Strategies for Implementing a Mass Digitization Program [i]

Erik Moore

University of Minnesota

Introduction

In 2007, OCLC published the report Shifting Gears: Gearing Up to Get Into the Flow to bring to the forefront a much needed conversation about digitization of archival collections, and access to the rich content accessible only through paper or other analog formats.[ii] The authors emphasized that any successful large digitization program would focus on access and quantity. They challenged archivists to rethink policies, procedures, and technologies that either slowed the process of mass digitization, or were unfriendly to the implementation of a rapid capture program. Recent articles, blog posts, and columns demonstrate that we as a profession continue to grapple with ways to implement digitization programs that are both sustainable and efficient.[iii] The strategies offered in this paper highlight a practical program for the mass digitization of organizational archival records using a rapid capture process that is replicable regardless of the size or resources of the repository. It will review the establishment of a rapid capture workflow at the University of Minnesota Archives; provide details on how it functions, including equipment information, scanner settings, and workflow procedures; explain the selection process for scanning; describe how it has helped to create inreach opportunities; and finally, examine how it has changed not only daily operations, but the perspective on what it means to provide broad access to the collections.

In 2008, the University of Minnesota Archives developed a low-cost, in-house solution for routine mass digitization of university publications, reports, and records. This programmatic effort facilitated access to the rich history found in the content of press releases, self-surveys, course bulletins, minutes, and more. At its core, the in-house scanning effort represents a recovery of information already in the archives, and a further commitment to its on-going preservation and use. The program incorporated recommendations from the National Archives and Records Administration (NARA) and the previously mentioned OCLC report, Shifting Gears. These reports advised integrated digitization activities provided the best means to achieve large-scale conversion of analog materials to digital formats for online access.[iv]

In order to be both successful and sustainable, the program recognized several decision points to better integrate the digitization activities. First, archives staff identified campus partners, both within the University Libraries organization, as well as across campus, that would be key to the program's realization. The process of establishing the scanning program as an in reach activity allowed staff to better position the work as satisfying a broader institutional need rather than a side activity. Second, staff determined equipment and technology needs that met several requirements, including low-cost, replaceable parts (consumables), low-barrier for use, speed of scanning, and considerations for how the technology treated the archival materials. Archives staff then set parameters for the selection and description of the content to be scanned. Preference is given to entire collections, series, or volumes that do not require item level review. Description is minimal and leverages existing metadata when available. Next, staff determined how to handle the archival materials both before and after scanning. This included defaults for scanning quality and formats. Since the scanning operations are considered destructive in nature (i.e., bindings or other permanent fasteners are removed), staff established a decision-making process on how to handle the materials after reformatting. Finally, staff determined the primary access point for the digitized material would be through the institutional repository; however, depending on available resources, access systems may vary and can employ a wide-range of options as discussed below.

Rapid Capture in Practice

The two basic goals of a mass digitization program are to scan what is useful and make delivery the objective. These two goals also define successful scan-on-demand programs. The primary difference between the two programs is that in a mass digitization workflow, selection is based on a longitudinal understanding of commonly used resources that would be appropriate for digitization, whereas digitization on demand relies mainly on user requests. It is helpful to distinguish between the two by thinking of the former as scanning in-demand materials rather than requested or on-demand; however, implementing either program does not negate or inhibit the use of the other.

Due to the similarities in the programs, an adopter of a scan-in-demand approach can leverage the workflows already provided for on-demand programs as a starting point. In 2011, OCLC produced a report that provided useful guides for on-demand workflows.[v] In the report, the authors provide a three track matrix on how to process an on-demand digitization request. The inside, middle, and outside tracks represent the breadth of complexity that may or may not be applied to creating a digital surrogate. The report encourages those implementing the matrix to jump between tracks when necessary to remain within institutional practices or confines.[vi] This type of adaptation encourages both adoption of the practice and makes it more sustainable.

The authors also remark on the second goal of the program – access – noting that “in the context of reduced resources and shifting user expectations of online access, a quick and easy way to deliver requested digital reproductions has become an imperative. User requests must not be bogged down by fine-tuning images and metadata.”[vii] Any type of rapid digitization service, whether on-demand or in-demand, should provide a digital surrogate that satisfies the user’s need. It should not be over thought, over described, or over scanned.

