

Books Are Stubborn Things

Amy McCrory; The Ohio State University Libraries; Columbus, Ohio/USA

Abstract

"Facts are stubborn things, and whatever may be our wishes, our inclinations, or the dictates of our passions, they cannot alter the state of facts and evidence." –John Adams

This paper addresses the complexities of book digitization. In many ways, books defy easy categories that would allow the uncomplicated workflows many people assume are possible. I will examine how this unavoidable fact impacts planning and implementation of regularized book digitization, highlighting common problems and describing how they may be resolved or avoided altogether through careful planning. The paper includes a description of how my organization used a pilot project as an opportunity to build a coordinated workflow encompassing multiple activities across several departments within a large university library.

Introduction

Book digitization is complex, more so than people often assume. Successful digitization projects are built on careful consideration of many possible complicating factors. These include the physical properties and condition of the books themselves; criteria for selection of texts; the completeness of the catalog metadata; management of data generated during work processes; and, within libraries and other institutions, the cross-departmental activity that must be sequenced to produce stable, coordinated, repeatable workflows.

This paper is focused on a category of books that does not have a widely recognized definition. They are often rare, in the sense that very few copies exist in institutional collections, according to records available in consortial catalogs like WorldCat. But they are not identified as rare books in the same way as Renaissance era Bibles or first editions of famous novels. Rather, they are books published primarily in the 19th and 20th centuries and held in the so-called general collections of libraries, circulating among users for years, with some eventually being pulled from the stacks and moved to high capacity closed depositories due to deteriorating condition or the discovery that they are among only a few existing copies. In the library and cultural heritage professions they are often referred to as "medium rare" books. [1] Many defeat a one-size-fits-all approach to digitization due to their physical properties.

Background

Digitization of bound book collections has been ongoing in libraries for decades, with the goal of creating digital repositories of substantial portions of library collections. During this time, efforts such as the Million Books Project, Google Books and the Open Content Alliance, along with in-house digitization in libraries and archives, have produced millions of digital copies, sharing content to a worldwide audience. They have also brought to the surface complications in digitizing bound texts. These problems have been

systematically examined, described and critiqued in various publications, pointing to the need for more awareness of common issues encountered in such projects. In 2013 Paul Conway identified tight bindings and binding-side margins of less than 1cm as the greatest predictors of digital image errors found in a survey of Hathi Trust collections, and cited Google's decision not to digitize foldouts as a significant reason for missing content. [2] Google later revised its practice to include foldouts, though there is an upper limit to the size of paper media they will scan. [3]

Book Digitization at OSUL

The Ohio State University Libraries (OSUL) have conducted book digitization projects since the late 1990s, starting with outsourced work and later adding inhouse digitization queues. OSUL is a Google Books partner, with titles from its collections deposited in Hathi Trust. Among the books Google identifies for digitization based on catalog metadata, some are excluded due to fragile condition or page dimensions that exceed the size Google will accept. OSUL retains records of titles identified but not digitized by Google; these records form a base from which OSUL subject librarians and collections curators may select groups of titles to be included in digitization project proposals. Criteria for selection have included specific languages as well as subject areas identified for their research value. Books digitized in-house are uploaded to the Internet Archive, and are copied over to Hathi Trust via a script run every month.

Metadata Issues

During these projects, several issues surfaced. These included the reality of catalog records created over the span of many decades. Some were incomplete with regard to present day requirements for minimal MARC metadata elements set by consortial digital repositories. Hathi Trust, for instance, provides a publicly available list of required fields. [4] If these are not present, the corresponding digitized texts will be rejected and a message requesting remediation sent to the contributing institution. This leads to unscheduled remediation activity which is inefficient for the Cataloging department. Other issues discovered included older records containing English-only versions of titles, authors and imprints for Arabic-language books, which, had they not been remediated with the addition of Arabic script, might have made the online texts less likely to be discovered by international audiences.

Materiality: physical complexity of books

Prior to image capture of books selected for digitization projects, OSUL's Digitization and Conservation departments meet for a physical review of the materials. The reviews typically identify the following:

- Issues such as fragile paper, broken bindings, or other deteriorated condition necessitating stabilization of the material by conservators prior to digitization.
- Books encased in plastic wrap ("shrink wrapped books"). In a previous era, a common method in some libraries for housing disbound brittle books was to wrap the separated

pages in plastic to keep them together. Removing the brittle pages from the wrap requires care, as does digitizing them. Afterward, Conservation makes new enclosures for them out of acid free board.

