All Text Considered: A Perspective on Mass Digitizing and Archival Processing[1]

Larisa K. Miller

ABSTRACT

Amid stagnant or diminishing resources, archival repositories are increasingly expected to digitize entire collections for the Internet. Yet many cannot eliminate their backlogs of unprocessed collections, let alone digitize and post them online. This article explores the idea of coupling robust collection-level descriptions to mass digitization and optical character recognition to provide full-text search of unprocessed and backlogged modern collections, bypassing archival processing and the creation of finding aids.

KEY WORDS

Archival Theory and Principles, Online Collections, Processing, Technology

With the shift from boutique to bulk digitization, archivists face even greater expectations. Creators and donors of collections increasingly demand digitization as a condition of transfer. Users want collections to be fully available online, as do many funders. Our explanations of copyright and privacy gain little traction when pocket-sized technology and social media have democratized the Internet, and researchers who swiftly snap away with digital cameras in reading rooms do not buy our protests about the high cost of digitization. Most of us have not eliminated processing backlogs, let alone met the rising expectations of the digital age.

We are not deaf to these demands and we strive to address them. We want our collections to reach the widest possible audience, so we are shifting from item-level digitization projects to folder-level ones. This not only cuts metadata costs, it ensures that a document retains its context within a folder, where related documents may shed light on it. Only collections that have been processed benefit from this approach, however, and the cost of archival processing remains a significant hurdle. There is the possibility of scanning entire collections as they are accessioned, but would that not just add to the demands on our limited resources?[2]

A couple of years ago a colleague asked, "What have you stopped doing?" It was a provocative question posited in light of the seemingly ever-increasing demands for repositories to do more as resources grow tighter. What if we could stop doing some other activity and redirect those resources to digitizing? We have already stopped microfilming in favor of digitizing. But even as we never microfilmed all of our collections, we still are not able to digitize all of them. Could we stop doing something else—a standard operation applied to all collections—and redirect those resources to the effort? Might archival processing—the "arrangement, description, and housing of archival materials for storage and use by patrons"—be that task?[3]

Digitizing in lieu of archival processing would enable full-text searching of modern collections containing machine-readable text, as well as browsing by collection and box, while preserving the fundamental archival principles of provenance and original order. It might facilitate and expand the use of our collections while giving our users what they want and continually request. It would allow us to say yes to creators who insist their collections be digitized as a stipulation for donation. While enhancing access, it would better preserve the originals, which could be largely withdrawn from the danger of handling inherent in reading room use. In some cases it might also reduce unprocessed backlogs, making more collections available sooner.

Before we explore this model, let me declare my assumptions, chiefly that most incoming and backlogged collections consist of modern textual materials, those created in the twentieth century and probably containing a large

percentage of typed and printed text on paper, as opposed to handwritten text or other formats like photographs. Whether these modern collections are personal papers or organizational records is probably incidental to the print characteristics they are presumed to have. This model may also be most applicable to personal papers because the legal requirements for government and some business records may stipulate specific outcomes beyond what is suggested here. Finally, collections that have already been processed are beyond its scope, though they may benefit from inclusion.

Digitizing without Archival Processing

How would this model work? Incoming modern collections, as well as those already accessioned but still unprocessed, would be cataloged at the collection level, ideally at the single-level optimum or added value levels of our description standard, *Describing Archives: A Content Standard* (*DACS*). Many repositories already do some version of this; more than half of archival collections in a 2010 survey had an online catalog record.[4] Cataloging documents provenance. It supplies name, subject, and other access points. It also provides other useful data, including an abstract, date span, and extent statement. Most of this information can be collected as part of the appraisal process—before the materials are acquired—by interviewing the creator or donor and surveying the collection. We already perform these actions to evaluate a collection's research value and make an acquisition decision. Repositories might develop an appraisal or accession form to guide this process and encourage robust documentation. The collection-level record would be the sum of the metadata created by archivists under this model; no additional description would be performed.

The contents of each container would then be digitized in its existing order as received from the donor. Any folders bearing labels would be digitized in place, serving as targets for their contents. This would preserve the true original order as received, with all its imperfections. The digitized page images would be packaged in box order, with each "box" of digital images maintained as a separate digital file. The images would be run through an automated Optical Character Recognition (OCR) program. The OCR produced would not be corrected because cleaning it up is resource intensive. Also through scripted processes, the "boxes" of digital page images would be loaded for delivery to users.

