Digitization Standards at the National Archives and Records Administration

Jeffrey Reed, Kate Murray, and Martin Jacobson, NARA, College Park, Maryland, USA

Abstract

The US National Archives and Records Administration (NARA), through the Digitization Planning Branch in the Office of Innovation, is developing flexible and appropriate agency-wide standards for digitization in order to advance NARA's goal of making its holdings more available. Recognizing that there is no single answer to the question "what format should I use," this effort uses the One Touch and Fit For Purpose concepts to document file format and attribute choices that will meet a variety of known and expected uses for the digitized material.

Focusing on a collection of use cases, this paper examines the essential components of the effort: characterizing customer groups including file creators and file consumers, defining the intended uses for digitization products, designing digitization products to satisfy these needs, and finally, packaging the standards information at an appropriate level of complexity for a variety of user communities. The effort reflects a broader agency-wide approach to systematic digitization and acknowledges the growing effectiveness of distributed digitization including utilizing commercial partners, crowd-sourcing and other community-driven initiatives.

Previous Work

Over the past 15 years, the National Archives and Records Administration (NARA) has advanced through several major milestones in formalizing digitization specifications. An early project was the Electronic Access Project (EAP), a pilot for large scale digitization of NARA holdings which created a core set of 124,000 images expressly to populate NARA's online catalog. The *NARA Guidelines for Digitizing Materials for Electronic Access* [1] developed for this project established digital imaging requirements to create master, access, and thumbnail files from paper and film based records. Because preservation copying was not a goal of the EAP project, specifications for preservation copies were not addressed. The EAP guidelines were revised in 2004 to become the *Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files – Raster Images* [2]. While the 2004 guidelines added more informational content, focusing on concepts and good practices for producing good master images that could be used for a variety of reproduction purposes, they were still noted as not being appropriate for preservation reformatting.

In 2010, the Digitization Services Division published the Products and Services (P&S) web portal which defines digital and analog products for all media types produced by the in-house reformatting labs and included preservation copies as a product[3]. P&S was a return to more prescriptive specifications, as compared to the 2004 guidelines' options, but the specifications were limited to activities within the division. Simultaneously, NARA staff from the static and dynamic media labs were participating in the Federal Agencies Digitization Guidelines Initiative (FADGI), a collaborative effort led by the Library of Congress to define common guidelines, methods, and practices for federal agencies digitizing cultural heritage materials [4]. NARA capitalizes on the technical expertise within FADGI to inform internal standards development and stay in sync with the federal segment of the digitizing community.

Revisiting Digitization Standards

As NARA refines its overall strategy for digitizing and making available its holdings, an important component is determining appropriate file format and attribute standards for still image and audiovisual digital products that are suitable for NARA's array of uses. Knowledge and implementation of digitization standards are uneven across NARA. Some areas have well-defined standards in use (such as P&S), others use less formal guidance, while still others have an ad hoc approach to creating digital files and do not implement consistent standards at all. The situation is further complicated by the fact that NARA increasingly utilizes various means for digitizing its holdings, including in-house operations of varying complexity and resources, external vendors, commercial partners, citizen archivists, and GOCO (Government-Owned, Contractor-Operated) operations. The result is a diverse set of content creators manufacturing an inconsistent and changeable set of digital products that may not satisfy current and future needs. This multi-faceted landscape taxes resources needed to create, maintain, or provide access to the digital products in coordinated and sustainable ways.

The goal of the Digitization Planning Branch's Digitization Standards Project is to create an easy-to-understand, widely-available, practical, flexible, and policy-backed set of standardized digitization products to be used by all entities digitizing NARA records in an official capacity. This agency-wide approach will help NARA make the most of its digitization efforts by defining appropriate technical specifications for a variety of intended end uses, and by creating consistent products around which more efficient production workflows and tools can be built. Preservation, access, and management of the files also will be simplified because there will be less diversity in the resulting digital products. While the definition of the standards set is an essential component of the project, equally as important is the tailoring of the standards information so that it can be accessed, understood, and implemented by the diverse set of content creators.

Ultimately, the goal of the project is to provide NARA's digitizing entities with the appropriate information, delivered in a variety of packages matched to different user profiles, to make informed decisions about technical specifications in order to advance NARA's goal of making its holdings more available.

Essential Concepts

The Digitization Standards Project is centered on two essential concepts that work together and tie into NARA's digitization strategy: "One Touch" and "Fit For Purpose".