- Tight bindings, particularly in the case of narrow margins on the binding sides of pages.
- A high page count, as this may mean full page image capture will be difficult due to the sheer bulk of the bound paper.
- Books larger than 21 x 29 cm (A4 paper), the largest size accommodated by our most-used inhouse book scanners.
- The presence of foldout or pocket maps, charts, and other oversized illustrative material. These books may be digitized on two types of equipment, a standard scanner for the regular run of pages and a large format scanner for oversized material.
- Books that are extremely fragile; books with color-critical illustrations.

These criteria determine which books will be routed to Conservation for treatment, either before or after they are digitized, or both. This review process has been conducted between the Conservation and Digitization departments for many years and so has become habitual and rather organic, but it is not always transparent to colleagues who don't work in either of these departments. In particular, the iterative nature of the work has not been well understood. Items may pass back and forth between the two areas more than once, as they may need, for instance, stabilization prior to digitization and rehousing after.

During the same review of physical condition, Digitization staff make decisions about which image capture equipment to use, based on the dimensions of books and any foldout material, as well as their robustness or fragility and the presence of any color-critical elements. The choices usually fall into the following categories:

- Internet Archive TableTop Scribe scanner: Books 21 x 29 cm or smaller; books that are robust or moderately fragile.
- Bookeye 5 V1A large format book scanner: Books between 21 x 29 cm and 63 x 91 cm; books that are too fragile for the Scribe's glass platens; oversized foldout or pocket media from a 21 x 29 or smaller book.
- Phase One iXG: Extremely fragile books; books with color-critical illustrations.
- Context HD Ultra Roller Scanner: Pocket media that exceed 63 x 91 cm in size.

Choosing and standardizing the correct software settings are important too, with decisions made toward even lighting across pages and a linear or near-linear curve, usually with black point adjustments for readability of the printed text.

Finally, physical attributes of books have an impact on post-processing work, particularly compliance with digital accessibility requirements. Faded, broken or thinly printed text can reduce the

success of automated OCR, while eclectic page layouts challenge automated tagging of reading order.

Successful models for book digitization require a thorough understanding of the materiality of books and the need to specify variations in practice in accordance with commonly encountered conditions. Planning of projects and workflows should include the input of digitization specialists who do the hands-on work. Otherwise, projects will be designed to fit the assumptions of non-specialists who do not engage with the reality of books as a specific form of media.

Book Digitization Strategic Initiative

In 2019 OSUL's Assistant Dean for Content & Access began discussions toward forming a more strategic approach to selecting items for digitization. Subsequent discussions shaped the idea further and it was ultimately approved as an initiative, Leveraging Book Digitization for Strategic Collection Building. In 2022 OSUL's Collection Strategist and Preservation & Digitization Strategist worked together to proceed with the initiative, outlining its requirements and parameters and identifying project stakeholders. The pilot began in 2023 and is underway now, with completion expected later this year. The description that follows is a preliminary overview of a project whose final report has yet to be issued.

Implementation of the Pilot Project

In June 2023 representatives from departments involved in the pilot project met to plan the work and write a project scope and charter, which was circulated for review, edited, and signed in August. Data analytics and item-in-hand reviews were conducted in October and November. Conservation and Digitization activity began after this, with the flow of books staggered as some copyright and condition decisions were considered and resolved among the departments. These issues slowed the onset of digitization, which in a project whose only aim is quick output of digital content would be viewed as a problem. Since this was a pilot, they were welcomed as an opportunity for all project partners to gain understanding of the complexities involved in working on "medium rare" book projects.

Criteria for selection of items to be digitized

Criteria were written for inclusion of books in the pilot project. Catalog records were filtered according to these data points:

1. Printed books in the public domain, that is, works published more than 75 years ago.
2. Books from the General Collections located in the University Libraries' Book Depository (books held in the library's Special Collections were determined out of scope for the pilot).
3. English language books.
4. Books whose catalog records contain the subject heading "Ohio" in the MARC 600 field.
5. Books not already available in the Hathi Trust collection of digitized texts.

Starting with the list of titles excluded from Google's scanning process due to fragile condition or oversized format, Collections Strategy generated a report from the library's catalog, applying filters that matched the criteria for items 1-4 listed above. The resulting records were exported into an Excel file. This was merged with a list of titles already held in Hathi Trust that overlapped with the filtered records. The output added a column titled "In Hathi" with yes/no designations. The final Analytics Report containing books to be digitized in the pilot project included only the 246 titles with a "no" value in that column.

Complications in the process were minimal, with one exception: MARC records allow for publication dates to appear in more than one possible field, in some cases in conjunction with other values. In order to have a single, sortable and searchable date field for each record, some data cleanup was required. It was then possible to exclude items not yet in the public domain from the selection list.

A script was also run to check for the presence of data fields specified in the MARC 21 Minimum Level Requirements which Hathi Trust requires for all submitted bibliographic records. If any required field lacked data, this would be noted and the item would be sent to Cataloging for remediation. This eliminated the problem of missing content causing digital files to be rejected from ingest.