At the University of Minnesota Archives the practice of rapid capture generally employs the following model. A single staff member identifies or selects the material for scanning. This is based on the staff member's familiarity with the content, or understanding of how it provides for a known user need. Selection is generally done at the series level. Scanning is done to a set of preset standards. These standards might be considered the lowest common denominator, unless particular aspects of the material require a change to the default scanner configurations. Bound volumes are scanned as a single digital object; likewise, a single folder of content is also reformatted as a single digital file. Bundling also occurs by logical groupings by year. For example, a set of press releases are scanned as a single object based on the calendar year associated with the news releases. All description created for the digital files is minimal. Descriptive elements are Dublin Core based and include author/creator, title, date, and type. Keywords, descriptions, or other information not pertinent to the content is generally not applied. The quality control focuses on trusting archives staff to follow the existing guidelines and procedures. This includes following several checklists for pre and post scanning operations. If a problem occurs, there are also steps to follow for checking in or getting advice from more experienced staff. There is no separate quality review outside of these steps. Finally, delivery is through the institutional repository. Although not all institutions or archives have access to such a system, alternatives are available and discussed below. In comparison to the tracks of service provided in the Scan and Deliver report, the University of Minnesota Archives rapid capture workflow includes three inside tracks (selection, scanning, and resolution), one middle track (metadata), and two outside tracks (large-scale and delivery).[viii]

The remaining provides more details about the process and strategies of establishing a mass digitization program, including the selection of equipment, settings and standards, workflow practices, staffing needs, and points of access. It is specific to the program established at the University of Minnesota Archives, but outlines key decisions, choices made during the process, and alternative options to allow for adaptation to localized needs or resources.

Selection

Initial selection for a mass digitization program should focus on materials that are information-rich. These materials can be identified by several factors. First, there is a known and documented sustained use of the materials by users either on-site or through off-site reference services. Frequently requested photocopying or scanning of the materials is a good indication. Second, the materials requested represent, or are part of, an entire series or collection rather than unique individual items. In many cases, the materials that answer common reference inquiries are serial publications or archival document types such as annual reports, minutes, or bulletins. Identifying the larger series for these materials is a good way to predict needs going forward and satisfy an identified use. Additionally, serial publications tend to be broadly distributed, resulting in two benefits for digitization programs; either researchers tend to look for materials that have been seen or referenced before, or there are likely extra copies available for destructive digitization. Finally, the focus on series, or entire collection scanning, removes an element of selection that can slow down the digitization process. This ties rapid capture scanning to our intellectual practices of arrangement and description at the series level.

In addition, selection practices should be aware of potential copyright issues associated with the materials. For institutional archives, a focus on institutionally produced or published materials can help identify available materials where copyright is held by the institution. For personal papers collections or manuscript repositories, a more careful review of donor agreements or intentions may be needed to determine if there is a risk tolerance in making digitized material freely available.

Finally, paper-based document types or serial publications that are now produced and/or accessioned in born-digital formats make good digitization candidates. Digitization allows for the creation of a single run of the materials to be available in a single location, regardless of the fact of whether the materials were born-digital or digitized. This is an inreach activity that the archives can promote and gain stakeholder buy-in as they build the program.

At the University of Minnesota Archives the material selected for scanning is largely 20th century, mass produced and distributed, and published by the University. It is considered informational in value with no artifactual or intrinsic value. The early emphasis for selection was on ready reference materials available to archives staff or in-person researchers. These materials included bound volumes of Board of Regents' minutes, minutes of Senate committees, annual reports of the University, course catalogs, and course schedules. These materials were regularly used to answer walk-in, phone, and email questions. As the program developed, archives staff identified materials that were information-rich, but less commonly used due to poor access points. Examples of these materials include press releases dating back to the 1920s, staff and alumni magazines and newsletters, and departmental self-surveys and program histories. All of these materials were uncataloged or otherwise not described in online systems. The mass digitization of these materials not only made them available for the first time through an online descriptive access point, but made them available in their entirety.

