https://doi.org/10.2352/issn.2168-3204.2023.20.1.35

Gimme Three Steps: A Mass Digitization Method at the Smithsonian

Nathan Anderson, Jeanine Nault, Luis J. Villanueva; Digitization Program Office, Smithsonian Institution; Washington, DC, USA

Abstract

The Smithsonian Institution Digitization Program Office's Collection Digitization team develops and designs a "three-pronged" workflow approach to mass digitization of museum collections, called the Physical, Imaging, and Virtual Workflows. This approach addresses proper handling of objects, optimizing capture throughputs, and streamlines the processing and delivery of images through automation. The Physical Workflow Design defines the production space and safe movement of objects from storage to the digitization production space; the Imaging Workflow Design defines the technical specifications, file deliverables, and the results of our 'Item Driven Image Fidelity' (IDIF) testing; and finally, the Virtual Workflow Design defines the lifecycle of the digital file, from creation to online access, describing the various data processes required for success.

Motivation

In 2013, the Smithsonian created a new office, the Digitization Program Office (DPO), a central service unit within
the Institution's Office of the Chief Information Officer. DPO's three main teams include: Policy and Analysis, which oversees our digitization metrics and identifies policies and plans that address the Smithsonian's digital assets; 3D Digitization which captures subsets of Smithsonian collections in 3D for use in novel ways in education, research, and by the general public; and finally, Collections Digitization (formerly "Mass Digitization"), which supports Smithsonian museums in their efforts to digitize collections as comprehensively, quickly, and cost-effectively as possible and by pairing up the digital surrogates created with the collections records stored in the Smithsonian's collections information systems. In doing so, we advance the mission of the increase and diffusion of knowledge by making as much of the Smithsonian's collection available online as possible.

Problem

The major hurdle for mass digitization at the Smithsonian is *scale*, given the enormous size and variation of the collections held in the twenty-one museums and the National Zoo. The Smithsonian Institution holds the largest museum collection on the planet, with an estimated 155 million specimens, objects, archival items, and library holdings. Given the size and extent of our collections, how can we efficiently digitize them? How can we make them available without sacrificing quality, and in ways that enhance their use?

Approach

When we talk about mass digitization, what we are talking about is an approach whereby we digitize hundreds, thousands, or millions of items (museum objects or specimens of diverse sizes, shapes, and imaging needs) while also maintaining an elevated level of object care and safety, as well as image quality.

We have developed a three (3) step workflow approach to properly handle our objects, optimize our capture throughputs, and streamline our automated delivery of images: physical, imaging, virtual workflows.

The three-pronged workflow approach follows the guiding principles for mass digitization:

- Build comprehensive end-to-end workflows.
- Work at high volume to achieve economies of scale.
- Relentless pursuit of efficiency
- Implement robust project/process management.
- Item Driven Image Fidelity (IDIF)
- Build institution-wide infrastructure for repeatable, sustainable results.

This three-step workflow approach, coupled with our robust project management style, has led to the digitization of over five million of the Smithsonian's 155 million items.

Our method of project management includes the following four (4) stages:

- *Resource Coalition* (2-4 weeks): a series of meetings with senior stakeholders to answer the who, what, where, when, and how of a project, as well as the resources necessary to do so, such as staff time, supplies, equipment, and funding.
- *Long Term Preparation* (2-6 months): museum collections staff prepares for digitization, including physically prepping the collection such as rehousing or barcoding, as well as Prepping catalog records, which may include record creation or updating metadata. Meanwhile, the DPO project manager works on administrative tasks such as contracting the digitization vendor, drafting, and circulating a Memorandum of Agreement (MOA) between senior stakeholders, as well as workflow development and testing.
- *Pilot* (3-4 months): the first time we work with a museum, we perform a short pilot to evaluate workflows and practice for longer term production projects. The goal of the Pilot phase is to develop the efficient workflows necessary for a successful mass digitization production

project, as well as provide an immersive mass digitization experience in a low stress environment.

- *Production* (2-24+ months): While the planning for the Pilot phase is comprehensive and in-depth, the amount of digitization is limited to 1-2 weeks to reduce the stress that is sometimes associated with larger and longer production projects. Additionally, the shorter Pilot phase project allows for quick turnaround of lessons learned before undertaking a larger Production phase project.

As evident above, prior to production starting, a significant amount of planning is necessary; this planning is documented and administered via a Gantt chart with more than sixty (60) individual tasks to be completed over the course of a given project.