One Touch, in this context, refers to limiting repeated contact with original content by scaling digitization capture standards to be robust and flexible in order to satisfy a variety of intended end uses, including long term preservation and flexible access options. On a practical level, One Touch means that in order to maximize the impact of digitization efforts, NARA will digitize the content and store the information at a sufficient quality level and within appropriate file structures to satisfy as many known and expected uses as practical (80% or more) for a given item or collection through either direct employment of the master file or through derivatives created from that file.

One Touch mandates that before digitizing to the specifications that meet immediate needs, careful consideration is taken of the other possible needs and a product is selected based on expected uses over the long term. This concept applies to the selection of the master file intended for re-use and in many cases, less demanding needs may be filled by derivative files which are also defined and tested for fitness for purpose. An important aspect of One Touch is that it is not a proposal to digitize all materials at the maximum capture levels to cover any possibility. One Touch helps NARA strive to digitize at an appropriate level and avoid gold-plating or underestimating information capture needs.

Fit For Purpose (FFP) means that a product will satisfy the customer's defined purpose. When applied to digitization standards, this means that content is captured in digital files at appropriate specifications to satisfy specific technical and consumer-driven needs. Although a simple concept that has appeared in this and other communities, it is important for NARA to formalize its usage in creating digitization product standards. There is often uncertainty about the uses a digital copy can fulfill at the time of creation or when a file is discovered. While P&S connects products to general purposes (preservation, reproduction, or distribution), this project will more closely match standard products to specific uses that come from daily production. In addition, product samples will be tested across user groups to ensure agreement about the fitness of the product for the defined purposes.

The digitization FFP concept has three major components:

- Customer Groups
- Intended Uses
- Digitization Products

The relationships and interactions between these components are explored in specific use cases.

Customer Groups

Digitization standards information has two major customer groups: file creators and file consumers. File creators include internal NARA staff such as the established in-house reformatting labs of the Digitization Services Division and in several Presidential Libraries, as well as the scanning operations in archival units and regional facilities across the country. There is a significant community of external file creators as well including commercial partners such as Ancestry.com, commercial vendors, organized Citizen Archivists groups, and Reference Room patrons scanning for their own personal use. File consumers also come in both internal and external flavors. Internal file consumers include NARA websites, NARA's Online Public Access catalog (OPA), exhibits staff, specialized NARA staff including social media staff and archival processing, and finally data repositories such as the Electronic Records Archive (ERA) and other storage environments. External file consumers include direct-to-customer requests and social media outlets such as Flickr, YouTube, FourSquare, Tumblr, HistoryPin, and Storify. Internal or external, creator or user, each group has expectations for digital copies that will be addressed in the creation of the standards.

Intended Uses

In order to better understand and organize the intended uses, categories of use were distilled from many specific examples. The categories are based on common characteristics found in the encompassed use cases as well as properties of the copies needed to fulfill the use. NARA's intended uses for digitized material are summarized into four basic categories.

- **Interpretation of original content** in which the use of the copy is to see or hear the information conveyed in the original item, such as the words written or spoken, or the scene photographed. Additional information unnecessary for understanding the content may be omitted or unlike the original, so the user experience may be different from interacting with the original.
- **Representation of original item** in which the use of the copy is characterized by the user's desire to see or hear a copy that represents the original item beyond just the content in order to approximate the experience of interacting with the original. Although the copy may not provide all of the functionality of the original item, the user is provided a sense of the look and feel of the original.
- **Functional copy of original content**, which demands a surrogate that can replace the original item, a virtual copy. Although all of the properties of the original item cannot be replicated, all of the expected functionality would be retained.
- **Alteration of original content** is unified by the need for copies that change the representation of the original to improve its functionality or derive additional information that would not be discovered through normal interaction with the original.

It is important to note that the use categories are a helpful construct that supports discussion and decision-making, but they do not provide a defined path to a particular product nor is there an exclusive relationship between a use category and a product.

Digitization Products

The primary outcome of this project is a suite of digital copy specifications. Rather than start from scratch, P&S will be analyzed and adjusted for NARA-wide use. In P&S, NARA modeled its digitization product categories on the three-product set that it uses for reformatting traditional materials: preservation master, intermediate or reproduction master, and use copy. However, the uses of digital material don't perfectly align with those of traditional format material where preservation copies are completely inaccessible and reproduction masters can be either a child of a preservation master or be the highest quality file in a set without a parent.

To clarify these different roles, NARA is slightly revising its categorization of digital product sets to better delineate digital products based on certain criteria. Digitization products will be divided into three groups: Master Files, Digital Intermediates, and Distribution Copies. Because of the ability to clone and deliver any product without perceivable loss, all products serve the primary mission of increasing accessibility of NARA's holdings, either immediately or eventually.