Once digitized, the original collection materials could be stored where shelf space is least costly and called upon only rarely when needed. Rehousing of the collection is incidental; as the More Product, Less Process (MPLP) method asserts, the storage environment can be relied on for preservation.[5] If rehousing is performed, it would maintain the arrangement of materials as received from the creator and would occur before the collection is digitized. In this way,

the original of a digitized page is correlated to its box and could be located if necessary.

Users would be able to perform full-text searches by keyword, much as they do with Google. This is what they are accustomed to and expect. They could choose to search across all archival collections or within a single collection. When selecting a search result, the user would be taken to a digitized page image containing the search term. The page image would be situated within its "box" of digitized page images, so users could browse adjacent pages that might be relevant. Users could also select a particular collection by the access points or other metadata created by an archivist in the collection-level record. These include the creator name and subject headings, which are much like those assigned to a book and thus familiar to most users. Users could then choose a box and begin to browse the pages in context with their neighbors.

This is neither revelation nor new model. It is the format of Google Books and similar large-scale book digitization projects. Its end product is a digitized archival collection that is essentially equivalent to a digitized book.

Reviewing the Literature

The linchpins of this proposal are rapid, high-volume digitization and application of OCR to enable full-text searching of archival materials. What does our professional literature reveal about them?

Most mass-digitizing projects involve books and newspapers, but digitization of large segments or entire collections of archival materials is occurring and is tied to a shift away from item-level metadata. In the 2000s, selected series of the John Muir Papers at the University of the Pacific Library were digitized. In a blend of large-scale and selective approaches, about sixteen thousand pages of various formats were scanned, including microfilm copies of handwritten letters. Existing descriptions were used, though full-text transcriptions were created for the letters. Speed of capture was less of a focal point than the effectiveness of the metadata, though the project made clear that capture at lower bit depth and resolution can speed production without inhibiting use.[6]

Driven by digitizing rates that were outstripping item-level metadata creation rates, in 2010 the University of Alabama Libraries digitized the Septimus D. Cabaniss Papers (31.8 linear feet, 46,663 scans) while comparing the efficiency and cost of item-level workflows to new workflows based on the online finding aid. The workflows spanned eight steps, from preliminary description to scanning and quality control, uploading to the web, and metadata remediation. The new method cut the time to create 100 scans from 825 to 434 minutes and cost 32 percent of the item-level method.[7]

To spur the scale of digitization of special collections, in 2011 OCLC reported on nonbook digitization initiatives being done "at scale." The report directed attention to outsourcing choices, throughput rates, and bottlenecks in the processes of nine organizations that were digitizing multiple manuscript collections totaling tens or hundreds of thousands of pages. Rates of four hundred, five hundred, or fifteen hundred images per day from originals, and forty-eight hundred images per day from microfilm, were among those reported.[8]

Millions of documents in various formats at the John F. Kennedy Presidential Library are being digitized by creating low resolution, web-ready JPEG files and providing access through online finding aids. At the Archives of American Art, 76 collections totaling 512 linear feet were digitized from 2005 to 2009; new equipment received in 2010 could scan a page in color in about one second. The National Archives of the United Kingdom has digitized more than sixty million documents using various methods. One project, the 1911 Census, used five automated scanners running around the clock to create about forty thousand images per day. The images were electronically transmitted to the Philippines, transcribed, and returned in five days with full quality assurance. For every document the archives delivered to researchers on site, more than 150 were delivered online.[9]

Thousands of archival collections totaling millions of pages in the Southern Historical Collection at the Library of the University of North Carolina, Chapel Hill, are being digitized in their entirety using the online finding aid for metadata and web interface. In 2009, the staff projected a sustainable digitizing rate of 88,550 scans per year, based on one scanner running twenty hours per week. In a concurrent project to digitize the Thomas E. Watson Papers from 2007 to 2009, actual digitizing rates of about 1.5 minutes and 1.2 minutes per page were achieved for two records series totaling sixty thousand pages. The time included mounting the item on the scanner, positioning four sides of a crop box on the screen, scanning, naming the file, capturing technical metadata, and moving it to permanent storage.[10]

These large-scale archival digitization projects use as metadata the folder-level descriptions in the finding aids that were created during archival processing. The project involving the Southern Historical Collection consciously excluded OCR technology because most of the documents to be digitized were handwritten.