Equipment

In order to establish a scanning program, it is necessary to identify an affordable office document scanner that satisfies several basic needs and many preferred requirements. First, archives staff identified a need for equipment that was able to adjust to the size of the original paper. Most economy desktop scanners with an automatic document feeder (ADF) can take a legal sheet as the maximum feed size

(8.5×14 in.). This size will satisfy most modern office paper sizes. Second, the paper should feed flat through the ADF and not curve around the light bar. Third, the scanner should allow for duplex scanning with a single pass (no retracting/re-feeding). The latter two requirements reduce the risk of item crumpling or tearing during the scanning process. Additional considerations for identifying an appropriate scanner include understanding how many scans in the machine’s lifetime the equipment can handle, and whether or not that will serve the program’s needs. A preference should be given to machines that have replacement parts in order to maximize the equipment’s lifespan. Finally, since not all materials are well-suited for an ADF, desktop machines that include a built in flatbed scanner, allow for more flexibility during the course of scanning, including an option to scan delicate mixed materials (e.g., photographs, onion skin paper, etc.) in the same workflow as the rapid capture process.[ix]

In 2008, the Fujitsu fi-6230 was the only scanner that met all basic and preferred requirements and was available for less than $1,400.[x]

Scanner Settings

The scan quality and file format outputs should emphasize speed and access. In this example, settings for all scanning meet a certain threshold but could be adjusted easily if an item or set of materials required configuration changes outside of the presets. Documentation for scanning operators includes the following guidance for selecting appropriate settings.

Bitonal (Black & White) – This is used for most typed or handwritten documents. This provides the cleanest image for contrast and printing. Resolution may be adjusted to a higher ppi depending on the type size to improve readability and Optical Character Recognition (OCR).

Resolution	300×300/400×400
Image Mode	Black & White
Black/White	Static Threshold
Brightness	128
Threshold	128
Contrast	128

Grayscale (8 bit) – This is the appropriate setting for color documents where the shades of colors are drastically different, but the colors themselves are not essential to the context or readability of the content.[xi] This setting will show the white paper as pale gray.

Resolution	300×300
Image Mode	Grayscale
Brightness	128
Contrast	128

Color (24 bit) – This setting is only used in cases where there are many colors and they are essential for the reader to be able to understand the document or to read text overlaid on a color background. A color document is much larger and slower to scan than a Black & White or Grayscale document.

Resolution	300×300
Image Mode	24bit Color
Brightness	128
Contrast	128

Individual master files of the scanned materials are not created. Instead, the digital surrogates are saved directly as a PDF/A file format.[xii]

Digitization Workflow

Since the digitization process relies on sheet fed scanning, all materials need to be loose-leaf and free of all fasteners. The process is generally termed destructive scanning since the scanning workflow is based on ability to sheet feed the materials through an automatic document feeder (ADF). The bindings of bound materials are cut and the edges are shaved to run loose through the ADF. Some items can be unbound on site with a paper cutter while others need to be sent to a bindery service that will remove the materials from the binding and shave off any remaining glue or fiber.[xiii] Once unbound and scanning is complete, items are not re-bound. Items are either placed in folders and boxed, or recycled.

Before scanning begins, the materials are prepared by removing all staples, paper clips, or other fasteners. Items that had a binding or staples removed should also be checked to ensure that all pages are loose and edges are smooth. This can be done by the scanning operator or by another individual given the task of prepping the materials.

Next, the operator checks the paper for size and color irregularities. If there are smaller pieces of paper mixed in with letter size paper it is centered in the ADF prior to scanning. If there are multiple colors then a decision is made regarding the best scanning setting to use.

Individual scanning projects may use a name rule to produce similarly named files that include incremental changes with a prefix or suffix. This is often helpful for projects that involve scanning folders of archival material that are assigned an identifier such as a folder number. For projects that require unique file names for each completed file, a name rule can be disregarded and the operator can add the file name when finished.

At the end of a scanning session, all files receive OCR as a batch operation. For this process, Adobe Acrobat Pro or other software allows for the processing of multiple files. Depending on the quantity of files needing text recognition, it can be a good practice to let this run on a computer that is not in use, or can be left to run over night.

Estimating an average for the output of scanned materials is difficult to determine. Scanning times vary due to the effort needed to address different paper size, fastener removal, and paper color configurations. It is also contingent upon the experience of the scanning operator. The highest recorded rate of scanning at the University of Minnesota Archives was 960 individual pages per hour. A more realistic average is 500-600 pages per hour.

Access

A mass digitization program should also include a means to provide delivery of the digitized content. The means of delivery may take different forms, and will likely be dependent on local resources. Leveraging any existing delivery platforms will aid in adoption of the program and is likely to be the least expensive option. Goals for a delivery platform should include a simple process for staff to upload, and ease of access for your target audience or general users. If possible, use existing description or other metadata to make access and delivery less time consuming to prepare. One example is to link PDF files of scanned materials to the box or folder level description in an online finding aid. This can be done through the use of the Digital Archival Object (DAO) tag in Encoded Archival Description (EAD), or through simple HTML linking in a non-EAD online finding aid. Other options for delivery include expanding the use of an image based repository (e.g., CONTENTdm) to include PDF formatted materials, or to use online exhibit software such as Omeka, or a content management system like WordPress to create an online repository. Likewise, if your institution has an institutional repository in place, consider using the platform as a delivery mechanism. The University of Minnesota Archives makes the majority of its scanned material available through the University Digital Conservancy, the University's institutional repository. The Digital Conservancy serves as the "digital arm" of the University Archives and provides a home for administrative digital content, including official organizational records and publications produced by the University. Most of the scanned content is complimentary in nature.