Further possible exclusions from the project included:

- Books that were reprints of earlier titles.
- Derivative works, such as translations.
- Books containing third party materials as inserts. Determining this required manual item-in-hand reviews by library staff.

All of the above conditions presented the possibility of additional or unknown rightsholders. Such items were sent to Copyright Services for additional review and a decision about whether to include or exclude them.

Manual reviews

Existing Conservation and Digitization manual review workflows were incorporated into the pilot project. In addition to identifying Conservation work to be done on the items and guiding choices about digitization equipment selection, these reviews allowed special cases to be detected. An example is multipart books referred to as "bound-withs" in cataloging practice. "A bound-with book is a single print volume... comprised of two or more separately created works that were bound together after publication," often by an individual collector or scholar, and "their uniqueness, rarity, and teachability give them special significance, particularly in academic libraries and special collections." [5] They are often distinguished by the varying colors, weights and textures of the paper on which the different titles are printed. A book like this may have originally been cataloged as a single title without an analytic record, obviating the additional titles between the covers. When it enters a library's digitization workflow there is an opportunity for catalogers to create more detailed records, making the full contents findable, so long as they are given the chance to teach digitization staff how to identify bound-withs.

Workflow development

To recap, seven departments in disparate locations worked on the various phases of the pilot project. A workflow plan was needed in order to make departments' work transparent and understandable to one another. The tasks to be built into the pilot workflow included:

- Materials selection (In this case, pilot project partners in lieu of subject librarians and curators)
- Inventory creation via data analytics (Collections Strategy)
- Retrieval, logging and transport of materials from the Book Depository (Depository Services, or "STX")
- Rights evaluations (Copyright Services)
- Manual review for condition (Conservation and Digitization)
- MARC record remediation (Cataloging)
- Stabilization and rehousing of fragile items (Conservation)
- Image capture, processing and uploading to Internet Archive (Digitization)
- Transfer of files from Internet Archive to Hathi Trust (Digital Initiatives)

The Digitization department developed a visual workflow plan with input from the project partners. After trying a few approaches to diagramming the flow, it was decided a swimlane chart would be the best fit. A swimlane configuration shows which department is responsible for each task or set of tasks, useful when the workflow is spread across multiple departments, and allows for some iterative process cycles rather than an entirely linear path. Throughout the pilot, adjustments were made as we conducted the work and discovered processes that needed revision. The pilot gave visibility to the overlapping nature of much of the work and gave project partners the opportunity to think about where to divide and where to consolidate processes. At the time of this writing, the workflow shown below is still in draft form, awaiting final edits.

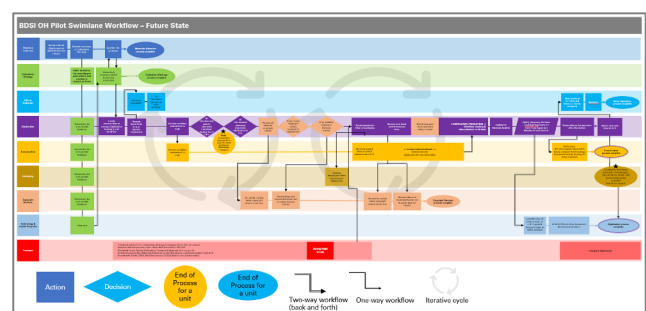


Figure 1: Draft of the book digitization workflow plan (not yet finalized). Credit: Madeleine Fix and Sarah Collier.

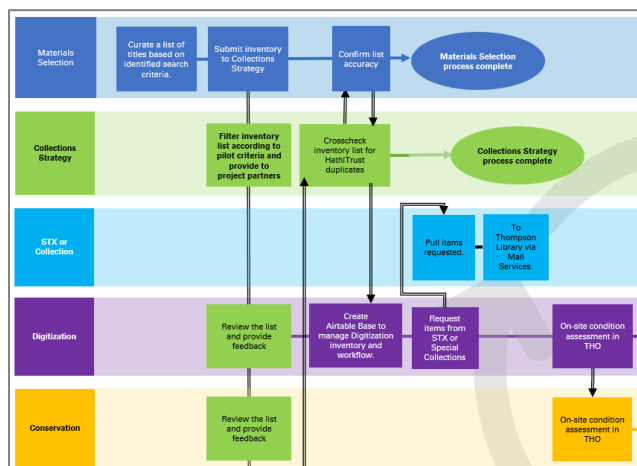


Figure 2: Detail of the draft workflow plan.

Data management

Tracking the work of book digitization requires the capacity to store and manage multiple types of data at the item level, including, among others: identifying information about the items taken from MARC records; details about physical features of the books that digitization specialists need to know; the current status of work being done; equipment used for capture; number of pages digitized; URLs of the digitized copies; and the current location of each of the books, which are transported to different locations during the life of the project, with library staff responsible for knowing where they are and ensuring they are safely returned to storage.