What about other projects that explored OCR? In the late 1980s, the National Archives and Records Administration (NARA) tested the ability of various OCR vendors to read archival documents. The error rate was unacceptably high, so NARA's focus turned to OCR of finding aids rather than historical materials.[11]

In the early 1990s, the Black History Archives Project spearheaded by Virginia Commonwealth University scanned and applied OCR to a variety of

modern historical materials on loan from their owners. The resulting database relied chiefly on keyword searching of uncorrected OCR. Details of the OCR technology and search capabilities were not discussed. The project was deemed a success even though the OCR was acknowledged as being "wildly inaccurate" at times.[12]

In the late 1990s, the application of OCR to digitized Civil War–era imprints was a major disappointment to the Hoole Library at the University of Alabama. A "high-end, 'teachable' OCR package" was not able to convert the "uncommon and infinitely flawed typefaces of the nineteenth century," causing the library to scale back its project.[13]

While not technically archival materials, historic newspapers may have enough similarities to modern archival documents to be germane to this review. Much like machine-generated text in modern archival documents, newspapers present a variety of fonts, deteriorated paper, faded print, bleed through, and fragmented or touching characters. In the early 2000s, the Utah Digital Newspapers Program found that OCR tools were more accurate at reading characters than recognizing words, but that redundancy of words in newspaper articles can increase search success. Filtering the results through a two-million-word dictionary and a Utah place names dictionary ensured that only valid words appeared in the final text. Two-word proximity searches proved to be the best search strategy to address the output created by the OCR program.[14]

Not long afterward, an OCR project involving nineteenth-century British newspapers at the British Library confirmed that character accuracy exceeded word accuracy and found that significant words and words that began with a capital letter were 5 to 10 percent less accurate than word accuracy. Thus, proper nouns, names, and place names are more challenging for OCR technology. Two-thirds of the forty-eight newspapers titles in the test group achieved character accuracy of at least 80 percent, half had word accuracy of 80 percent, and only a quarter yielded more than 80 percent significant word accuracy. Eighty percent recognition was the threshold at which the fuzzy processing of search engines could produce a search accuracy of at least 95 percent.[15]

The National Library of Australia's project to improve OCR accuracy for Australian newspapers tested several methods, including using Australian dictionaries, using grayscale rather than bitonal files, comparing image optimization software, or correcting OCR text manually. The first three methods were not particularly effective, but the last was considered worthwhile and feasible if public users could be recruited to do the corrections.[16]

OCR is still far from perfect, so research continues. The Text-Induced Corpus Clean-up (TICCL) system was tested at Tilburg University in the Netherlands for automatic postcorrection of OCR errors in Netherlands newspapers published from 1918 to 1946, as well as other types of materials. Rather than "training" the

OCR software, TICCL centers on a correction algorithm and yields particularly good results with at least some names, as opposed to common words.[17]

Another recent project looked at open source OCR workflows for historical materials. It proposed a modular approach that customizes the workflow according to the characteristics of the particular source materials. At present, the best solution combines commercial character-recognition engines with open source tools for tailored preprocessing and layout analysis.[18]

It appears from this review that the true accuracy of OCR applied to historical printed materials is far less than the near-perfect rates advertised by the vendors of OCR products. In mass digitizing archival materials, archivists seem to be most focused on the descriptive component and on synching digitized materials to finding aids.

Is Digitizing Faster than Processing?

The crux of this new proposal to digitize in lieu of processing rests on metrics. There is no point in considering it if archival digitizing rates are not competitive with archival processing rates. How do these rates compare?

Processing rates vary tremendously. The Beinecke Library at Yale provides a table of estimated processing times based on the origin and time period of creation. Looking at the time to process a cubic foot of post-1900 collection materials, it estimates 1.1 days for state government records, 1.25 days for business records, 2.25 days for local government records, 3.5 days for personal papers, and 3.25 days for mixed types.[19]

Mark Greene and Dennis Meissner also found great variability in archival processing rates. Their review of NHPRC grants yielded an average rate of 9 hours per foot. When they surveyed processing archivists about processing rates, the average of the results was 14.8 hours per cubic foot, though the most frequently cited figure was 8 hours per foot. The pair went on to advocate for a much faster rate of 4 hours per cubic foot for large twentieth-century archival materials if their MPLP method was used.[20] It is difficult to establish a single rate for purposes of evaluating this model. Because processing archivists themselves posited a rate of 8 hours per foot, I will use that as the benchmark.

How many pages are in a cubic foot? There are 15 linear inches of letter-size pages in a cubic foot, which is comparable to 3 manuscript boxes. If there are 800 pages in a manuscript box, that makes 2,400 pages in a cubic foot. If the pages are double-sided, the actual pages to be digitized doubles to 4,800. The number changes if the pages are legal size, because then there would only be 12 linear inches in a cubic foot. I will assume an average collection is letter size with half of its pages double sided. That yields a figure of 3,600 pages per cubic

foot. The processing rate of 8 hours per cubic foot thus converts to 450 pages per hour. Can digitizing compete with that?