Physical Workflow Design

Our workflow design development usually begins with the Physical Workflow Design, which includes the following:

- Item Inventory and Collections
- Storage and Digitization Floorplans
- Item Handling Guidelines
- Movement Plans
- Object Placement Guidelines
- Equipment and Supplies

We are documenting the physical collections, including their safe movement, and handling, before and after digitization. The item inventory and collections outline the size and quantity of collections. We have throughput rates for daily digitization calculated based on the physical aspects of the collections such as their size (from macro, like a bumblebee to oversized like a chair) and dimensionality (flat like posters or non-flat like coins). The storage and digitization floor plans outline where and how the collections are stored and digitized. We do collections visits with museum staff and document these through photos, room measurements and accessing floorplans. We then mockup the production space floorplan to scale. The Item handling guidelines outline the safe handling of the collections. Museum staff trains our vendors in proper handling onsite before we begin digitization. The movement plans outline the path the collections will travel from their storage to the production space. It is important to not only document this path, but to walk the path and verify that the project team can move efficiently. Once we have safely moved and handled the collections safely, we document how to properly position them on the imaging stage to get the correct, consistent image.

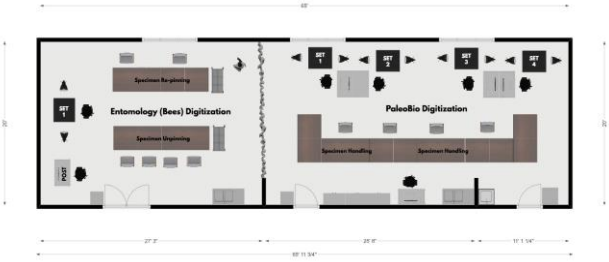

Figure 1. *Floorplan diagram of concurrent mass digitization projects at the National Museum of Natural History.*

Imaging Workflow Design

The Smithsonian Institution's approach to digitization is comprehensive. Specifically, the processes necessary to generate "research-worthy" digital images that stand the test of time. Digital images showing insufficient detail or that are too low in resolution make for a frustrating online experience for the public and researchers alike but capturing at excessive resolutions taxes our storage & network infrastructure unnecessarily. To ensure that we are achieving the optimal resolution for any Smithsonian managed digitization project, we go through a rigorous validation process we call Item Driven Image Fidelity (IDIF). IDIF validation begins with collections staff identifying a representative sampling of the collection that contains the smallest details.

Using an example collection item, the smallest details necessary to be resolved are measured at the micron scale and documented. This measurement translates into PPI (pixels per inch) resolution required to photograph the item's detail. We then validate our PPI resolution through spatial frequency response (SFR) testing. The IDIF analysis results in a project-specific imaging standard by which we measure capture quality for all Smithsonian managed digitization projects. Our process ensures that fine specimen or item details are resolvable for remote researchers to do meaningful work with collections even if they are half a world away.

Figure 2. *Specimen Drawer, National Museum of Natural History, Dept. of Paleobiology*

In this example from the Department of Paleobiology at the National Museum of Natural History it is a vial containing fossilized seashells.

Next, the curator chooses the best represented specimen from the collection and identifies the smallest detail needing to be of resolvable resolution for a research quality image. For bivalve shell identification, taxodont teeth along the hinge plate need to be of resolvable resolution for research. "Resolved" has a specific meaning as defined by the Rayleigh Criterion. The *Rayleigh Criterion* was originally formulated for determining the resolution of two-dimensional telescope images but has since spread into many

Figure 3. *Isolated specimen measuring 3 x 6 mm, National Museum of Natural History, Dept. of Paleobiology*

other arenas in optics. It is defined in terms of the minimum resolvable distance between two points in sources of light and is generally accepted criterion for the minimum resolvable detail of an item.

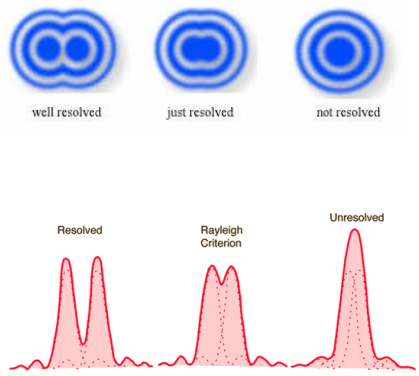

Figure 4. *Rayleigh Criterion illustrations*

For transmissive items, like 35 mm slides or 120 mm color transparency film, such as we digitized with the Archives of American Gardens' Garden Club of America collection, we use published research and established standards on the resolution needs for capturing the microstructure of film, which negates our need for microscopic investigation. Accepting the conclusions of previous research, we assume that the artifactual information in color film ranges from 5 to 20 microns in size and base resolution specification on this range. For color film, the fundamental image particles are color dye clouds, which range from 6-15 microns in diameter.[1]