Staffing

Implementing a mass digitization program requires thoughtful changes to daily operating procedures, and may require either additional staffing or reconfiguration of existing staff and/or duties. Installing scanning software and drivers, as well as creating presets and default configurations, requires an intermediary knowledge of software installation and computer systems. This level of knowledge may be available either through existing staff or IT support. Familiarity with software and computer configurations is not required of the scanning operators. Most desktop scanning operations are repetitive, require entry-level expertise, and can be learned on the job. Having detailed workflows, default scanner settings, and introductory training will ensure consistent quality, even if there is a regular turnover rate for scanning operators.

At the University of Minnesota Archives, a majority of the scanning work is completed by undergraduate student workers. The Archives traditionally employed student workers to provide assistance in collection re-boxing, basic processing, shelving and retrieval, photocopying, and newspaper clipping. It seemed evident that clipping daily newspapers and press releases for vertical files became less effective, especially in light of the propensity for staff and users to use Google to answer basic questions. Likewise, due to the types of material that were selected for digitization, it was expected that the number of photocopy requests would diminish as more material would be available in PDF format. Routine clipping of newspaper sources ceased and these hours, along with other duties, were redirected to rapid capture scanning.

Preservation and Storage

It is important to make decisions regarding the storage and preservation of the digital surrogates. Although these are designed to be access files, it is probably a wise decision to arrange for the digital files to be backed up or available for replacement, if there is a system loss, or file corruption. Again, leveraging existing file back-up systems or replicated storage options are the best in order to work within in the normal operations of your institution and create as little overhead as possible for the program. If the materials are uploaded to an online exhibit service or institutional repository, it is possible these systems or agreements come with certain assurances regarding file preservation. Identifying existing services may lessen the need to create a separate preservation storage environment.

Outcomes of the Digitization Program

Since implementation of the program, approximately one million digitized pages are now accessible through the open access, full-text searchable institutional repository at the University of Minnesota. This enhances the Archives' capability to serve external audiences and to provide internal support. It offers opportunities to reach new potential users, and mitigate the obsolescence of material in the collections by closing the gap between analog and digital, discovery and access.

Rapid capture scanning, once adapted for local implementation, changes the nature of archival reference services – ease of access, improved discoverability, and placement within the user's process, not our own, are the benefits. Statistics show there are thousands of downloads per month of the digitized content. If each download were compared to folders in the reading room, it becomes clear that the availability of the scanned content is satisfying user needs that are happening elsewhere. It is providing instant satisfaction without the need to always interact with the archives. The vast majority of these users do not contact us for follow-up questions or seek to see the "originals." They find what they need and move on, staying within the flow of their own research.[xiv]

Additionally, materials reformatted and made available as full-text searchable documents reveal that there is greater informational value as a digital format. Keyword searching across a broad range of documents and publications allows for users to identify sources of information that would not be available through traditional indexing or description work. Examples include the ability to identify the first mention of a name, phrase, or event, or to corroborate a piece of information by searching across unrelated archival sources at the same time. Reformatting into a digital format releases this hidden information from its single access point – the physical archives – and allows for greater functionality. This is the recovery of information through mass digitization.

And, if digitization enables us to close the gap between discovery and delivery for external audiences, it stands to reason that internal partners benefit too. New partnerships with university offices originated from the University Archives' capacity to pair their born-digital records with digitized content of hardcopy archival holdings in a manner that is understandable, transparent, and serves their needs. The mass digitization program became an inreach opportunity to better support university departments and offices by providing them with full access to their historical records. Examples of this include digitizing full runs of Board of Regents' minutes, campus-wide newsletters for University Relations, and a report series for the University Senate office. With full access to past records and reports, these offices gained a first-hand understanding of the role of the University Archives and the direct benefit the Archives provide to campus. The result is more offices now understand the value of permanently retaining their born-digital materials and trusting the University Archives to preserve and make them available. The Archives became a sought after partner on campus, bringing value and solutions to the table, by digitizing hundreds of linear feet of paper material, and increasing the efficiency with which it could capture historically significant born-digital information that is otherwise difficult to acquire.