Digitizing equipment can more than exceed this rate. For example, one vendor's planetary book scanner can process a sheet in 2.7 seconds (that is 1,333 sheets per hour) while another's overhead camera system claims up to 700 pages per hour. A do-it-yourself scanner with open hardware and software can capture up to 1,200 pages per hour. Looking to the future, experimental technology can scan a 200-page book in less than a minute by flipping through the pages under a camera.[21]

Real projects probably provide more accurate rates. An operator digitizing books for Google can scan about 50 books a day. If each book is 150 pages, that is 7,500 pages per day or 938 per hour. The Internet Archive digitizes a book in 30 to 60 minutes, depending on its length. That yields a lower rate of a few hundred pages per hour.[22]

The type and condition of archival materials put severe brakes on these speeds. Archival collections mix paper sizes and types of varying strengths and conditions, as well as single- and double-sided sheets; they may be interspersed with unusual formats. They also contain fasteners that take time to remove. Yet even with these problems, some repositories are achieving rapid digitizing rates for selected archival materials. Princeton's Mudd Library is scanning 300 pages per hour on a photocopy-style scanner, while a project at the University of Minnesota is digitizing disbound university publications with a sheet-fed scanner at a rate of 500 pages per hour. The Internet Archive charges 10 cents per image to capture books and 25 cents per image to capture archival materials, which suggests a fairly close rate for these two types.[23]

These very exceptional digitizing projects are competitive with our archival processing rate of 450 pages per hour. Most digitizing projects are much slower. But mass book digitization projects are pushing technology and workflows ever faster, and they are trickling down into the archival world. It seems safe to say that digitizing rates within the archival world will continue to increase, though the physical characteristics of the materials will always serve as a drag on those rates. It may behoove us to reconsider our materials handling rules for digitization projects. Many twentieth-century collection materials are low in intrinsic value and sturdy enough to enable auto-feed scanning without damage. This sometimes occurs already; when I worked at the National Archives in the 1990s, the records declassification unit alone was authorized to use an auto-feed copy machine because of the huge amount of modern materials it had to duplicate. We might also consider one-time handling of a nature we might not repeatedly allow in the reading room if the end result is a digital surrogate that can stand in for most or all future use.

Reassessing the calculation through a different lens, the processing rate of 8 hours per foot may be too optimistic, and it does not distinguish between different types of collections. If we instead use the Beinecke's processing rate for personal papers, 3.5 days per cubic foot, the processing rate for the same 3,600 pages drops to 129 pages per hour. This makes more archives' digitizing projects competitive. The University of North Carolina's Digital Southern Historical Collection program achieves 90 images per hour, and the Green Revolution project at the University of Minnesota reaches 96 per hour.[24]

Tempering factors may also lift this model over the digitizing rate barrier. Repositories may be able to deploy more human resources to digitizing than to processing because many of the high-speed archival digitization projects cited above employ students. In addition, scientific or highly technical collections, as well as those in foreign languages, could be digitized without the need for processing by staff with specialized knowledge. The main bottleneck to increasing the number of staff digitizing, as compared to processing, is probably the cost of fast-throughput scanning stations, but even these could be stretched by shift work.

It is probably safe to say that digitizing has the potential to achieve rates comparable to those for archival processing of at least some archival materials. If the metrics appear possible, arguments turn to the value of archival processing and its codification in finding aids. Aren't finding aids the most effective method of describing manuscript materials for researchers?

How Will Users Manage without Finding Aids?

In all probability only archivists would ask this question. Studies repeatedly show that researchers have trouble understanding finding aids.[25] This was one of the first things I learned about archives—when I told my father I hoped to be an archivist, he replied that he never understood archival research until he read a book that explained archives. While we might dream of researchers who read a book to understand our system, we cannot expect them, especially if we wish to expand use. Providing our users with something more familiar than finding aids—something more like Google Books—could increase our relevance to the public, especially in today's information-driven economy.

Archival processing typically produces a finding aid with series descriptions and folder titles, which are the most detailed level of description created. Let's look at the folder-level metadata, which would be jettisoned if we digitized without processing.