Master Files are always managed for the long term and always serve preservation needs. The risks addressed, along with other characteristics, divide the master category into Low Master and High Master. Low Master files are made from low risk originals and minimize risk due to frequent handling or high use whereas High Master files are made from high risk originals and minimize risk due to inherent vice (obsolescence or deterioration) as well as handling.

The Digital Intermediate file is a workflow file that always has a High Master or Low Master parent. Its use is in translating the Master File into a Distribution Copy when the cost is too high to go directly from a Master File to a Distribution Copy but it is anticipated that future translations to new Distribution Copies are expected. This may include transcoding from a large data rich master such as a DPX sequence to a middle state file like an MPEG-2 before creating a web deliverable WMV and a DVD-ready file. The middle state MPEG-2 will take far less time and bandwidth to transcode than the DPX sequence but is robust enough to create a variety of Distribution Copies.

Distribution Copies are distinguished by their short term retention period. This category is created to meet the file requirements of specific delivery systems and end user capabilities. In general, these products emphasize small file sizes and faster transfer rates over high quality levels, or are specific, limited-use renditions. They may be made from source materials or may be a derivative of Master Files or Digital Intermediates.

Digitization Product Characteristics Summary

High Master

- **Amount of information contained:** High
- **Risk addressed**: High risk of original due to inherent vice (obsolescence or deterioration) AND/OR High value
- **Immediacy of use:** Could be immediately usable OR Need to create Digital Intermediate and/or Distribution Copy
- **Retention period:** Long term
- **Error tolerance/deviation from ideal aim:** Low
- **Source:** Original record material
- **Reference Target/Bars/Tones (embedded or linked):** Required
- **Ease of creation (tools and expertise):** Difficult/high
- **Use Summary:** Use Neutral
- **Can this satisfy the Function intended use?** Yes
- **Can this satisfy the Representation intended use?** Yes
- **Can this satisfy the Interpretation intended use?** Yes
- **Can this satisfy the Alteration intended use?** Maybe

Low Master

- **Amount of information contained:** Intermediate
- **Risk addressed**: Risk due to frequent handling or high use (loss, theft, wear, and mishandling)
- **Immediacy of use:** Could be immediately usable OR Need to create Distribution Copy
- **Retention period:** Long term
- **Error tolerance/deviation from ideal aim:** Medium
- **Source:** Original record material
- **Reference Target/Bars/Tones (embedded or linked):** Not required
- **Ease of creation (tools and expertise):** Less difficult/medium
- **Use Summary:** Defined Use
- **Can this satisfy the Function intended use?** Maybe
- **Can this satisfy the Representation intended use?** Yes
- **Can this satisfy the Interpretation intended use?** Yes
- **Can this satisfy the Alteration intended use?** Maybe

Digital Intermediate

- **Amount of information contained:** Intermediate to high
- **Risk addressed**: N/A
- **Immediacy of use:** Could be immediately usable OR Need to create Distribution Copy
- **Retention period:** Long term
- **Error tolerance/deviation from ideal aim:** Low
- **Source:** Master File
- **Reference Target/Bars/Tones (embedded or linked):** Not required
- **Ease of creation (tools and expertise):** Less difficult/medium
- **Use Summary:** Defined Use
- **Can this satisfy the Function intended use?** Maybe
- **Can this satisfy the Representation intended use?** Yes
- **Can this satisfy the Interpretation intended use?** Yes
- **Can this satisfy the Alteration intended use?** Maybe

Distribution Copy

- **Amount of information contained:** Variable
- **Risk addressed**: N/A
- **Immediacy of use:** Immediately usable
- **Retention period:** Short term
- **Error tolerance/deviation from ideal aim:** High
- **Source:** Master File OR Digital Intermediate OR original record material
- **Reference Target/Bars/Tones (embedded or linked):** Not required
- **Ease of creation (tools and expertise):** Not difficult/low
- **Use Summary:** Specific Use
- **Can this satisfy the Function intended use?** No
- **Can this satisfy the Representation intended use?** Maybe
- **Can this satisfy the Interpretation intended use?** Yes
- **Can this satisfy the Alteration intended use?** Maybe

Specific Use Cases

We have selected several examples from the large collection of use cases we have gathered so far in order to illustrate how these pieces fit together. Although the digital products described in the P&S web portal will be used as the foundation for the standards in this project, review, analysis, and adjustment hasn't occurred yet so the following examples will refer only to the product categories listed above and not the full product specifications. While the examples depict a rationale behind

selecting solutions for needs based on these concepts, a formalized decision process is not in the scope of this project.