Material/Item Type	Micron Measurement	Resolvable Resolution
Halftone print	Halftone dot: 30 microns	1280 PPI
Rotogravure print	Smallest mesh detail: 30 microns	1280 PPI
Collotype print	Smallest reticulated pattern shape: 25-30 microns	1280 PPI
Silver gelatin photographic print	Printed grain: 25-30 microns	1280 PPI
Botany pressed specimen	Fern specimen: Cibotium glaucum spores: 42 microns	907 PPI
Orchid living specimen	Orchid pistol hairs: 60 microns	635 PPI
Engraving line (stamps, currency)	Width of engraving line on stamp: 50 microns	762 PPI
Thread stitching (textiles)	Width of thread: 300 microns	127 PPI
Transmissive film grain	Width of film grain: 5.3 microns	4800 PPI

Virtual Workflow Design

The Virtual Workflow includes the steps in a Mass Digitization project after imaging, where the images are checked, validated, and the data of the objects are populated to the collection systems. Like the other workflows, planning and discussions about the way we handle the digital data makes the process run smoothly and reduces the number of errors. Due to the diversity of the collections at the Smithsonian, we try to use similar workflows where possible, while allowing for customized steps where needed.

One of the standard steps in all the Mass Digitization projects is the tracking and verification of the digital files. We track the files using Osprey, a web-based application that displays the status of all the files in a project. The system checks that all the images meet the requirements of the project and stores the results in a database. The dashboard allows the collection staff, project managers, and other stakeholders to see summaries of the results of these checks and the status of each individual image. In addition, the system keeps track of other data-related tasks like data extraction and synchronizing with other systems.

One of the details with more variability between projects that we deal with is the tracking of identifiers. Some collections are fully cataloged while others might not have a single digital record. For either case, we aim to reduce the amount of handling and data entry during a mass digitization project. The time invested in these tasks during the digitization process reduces the number of objects digitized per time, increasing the total cost. When necessary, we have used barcodes and automated methods to gather and track identifiers.

An example of a collection that was fully cataloged was the National Numismatic Collection's Russian coins and medals, Lilly, and Straub collections of the National Museum of American History. The collection did not use barcodes, so we used Virtual Barcodes, an application we wrote for these cases. The vendor searched for the identifier of each object in the system, which had a copy of the collection database. The application displayed a data matrix that encoded the object identifier so that the vendor could scan it to assign the filenames of each object. This process reduced the number of errors since each identifier had to exist in the database, but increased the time required to handle the object. In future projects, we will move this step to the Physical Workflow as part of the preparation of the collections and add permanent or temporary barcodes to each object to make this process faster during digitization.

Another example where we used the Virtual Barcodes was the digitization of about 30,000 carpenter and bumblebee specimens from the National Museum of Natural History. We needed to keep track of the scientific name assigned to each specimen without making this link permanent in the image. Each tray in the collection had the specimens grouped by scientific name. Before digitizing the first specimen of each group, the vendor searched for the scientific name in the application and used the barcode to store the database

identifier for the name in the image metadata. The following specimens were automatically assigned the same identifier until a new group of specimens with a different name was placed in the queue for digitization. Once we got the digital files, we extracted to species identifier from the image metadata and cleared this value from the file using an automated script. This step was required because the museum staff did not want to have identifiers for the scientific name in multiple places since these names can change due to new knowledge from research in the collections. We delivered a data file with the specimen number and the scientific name identifier to the data manager of the collection to load into the collection system. With some preparation, we were able to keep track of the taxonomy of all the specimens in the project in a minimal amount of time.

Another step we have standardized is the use of sampling per lot for quality control (QC) of the images. Due to the size of the collections, running QC of all the images is not a cost-effective option. We implement the ISO 2859-1:1999 standard to inspect the images by samples. As a comparison, the staff of the Smithsonian Gardens ran a parallel QC process in a project where they verified all the 40,900 items digitized. The manual check of all the images only found 323 images (0.79%) that required remediation. The small number of errors make evaluation of all the images too time intensive for the large-scale projects in Mass Digitization. QC by sampling allows us to digitize the large collections at the Smithsonian while maintaining a balance between cost and quality.

The last step that the Virtual Workflows include is the ingestion into the systems of record. We make sure that there is a clear path for the files, from the capture by a vendor to the Digital Asset Management System (DAMS) of the Smithsonian and the Collection Information System (CIS) of each museum. In most cases, the paths of the data are the same. The images are transferred using automated scripts that synchronize the files and the records of the objects between the DAMS and the CIS. This allows the museum to store and publish the digital images of thousands of specimens or objects automatically. The staff can concentrate their time and efforts into the curation of the collection instead of the manual management of the thousands of files generated in a Mass Digitization project.