The DACS description standard requires that titles include a creator name and a term indicating the nature of the materials. Creator names are valuable for access, but they are often inherited from higher levels of description. When

many folder titles share a single creator, the degree of differentiation within a collection description is limited. Correspondence series are probably most likely to provide a folder title involving a name other than that of the overall collection. Due to the inheritance of creator names, the "nature of materials" segment of the DACS title often carries the burden of description. But how many researchers think in terms of physical forms, or search for them?

Numerous studies indicate that users search by subject, yet DACS requires few subject terms.[26] Topical terms are added to titles only when the name and nature elements do not clearly identify the materials. This perpetuates our tendency to provide limited subject access to archival materials. We prefer provenance for access because archival materials are closely related to the creator's activities and difficult to categorize by topic. Identifying the "aboutness" of a letter can be problematic, and a document that covers multiple subjects can only be placed in one folder. As a result, finding aids are out of sync with how many users search.

Traditionally, users had to search for archival materials by starting at the collection level and then drilling down to series and then to folders. This provenance-based approach could have been rendered nearly irrelevant in the digital age were it not for Encoded Archival Description (EAD), which makes the entire text of finding aids searchable on the Internet. This has freed users from some of their most frustrating problems with finding aids. They can completely bypass the cumbersome and unintuitive provenance-based descriptions and instead search the full text of finding aids directly. Most users have already moved to this kind of search.[27]

Finding aids also severely restrict the pool of searchable text. Users can only search the very limited metadata of the finding aid, in which the few words of a folder title stand in for the content of dozens of documents—sometimes a hundred pages or more—within that folder. Searching only the metadata of the finding aid may falsely yield a paucity of results. In all likelihood, many users are doing nothing more than a keyword search of our finding aids by some topic. Wouldn't we rather have them perform that search on the full text of the materials?

We know that controlled vocabularies are essential to effective search results, but both finding aids and full text contain mostly uncontrolled words. While DACS identifies six broad categories of access points, it leaves both their definition and choice of thesauri to local decision, with one exception: that of names of creators in a single-level description or the top level of a multilevel description. Some local entities prescribe greater use of access points; "to promote content access," the Online Archive of California requires at least three controlled access headings, including name, occupation, and function, but not

subject.[28] When searching finding aids, users cannot rely on controlled access points other than certain creator names. This is no better than full-text search.

In the new model, users could search the collection-level metadata with access points created by archivists and/or the uncorrected full text of all the materials in each collection. Most OCR software can achieve accuracy rates of 98 to 99.9 percent. One popular product claims recognition of 189 languages and "up to 99.8% recognition accuracy," though that depends on such factors as document quality and scanner settings.[29] We know that character accuracy rates mask problems with word accuracy, especially accuracy of significant words. But significant words in a collection are likely to be repeated, which improves the chances of accuracy. In addition, the OCR software could perhaps be "trained" to recognize proper names and specialized terms anticipated to appear in a given collection. Furthermore, once users find a search term in the text, they can browse nearby to uncover related documents that might not have been accurately indexed by the software.

Series and folder titles, as well as the date spans tied to them, would be lost in the full-text model. Some of them carry rich meaning, while others are uninformative. Looking at hypothetical examples, the "Joe Seeland letters" file identifies and aggregates all of his letters and adds a date span. The "printed matter" file is much more opaque. A user interested in whatever topic—let's say submarines—may never choose to consult the "Joe Seeland letters" file or the "printed matter" file based on their titles, but might discover relevant documents if able to search their full text for the word "submarine." Depending on the nature of the query, some users would enjoy more success with full-text searches and others would find less. That is already how it works with both finding aids and full-text search.

Few users employ advanced search functions.[30] We marvel when their keyword search yields a million hits and they are content to view only the few topping the list. Thus full-text search by keyword will meet the needs of many users and be an end in itself. For those more discerning, the collection-level metadata could take on new value. To filter extensive search results, users might execute a search of the controlled subject headings at the collection level, or consult the biographical note and other collection-level information, which is easily overlooked when container lists are searchable. This could strengthen appreciation and understanding of one of our core principles, provenance.

This model might also provide more seamless integration of hybrid paper-digital collections. It is unlikely that we will arrange born-digital materials the same way we arrange analog materials. Opening, viewing, and arranging the contents of hard drives brimming with thousands of born-digital files is unrealistic; users will probably perform full-text searches of these materials. By digitizing the papers in a hybrid collection, the collection can be united

electronically. The context between the paper and born-digital components in this digital union may not be fully integrated, but at least both parts will be fairly equally accessible through a single search interface. Users will not need to switch gears between finding aids for boxes of papers and computer terminals for born-digital materials.