Conclusion

The outcome of these strategies for implementing a mass digitization program within normal archival operations, demonstrate that when discovery and delivery coincide it has a major impact on our ability to provide online reference services because “discovery happens elsewhere,” and that archives are not the sole distributor of our content.[xv] It allows for archivists and users to interact with collections in new ways, find new uses for old sources, and provide instant satisfaction for researchers and internal partners alike. The digitally reformatted materials provide user access and portability via search engines, printers, downloads, tablets, and smartphones. It does all of this in a manner that is sustainable and applicable to many archival programs.

About the Author

Erik A. Moore is the university archivist for the University of Minnesota Archives. Moore is also co-director of the University Digital Conservancy, the University of Minnesota’s institutional repository. He has advanced degrees in Library & Information Sciences and Historical Studies. Prior to his current position, Moore served as the Assistant University Archivist & Lead Archivist for Health Sciences; as the Archivist/History Project Director for the Academic Health Center; and as the Digital Program Associate at the Immigration History Research Center, all at the University of Minnesota.

[i] Program highlighted in Erway, Ricky. Rapid Capture: Faster Throughput in Digitization of Special Collections. Dublin, Ohio: OCLC Research (2011). Available at: http://www.oclc.org/research/publications/library/2011/2011-04.pdf.

[ii] Erway, Ricky, and Jennifer Schaffner. Shifting Gears: Gearing Up to Get into the Flow. Report produced by OCLC Programs and Research (2007). Available at: http://www.oclc.org/programs/publications/reports/2007-02.pdf.

[iii] See Larisa K. Miller. "All Text Considered: A Perspective on Mass Digitizing and Archival Processing." The American Archivist 76 (Fall/Winter 2013): 521-541. Jeanne Kramer-Smyth. "Digitization Quality vs Quantity: An Exercise in Fortune Telling." Spellbound Blog (Posted 31 March 2012). Available at: http://www.spellboundblog.com/2012/03/31/digitization-quality-vs-quantity-an-exercise-in-fortune-telling/. David S. Ferriero. "From the U.S. Archivist: Scanning the Past to Make Access Happen." Archival Outlook (July/August 2014): 26-28.

[iv] National Archives and Records Administration. Draft Plan for Digitizing Archival Materials for Public Access. (September 2007). Available at: http://www.archives.gov/digitization/plan.html. Erway & Schaffner, 5.

[v] Schaffner, Jennifer, Francine Snyder, and Shannon Supple. "Scan and Deliver: Managing User-initiated Digitization in Special Collections and Archives." Dublin, OH: OCLC Research (2011). Available at: http://www.oclc.org/research/publications/library/2011/2011-05.pdf.

[vi] Ibid., page 9.

[vii] Ibid., page 6.

[viii] Ibid., page, 8.

[ix] For additional information on rapid capture scanning programs for mixed materials (e.g., manuscripts, notebooks) or audiovisual materials (e.g., cassette tapes, nitrate negatives, microfilm), see Erway, Rapid Capture (2011).

[x] The Fujitsu fi-6230 is no longer available; however Fujitsu has more recent models available that satisfy these same requirements.

[xi] A 2009 study showed there were no benefits to scanning in grayscale rather than bitonal in terms of the success with OCR. See Tracy Powell and Gordon Paynter, “Going Grey? Comparing the OCR Accuracy Levels of Bitonal and Greyscale Images.” D-Lib Magazine (March/April 2009). doi: 10.1045/march2009-powell.

[xii] See http://www.digitalpreservation.gov/formats/fdd/fdd000125.shtml.

[xiii] This service incurs a minimal cost and adds additional time to the process. A workflow established for binder removal takes approximately two weeks to send a bin and have it returned.

[xiv] For a discussion about being in the user’s environment, see: Lorcan Dempsey. “In the Flow.” Lorcan Dempsey’s Weblog On Libraries, Services, and Networks (Posted 24 June 2005). Available at http://orweblog.oclc.org/archives/000688.html.

[xv] Erway & Schaffner, 7. For a discussion on the expectation of discovery and delivery to coincide, see Hanson, Cody, Heather Hessel, et al. Discoverability Phase 1 Final Report. University of Minnesota Libraries (2009). Available at: http://purl.umn.edu/48258.