Example 1: A researcher wants to find a relative's address in a census record. In this case the file user only needs informational content from the copy. He is not interested in the appearance of the page containing the information nor does he need a life-like copy. This specific need falls in the Interpretation category – content only. The user needs a copy that allows him to read the words contained on the page. The FFP concept prescribes there is a digital file captured at appropriate specifications to satisfy this specific need. The directly matching product for that need falls in the Distribution Copy product category which describes copy types that are optimized for delivery mechanisms and are only maintained as long as needed. Although very low on the information capture scale and permitting a large tolerance for error, this product is appropriate to satisfy this user's need without extra expense for unnecessary embellishment. Before making the copy to meet only the immediate need, the One Touch concept asserts that consideration of future needs be made and a product should be selected to minimize repeated contact with the original. It is quite likely that others will want access to the same item indefinitely, so a Master File should be created and maintained for the long term. The item, however, is not at risk of immediate loss of information from deterioration so a Low Master is appropriate for the low risk. Like the first researcher, other researchers may only be interested in the informational content but some have an expectation that they can see the document as it really looks and the exhibits staff will want to display a same size facsimile print in an upcoming exhibit. These realistic expected uses would be satisfied by the Low Master file. This copy captures the color of the document as part of a reasonably faithful representation of the page as an object. The spatial frequency is sufficient to resolve all of the written content as well as provide a sense of the paper texture and allow for small scale zoom or print enlargement. The increase in information captured and fidelity to the original now covers the expected uses that extend into the Representation category while still meeting the triggering request. If the data size or format is incompatible with delivery mechanisms, a Distribution Copy can be derived from the master to meet more specific needs. For this original, it is not anticipated that a fully functional replacement copy is needed nor is it expected that an altered representation would be needed that couldn't be derived from the master created (e.g. increased contrast for enhanced legibility), so a High Master would be unwarranted.

Example 2: A production company needs standard definition video material depicting family life on US Military bases for a TV documentary. Content is the primary need in this example as in the first, but unlike example 1, this use requires a copy that closely matches the original image and sound. This specific use fits in the Representation Category. A reduction in the informational capture from that contained in the original, such as loss of detail from low sampling, would still convey the substance of the material and could be useful for discovery, but the copy would be insufficient for the broadcast demands. The Low Master file would be a sufficient capture for this immediate need, but One Touch brings other considerations into the choice. The original is considered high risk because of the rapid obsolescence and deterioration of the media type; therefore it is paramount that the copy captures the highest amount of information to replicate, as much as possible, the full functionality of the original. Because of the high risk, a High Master file would be selected to satisfy the expected uses and a Distribution Copy would be created and delivered to the user to best match their request. A Digital Intermediate is unnecessary because the translation from the Master File is economical through automation.

Example 3: An archivist needs to copy a group of nitrate film negatives so they can be destroyed to comply with agency storage regulations. This specific use very clearly is within the Functionality use category because the copy needs to replace the original. The copy must carry forward as much information from and about the original to permit its disposal. The High Master is the appropriate product for this use and, because of the high information capture level, almost any other use can be derived from this master except some extreme, alternative imaging processes (e.g. X-Ray). In this example with film negatives, the High Master captures as high fidelity copy as possible which provides the replacement functionality, but like the original negatives, the copy is not a rendering desired by most end users. Most uses of the copy fit in the alteration category – a different appearance from the original is required. Special capture is not needed for this but a transformation (tone reversal and adjustment) will have to be done that is not currently automated. Therefore, the transformation output would not be a Distribution Copy but rather a Digital Intermediate to save as well as deliver. If necessary because of limitations of delivery mechanisms, a Distribution Copy may be created for specific types of access.

Example 4: A researcher needs to confirm typographical edits were made to a document. The changes are not visible within the document under normal viewing conditions but are mentioned in a related document. To reveal the hidden change, an enhancement of the appearance of the original is required. This use is not satisfied by a straight-forward image of the document under normal lighting conditions and would be categorized as an Alteration use. The document is not at risk from deterioration or obsolescence but the copy should be managed for long term use so a Low Master is the appropriate product category even though the imaging process will produce a copy that does not look like the original. In fact, it may not convey the visible content of the original at all. Because this specific use limits the other possible uses, higher information capture levels wouldn't be beneficial.