The Book Digitization Associate created an Airtable base for the project, importing the list of 246 selected titles. The software enables users to assign category tags to processes, equipment, work status, department names and other recurring data values, and to build authority lists of these values, creating a standardized vocabulary to be used in the project. User-friendly views of the data can be generated and customized for colleagues who want a snapshot of the project's status. We found these features especially useful in the context of a large organization, where regular communications and updates can be challenging. The data will be used to assess the pilot project and make projections for future efforts involving these types of books.

A Title	Location	Status	Camera
Atlas of the city of Dayton, Ohio : and adjoining territory. Compiled from o...	Vault/Storage	Awaiting IXG	IXG
Boiler inspection law and rules formulated Board of boiler rules ... for the ...	Conservation	Cons Pre-Scan	Scribe LTC
Report of the Commission Appointed to Inquire Into the System of Screen...	Thompson	Awaiting Return	Scribe THO
Report of E. McMillin showing quantity and kind of minerals along the line...	Conservation	Cons Post Scan	Scribe LTC
Rules of practice before the Commission in cases for claims, charges or de...	Conservation	Cons Post Scan	Scribe LTC
Logan County and Bellefontaine, Ohio, school survey	Thompson	Awaiting Return	Scribe THO
The tax laws of Ohio as codified and revised for 1915. General code. (Marc...	Thompson	Awaiting Return	Scribe THO
Report of the Investigating commission appointed to enquire into the cau...	Digitization	Uploading	Bookeye Scribe LTC
Summary of the climatological data for the United States : by sections ... s...	Digitization	Post Processing	Bookeye
Fundamental facts for new citizens	Digitization	Post Processing	Bookeye
Official circular issued to boards of education in Ohio : by the state commi...	Conservation	Cons Post Scan	Scribe LTC
Catalogue of the geological specimens, collected, on the late survey of the...	Digitization	Post Processing	Bookeye
A report to the Director of Public Improvements of the City of Columbus, ...	Digitization	Uploading	Bookeye Scribe LTC

Figure 3: Detail of the workflow tracker showing selected fields in the Airtable base.

Conclusion and looking ahead

At the date of this writing, work is nearing completion, with 28,047 pages digitized so far. Post-digitization conservation treatment and rehousing of books continues, and all completed items have been returned to the Book Depository. A project close report will be written and shared with OSUL's administration who will assess the project and decide whether to approve the workflow as a regular initiative going forward.

Issues to be considered in possible future iterations of this process include:

- Deciding how to approach titles that are eliminated from a given project because they are not yet in the public domain, but which become eligible in successive years. On what schedule should we revisit titles and how do we stage them for digitization without disrupting other projects underway at that point?
- Are there categories of published works whose catalog records should always be considered for a review of possibly obsolete or incomplete descriptions prior to digitization?

Overall, the pilot was useful in bringing transparency and a broader understanding to issues involved in selecting, assessing, treating and digitizing "medium rare" books and adding them to consortial repositories. The workflow plan and project documentation will, I hope, facilitate cross-departmental communications and work in future projects.

References

- [1] Q. Ferris, "Medium Rare: An Innovative Treatment Approach to the Space Between Special and General Collections." At the 45th Annual Meeting, AIC—Book and Paper Session, 2017. <https://resources.culturalheritage.org/conservators-converse/2017/06/23/45th-annual-meeting-book-and-paper-session-may-30-medium-rare-an-innovative-treatment-approach-to-the-space-between-special-and-general-collections-by-quinn-ferris/>
- [2] P. Conway, "Preserving Imperfection: Assessing the Incidence of Digital Imaging Error in Hathi Trust," in *Digital Technology and Culture*, 2013, 42(1), 17-30.
- [3] Mail your books for scanning - Google Play Books Partner Center Help. <https://support.google.com/books/partner/answer/3297525?hl=en#guidelines>
- [4] Hathi Trust Bibliographic Metadata Specifications. <https://www.hathitrust.org/wp-content/uploads/2023/07/Bibliographic-Metadata-Specifications.pdf>. Accessed on 5/14/2025.
- [5] Floyd, D., (2024) "Making Decisions When Cataloging Bound-with Books", *NASIG Proceedings* 38. doi: <https://doi.org/10.3998/nasig.6735>

Author biography

Amy McCrory received her BA in Art from Northwestern University (1986) and her MLS from Clark Atlanta University (1995). She currently works as Associate Director, Digitization at the Ohio State University Libraries, after working as their Digital Imaging Specialist from 2006 to 2017 and Program Manager for Digitization from 2018 to 2024.