It is hard to say with certainty that finding aids provide significantly better access than full-text search, which has only recently become possible. Archival processing and finding aids were developed when our predecessors could not keep up with item-level registration of documents. In a paper-bound era they produced an effective, elegant solution for access. The MPLP method is controversial despite striving to sustain processing in the face of more voluminous collections. It does not ask, "If digital technology was available when archival processing was developed, which method would we be using today?"

Can These Digitized Collections Go Online?

The content of modern collections is likely to contain personal information about living people and be protected by copyright. This information cannot be posted on the Internet, which is the ideal delivery tool for digitized materials and the place where users expect to find archival collections. Donor consent in transfer agreements helps immensely but does not cover the rights of third parties whose works are common in archival collections. Archivists are committed to openness, but we believe it must be balanced by respect for intellectual property rights and personal privacy. This most difficult problem cannot be entirely overcome by the march of technology.

Digitizing by archives is usually not considered a copyright violation. Posting digitized items online is equivalent to publication, which is a violation. One solution to the copyright problem is allowing access to digitized collections only in the reading room, but this does nothing for users who demand online access. A more liberal approach would involve providing older collections on the Internet—those in which the materials are, say, fifty years old or some other age at which infringement is deemed less risky.[31] More recent materials would remain limited to the archives reading room. Some type of virtual reading room where users request and receive access to specific collections, controlled by password and time limit, might also be possible. Such a space already exists at the Special Collections and Archives at the University of California, Irvine.[32]

From an openness perspective, the ideal solution to the copyright problem may be to post collections online by invoking fair use, building a case to support it, and supplementing it with a liberal take-down policy. Providing searchable full text may be the transformative use that qualifies for fair use, as a recent court ruling suggests.[33] Fair use is the rationale for online delivery of the output

of the mass digitization project involving the Southern Historical Collection at UNC.[34] In some cases it may be worthwhile to buttress this by going to our funders who insist on digitizing, explaining what we are doing, and requesting their financial support if the results trigger lawsuits.

Privacy must be assessed separately from copyright; it can be violated in the reading room as well as on the Internet, with numerous third parties unknowingly involved. Technology can help with privacy issues. Forensic software that identifies words of concern and patterns for other data of concern, like Social Security and credit card numbers, could be used to redact private data automatically via full-text search. An alternative to redaction might be a nondisclosure agreement signed by users. This approach is already used for some collections that are only available in reading rooms, and it may be gaining momentum; it was the focus of the 2012 annual meeting of the Privacy and Confidentiality Roundtable of the Society of American Archivists.[35] A click-through nondisclosure agreement might even prove acceptable for collections posted online, or available through an archives-controlled virtual reading room, especially if archivists are somewhat selective about which collections they make available this way.

These approaches could be mixed and matched to particular collections to achieve the greatest possible access while we exercise our responsibilities to protect data in collections likely to need it. Where we determine that collections cannot go online, perhaps we could instead provide the indexed words of the full text online but out of context, keyed only to their collections, somewhat like the database of five billion words from Google Books.[36] Users could perform searches, identify collections, review the collection-level descriptions, and then contact the repository. An archivist could then deliver digitized materials directly to that user.

What about All the Other Problems?

Relative to copyright and privacy, other problems may seem minor. They include workflows, storage, and preservation, and their costs. Some handwritten, visual, and audiovisual materials do not seem to benefit from OCR. Undoubtedly, there are other problems.

Digitizing rates may be fast, but that is just one step of the process. The full-text model eliminates one standard digitizing step—metadata creation, which is thought to account for a third of digitizing costs.[37] At large scale, logistical issues crop up at every remaining step, from removing fasteners to checking image readability, looking for missing pages, and uploading and indexing for delivery. The most rapid archival digitization projects previously mentioned note many postcapture workflow bottlenecks. Indeed, missing and unreadable

pages are problems with Google Books and fixing them is not efficient. Even if bad images are not corrected when identified, the full-text model might still support discoverability and satisfy a majority of users; as a fallback, the original documents could be made available in the reading room. If archivists think this model has merit, we will need to experiment with and develop workflows that fit our goals and budgets.

The sheer amount of data generated by mass digitization is a storage challenge, especially with the cost of redundant copies and regular backup. While digital storage capacity increases rapidly—Kryder's Law recognizes the "50-million-fold increase" in magnetic disk capacity since 1956[38]—the true cost of storage remains unclear. We lack consistent metrics for the human operations involved in long-term management of storage systems, and we do not know how economies of scale might reduce unit costs.

