for a scalable approach

The new catalog tools rely upon imaging standards to include scans of records that are "transcription" quality. While NARA's early digitization effort yielded the EAP Project Guidelines, over time non-specialists on consumer grade equipment are creating more and more digital content. Recognizing that professional imaging standards such as the 2004 Guidelines, and the FADGI Guidelines are not user friendly for most operations, the Digital Public Access Branch staff developed a user-friendly set of guidelines. These guidelines adapted the technical FADGI specifications with NARA's Agency Transfer Guidance File Format Appendix, to create a hybrid guide to submitting digital objects to the catalog. [11] These file specifications follow the _ac.jpeg specifications described above, but are written in a non-technical language to encourage adoption and conformance.

Achieving FADGI 4 star quality for the bulk of archival record digitization is not a practical workflow for a majority of records. NARA has given itself a goal to place 500 million digital objects into the catalog by the end of FY 2018. The challenge is to balance the addition of new scans into the catalog while maintaining image quality standards. The files created for the EAP project remain a viable resource 20 years later, however anticipating technological advancements, meeting citizen archivist expectations for more access, and meeting Institutional goals for the future, point to the need to maintain above average image specifications. While establishing guidelines, standards, and requirements are necessary parts of a digitization workflow, getting the products to the platform, and relating to an online audience is equally important. If imaging science is to have a larger impact on the archiving community it must be understandable to practitioners and meet the needs of citizens.

References

[1] Jessie Kratz, "The National Archives on the 25th anniversary of the World Wide Web", https://prologue.blogs.archives.gov/page/28, 2014

[2] Steve Puglia and Barry Roginski, "Nara Guidelines for Digitizing Archival Materials for Electronic Access" https://www.archives.gov/files/preservation/technical/guidelines-1998.pdf, 1998

[3] Steven Puglia, Jeffery Reed, and Erin Rhodes, "Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files-Raster Images", https://www.archives.gov/files/preservation/technical/guidelines-matrix.pdf, 2004

[4] Federal Agencies Digital Guidelines Initiative, "Technical Guidelines for Digitizing Cultural Heritage Materials", http://www.digitizationguidelines.gov/guidelines/digitize-technical.html, September 2016

[5] National Archives and Records Administration, "Social Media Strategy 2017-2020", https://usnationalarchives.github.io/social-media-strategy/introduction/, 2017

[6] National Archives and Records Administration, "Social Media Statistics", https://www.archives.gov/social-media/reports, 2017

[7] Photograph No. 111-B-4975 (Photographer Mathew Brady); "Crowd of citizens, soldiers, and etc. with Abraham Lincoln at Gettysburg"; Mathew Brady Photographs of Civil War-Era Personalities and Scenes, 1921 - 1940; Record Group 111: Records of the Office of the Chief Signal Officer, 1860 - 1985; National Archives at College Park, College Park, MD; https://catalog.archives.gov/id/529085

[8] Rob Crotty, "Rare photo of Lincoln at Gettysburg", https://prologue.blogs.archives.gov/2010/11/19/rare-photo-of-lincoln-at-gettysburg/, 2010

[9] National Archives and Records Administration, "Citizen Archivist Dashboard: Transcription Tips", https://www.archives.gov/citizen-archivist/transcribe/citizen-archivist/transcribe/tips, 2017

[10] Pamela Wright, Chief Innovation Officer, US National Archives and Records Administration, personal communication, March 1, 2017

[11] National Archives and Records Administration, "Bulletin 2014-04, Revised Format Guidance for the Transfer of Permanent Electronic Records", https://www.archives.gov/records-mgmt/bulletins/2014/2014-04.html, 2014

Additional Resources:

https://www.archives.gov/files/preservation/technical/guidelines-matrix.pdf
https://www.archives.gov/preservation/technical/guidelines.html
http://www.digitizationguidelines.gov/guidelines/digitize-technical.html
https://www.archives.gov/citizen-archivist/transcribe
https://www.archives.gov/social-media/policies/tagging-policy.html
https://usnationalarchives.github.io/opengovplan/
https://www.archives.gov/digitization/partnerships.html

Author Biography

Michael Horsley is an Electronic File Format Specialist in the Office of the Chief Records Officer, US National Archives and Records Administration. Previously he was a Digital Imaging Specialist in the Digital Public Access Branch, Office of Innovation. He has a BA in History from the University of Maryland and a Certificate in Digital Imaging from the Corcoran College of Art and Design. He has over 20 years of experience in the Cultural Heritage digital imaging field and has worked at the Smithsonian Institution and Library of Congress.