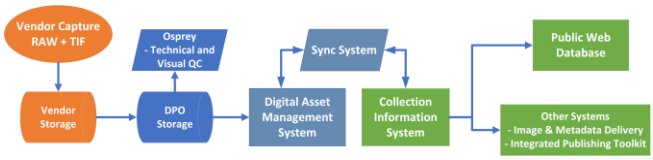

Figure 5. *Summary of the movement of the digital files from capture to the publication in the museum's systems.*

Results

The IDIF analysis results in a project-specific imaging standard by which we measure capture quality for all Smithsonian managed digitization projects. Our process ensures that fine specimen or item details are resolvable.

For the Paleobiology specimens, details in the **80-micron range** were identified by collections staff as the smallest details of interest in the specimen to be captured.

Therefore, the spatial frequency, or PPI (pixels per inch) required to capture the smallest measured detail of eighty microns is **317 PPI.** However, we must also consider phase correction in digital imaging (i.e., pixel elements do not align perfectly with physical detail). To account for that, we apply a 1.5 multiplier to our calculated resolution which results in an optimal resolution of 476 PPI.

Calculation: 25,400 microns/inch ÷ 80 microns measured= 317 PPI x 1.5 phase correction = 476 PPI

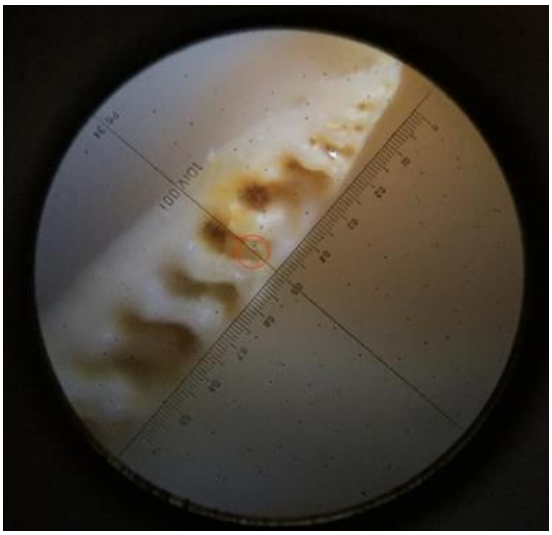

Figure 6a. *Magnified view showing 80-micron detail identifiable of taxodont teeth along the hinge plate of a bivalve specimen.*

Figure 6b. *Image at 100% showing resolvable specimen taxodont teeth along hinge plate.*

Conclusions

As detailed in the Smithsonian Institution's new strategic plan *Smithsonian 2027: Our Shared Future,* the number one area of focus was, "Ensuring a **digitally empowered Institution** with expanded virtual reach." The Digitization Program Office's Collections Digitization Program has a proven history of success with over 5.3 million specimens/objects digitized since 2014 and is uniquely positioned at the Smithsonian to ensure continued stewardship of creating and enriching digital surrogates. DPO Collection Digitization's efficient workflows and infrastructure and our seamless automated integration into IT infrastructure help improve consistency across all our projects. Ultimately this helps

researchers, educators, curators, historians, activists, artists, and students study these items in depth and in high resolution, regardless of their location.

References

[1] Establishing Spatial Resolution Requirements for Digitizing Transmissive Content: A Use Case Approach. Don Williams, Image Science Associates, Rochester, NY USA; Michael Stelmach, Consultant, USA; and Steven Puglia, Library of Congress, Washington, DC USA

[2] Film Grain, Resolution, and Fundamental Film Particles. Tim Vitale, Paper, Photographs & Electronic Media Conservator, Emeryville, CA USA

Author Biography

Nathan Ian Anderson received his B.F.A. in Photography from the Parsons School of Design in New York City and is a cultural heritage professional with over twenty-five years of expertise. As a photographer and program officer with the Smithsonian, Anderson is responsible for the digitization of over 500,000 objects, so he is no stranger to working with varied collections from a rare Tiffany vase to a 30-million-year-old fossil. A professional large format fine art photographer in his spare time, Nathan has participated in several shows nationwide, and his work has been internationally exhibited and collected.

Jeanine Nault serves as the Mass Digitization Team Lead in the Smithsonian Institution Digitization Program Office. She previously managed the digitization program at the National Anthropological Archives (NMNH) and served as the digital asset manager for the Veterans History Project (Library of Congress). Ms. Nault holds undergraduate degrees in English Literature and Anthropology from the University of Michigan and a graduate degree in Museum Studies from the George Washington University.

Luis J. Villanueva Luis joined the DPO in April 2018 to help enrich item-level records across all units of the Smithsonian. During his academic career, which went from searching for tropical frogs in Puerto Rico to listening to temperate and tropical soundscapes, he studied the way data and databases are used to analyze biological systems. Luis developed software packages to allow other researchers to analyze audio files as part of his PhD work at Purdue. From there, he managed and expanded a large spatial database of biodiversity information at Yale that was built from a variety of sources. Luis is now seeking to expand the tools and resources available for the collections to improve the data available on a massive scale.