To conserve storage space, we can trim the files created to the minimum needed for OCR and human readability—300 ppi grayscale files are much smaller than 600 ppi color files, for example. The files will also piggyback on the storage needed for born-digital archival materials, which will quickly outpace the number, size, and complexity of these digitized files. At this point, however, most large repositories in the United States lack the funding and infrastructure for managing born-digital materials, and few have significant amounts of such files. Then again, the majority of them have an institutional repository, which could conceivably become the storage site for digitized archival materials.[39]

The long-term preservation of digital materials is also up in the air. However, the formats we generate by digitizing textual materials will be more consistent and stable than most of the born-digital archival materials that we will be responsible for in the future. The target file formats we use are based on preservation concerns, and their ubiquity helps ensure that migration paths will be available. The homogeneous files we create will be less difficult to preserve than the born-digital ones created by everyone else.

These and other technological challenges are formidable. Yet they are chiefly problems of technology, where the exponential improvements noted by Kryder's Law can be applied to many areas. Moreover, even item-level mass processes, when automated, can be performed efficiently. Many economic sectors, including the Internet, retail, health care, science, and government, are harnessing big data with powerful analytical tools. The Internet Archive manages about nine thousand terabytes of data.[40] Google seems to have mastered the problems of storage, search, and speed. Its success may be nearly unmatchable, and the business model of most archives does not include commodification of resources. But as tools for managing big data are developed in other sectors, they will trickle down to archives, and implementing them only after they are proven may save some cost. Partnerships among repositories, or with government or

commercial entities—perhaps even Google or the Internet Archive—may help underwrite the costs of technology.

We may look to archives in China for solutions, though without knowing Chinese, it is difficult to gauge the full parameters of their programs. The Foreign Ministry Archives of the People's Republic of China has digitized all publicly available documents generated between 1949 and 1965; access is available only on computers in the reading room. Rather than full-text search, it appears that metadata are searchable. The Beijing Municipal Archives planned to digitize "10% of our paper archives, some microfilm information and all of the audio-visual archives by the end of 2005, for the convenience of on-line utilization." The National Library of China had digitized more than 180 terabytes of special collection materials as of 2008. Government involvement, labor costs, and many other societal factors are different in China, but much of the technology should be similar across national boundaries.[41]

Nonprint archival materials also present challenges. Audiovisual materials are more costly to digitize but their sturdy casings make them less vulnerable to robotic handling, assuming there is no sticky shed syndrome or other condition issues. One machine robotically digitizes multiple video streams, another digitizes eighty thousand videotapes in a year, and the Swedish National Archive of Recorded Sound and Moving Images was using automated workflows to digitize open-reel audiotapes at a rate of 1,500 hours per day in 2007.[42] For spoken-word video and audio recordings, speech-to-text software might produce searchable text comparable to that of OCR for text-based materials. Voice recognition software is improving rapidly; it is increasingly common in smartphones, and one popular software product claims "up to 99% accuracy out of the box."[43] The transcripts generated by these processes could be merged with the text recognized by OCR to provide seamless search points for users.

Handwritten materials present challenges to machine readability. Computer scientists are working on machine recognition of handwritten U.S. Census schedules, which present writing by many different people in a structured format that enhances readability. While their results are not usable yet, they are making remarkable progress.[44]

Crowdsourcing tools could conceivably assist with a variety of problems. Correcting errors in OCR, transcribing handwritten materials, and notifying repositories of image quality problems are among the many areas where end users could contribute. It may be difficult to recruit the public for these tasks, but there are success stories involving materials of genealogical interest. When the National Library of Australia began digitizing one million newspaper pages per year and sought users to correct text and tag data, the public responded in an unprecedented way, with thousands volunteering.[45] The FamilySearch website provides an easy way for individuals around the world to index records

from home.[46] An alternative approach gets users to decipher scanned words that OCR failed to recognize as part of web security and authentication measures.[47] Even if the public response is effective, developing these tools adds another layer to our technology challenges.

It is likely that even under this new model some collections would still warrant archival processing, which might be redefined as a boutique activity. Collections of photographs or those with many fragile or handwritten materials might fit this category, as well as especially significant or valuable collections. Then again, improvements in facial recognition software may someday make many visual materials accessible without metadata.