Packaging and Dissemination

An important part to ensure the success of developing agency-wide digitization standards is packaging the information so that it can be accessed and understood by NARA's diverse community of content creators and product users. The current version of the P&S web portal is information rich by design but overwhelming to some readers. The format specification pages include all relevant information about a reformatting product that might be needed to select or justify a particular format or attribute choice. However, now that P&S is expanding to serve all digitization entities and products at NARA, the information levels need to be adjustable for audiences of different interest levels and technical abilities. We have identified the core customer groups of internal and external file creators and file consumers and we are working to develop three main user profiles based upon the level of information relevant to each group. The Novice level is the most non-technical and includes expected users such as Citizen

Archivists and the general public. To serve this community, the "Fundamental Five" concept was developed to highlight only the five most essential pieces of standards information in easy-to-understand language and a simplified layout. Its goal is to present basic information in a straightforward way. The next level is the Proficient level which expands on the Fundamental Five concept and is appropriate for those with a little more experience but who are not yet experts. The goal of this information package is to cover the basics plus a little more. This would include customers such as archivists scanning on site with desktop scanners or emerging digitization activities in other areas of the agency. Finally, the Expert level includes all available information in detail. Customers for this level of standards information include the digitization staff in the College Park reformatting labs, established digitization programs in selected Presidential Libraries, and commercial partners such as Ancestry.com. The information content grows with each level and the tone gradually shifts from casual to professional. Readers will be able to move from one user profile to the next as their interests and knowledge expands. The architecture of the final web-based resource is not yet finalized but the infrastructure will need to be flexible and able to reuse data for the multiple presentations in an efficient and practical way.

Conclusion

Digitization strategy at the National Archives and Records Administration is constantly evolving. Because of the volume and variety of materials, uses, and actors, standardization is a vital keystone for success in meeting agency goals. Unification of approach across the agency will provide many benefits, but neither choosing a single standard to cover every possibility nor centralizing decision-making is a workable solution for our situation. Guided by the concepts presented here, the project team will assemble the file format and attribute specifications for digitization products to be adopted across NARA. With a flexible yet succinct rational to organize the specifications, consistent decision making can be made at any level. Future work will include review and analysis of existing specifications, development of relationships with other stakeholders to ensure needs and expectations are met, testing and finalization of the suite of specifications, and building the information delivery channels that will promote wide-spread assimilation. Many people have put hard work into paving the way before us. This project to align the diverse operations across the agency will build on those antecedents and is an important step in NARA's digitization evolution.

References

[1] Steven Puglia and Barry Roginski. NARA Guidelines for Digitizing Archival Materials for Electronic Access, January 1998. Available online: http://www.archives.gov/preservation/technical/guidelines-1998.pdf.

[2] Steven Puglia, Jeffrey Reed, and Erin Rhodes. Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files – Raster Images For the Following Record Types- Textual, Graphic Illustrations/Artwork/Originals, Maps, Plans, Oversized, Photographs, Aerial Photographs, and Objects/Artifacts, June 2004. Available online: http://www.archives.gov/preservation/technical/guidelines.pdf

[3] National Archives and Records Administration. Digitization Services Products and Services. Updated 2012. Available online: http://www.archives.gov/preservation/products/

[4] Library of Congress. Federal Agencies Digitization Guidelines Initiative. Available online: http://www.digitizationguidelines.gov/

Author Biography

Jeffrey Reed is a Digital Process Development Specialist in the Digitization Planning Branch at NARA. After earning a BA in Architecture from VPI&SU in 1991, Jeffrey began reformatting at Photo Preservation Services, Inc, a company specializing in preservation reformatting of photographic materials. He added digitization skills at Rieger Communications, Inc., a commercial service bureau, before joining NARA in 2002 as a Digital Imaging Specialist. Prior to his current position, Jeffrey supervised the Photographic Imaging Lab.

Kate Murray is a Digital Process Development Specialist in the Digitization Planning Branch at NARA specializing in standardizing and documenting moving image and audio formats. Prior to joining NARA in 2008, Kate has held positions at the University of Maryland Libraries, Emory University Libraries, the University of Cape Town Libraries, and NYU Libraries. Kate received her undergraduate degree in medieval literature from Columbia University and her MLS from the University of Cape Town.

Martin Jacobson is Chief of the Digitization Planning Branch at NARA. The branch's initial focus is on assembling relevant digitization standards for use throughout NARA. Martin's career in archiving began in 1994 at the Swedish National Audiovisual Archive where he held various positions including Head of Technology and Head of Archives and Preservation. He joined NARA in 2008 as Director of the Special Media Preservation Division, and has been in his current position since 2012.