Beyond all the practicalities is the question of our professional survival. If processing and finding aids become peripheral or are eliminated, what will happen to archivy as a profession? Original order and provenance, which this model maintains and could possibly elevate, should be sufficient to define us. Meanwhile, we have no shortage of challenges. We can work on effectively surveying collections during acquisition to create rich collection-level descriptions. We can partner with computer programmers on optimizing search algorithms for full-text coupled with collection-level descriptions, and we can assist users with sophisticated search protocols. We can collaborate with creators on collecting and preserving their born-digital materials, from their email to their dynamic social media presence. Capturing and preserving the historical and societal record will still be a full-time job.

Should We Consider Adopting the Full-Text Model?

The basic idea behind this model is simple. It asks whether mass digitization of archival materials would be more feasible if we could eliminate the resources—staff, time, money—invested in archival processing of collections before they become eligible for digitization. It would apply to modern collections that are incoming or backlogged for processing. By not processing them, we would have more resources to direct toward digitization.

Instead of the folder descriptions created during manual processing, uncorrected OCR would be applied as an automated task to the full text of the materials. Choosing between these options might hinge on how much and what kind of description are sufficient and adequate for users. The few words selected for a folder title stand in for the entire text of that folder. The full-text model embraces the entire text in that folder, but will not be 100 percent accurate or complete due to the limitations of OCR. What level of full-text word accuracy would allow access comparable to the access a folder title enables and can OCR achieve it? If the full-text model provides a discovery method that is more familiar and intuitive for most users, does that change the equation?

As we consider the full-text model, we should bear in mind our core beliefs and practical realities. We want our collections to be open and accessible to all without cost and easy to discover, access, and use. We want to be responsive to the researchers, records creators, and resource allocators who wish to see our collections digitized. We need to maximize our limited resources. As our resources stagnate or diminish while expectations for digitization grow seemingly exponentially, we must ensure our time is well spent. Can this model better achieve our goals and the expectations of our constituencies?

If we really do want to provide entire collections on the Internet, what steps are necessary to achieve it, and what steps add unnecessary cost? Archival processing produces an arrangement and a finding aid that enable access. Digitizing that processed collection is a second, separate activity that can only be performed at whatever digitizing rate is achievable. If we instead digitize without processing, we would produce a fully digitized, somewhat searchable collection for online access. The new model turns a two-step operation into a single activity.

Digitizing rates for archival materials lag behind those for books. OCR technology cannot produce 100 percent word accuracy on the range of machine-produced print in modern archival materials. We could dismiss this model today, but we should not don blinders to it. Technology is advancing rapidly. If digitizing coupled with full-text search seems at all promising, exploring it now allows us to be ready when the technology is. If we decide it is unrealistic or inferior to current methods, then we are better positioned to make our case to the many constituencies who expect to find collections online. If we decide the full-text model is possible and desirable, we can begin exploring the workflows, delivery methods, and other problems that remain to be solved; this will also help us ramp up for the stewardship challenges of born-digital archival materials.

We could begin incrementally. We might select modern collections that would most benefit from the full-text model—those that seem appropriate for the Internet from copyright and privacy perspectives, relying heavily on fair use. Or we might funnel collections received in good order to processing and direct highly disorganized collections, which are the most labor-intensive to process, to digitizing and OCR.

It is easy to reject this model because we are wedded to the traditions of processing and finding aids. Most of us cut our teeth as archivists by processing collections. But archivists are also inherently practical—aggregate description at its core is an approach to archival description based on the realities of large collections and limited resources. It is difficult to imagine collections without finding aids, but we should remain aware of and open to the entire world in which we operate. It is worth remembering the situation that drives this model. We have limited or diminishing resources, but we are surrounded by demands to

digitize our collections en masse, and those demands are not likely to subside. If they do, odds are it will be because we are doing it, or because our researchers, records creators, and resource allocators have given up on us and moved on.

Notes

[1] The author wishes to thank Jean M. Deken, Paula Jabloner, William E. Miller, and Zuoyue Wang for their comments on early versions of this manuscript, though their assistance does not mean they endorse the ideas it presents.

[2] Ricky Erway, "Supply and Demand: Special Collections and Digitisation," *Liber Quarterly* 18, nos. 3–4 (2008): 327.

[3] Richard Pearce-Moses, *A Glossary of Archival and Records Terminology* (Chicago: Society of American Archivists, 2005), 314.

[4] Jackie M. Dooley and Katherine Luce, *Taking Our Pulse: The OCLC Research Survey of Special Collections and Archives* (Dublin, Ohio: OCLC Research, October 2010), 46, http://www.oclc.org/content/dam/research/publications/library/2010/2010-11.pdf.

[5] Mark A. Greene and Dennis Meissner, "More Product, Less Process: Revamping Traditional Archival Processing," *The American Archivist* 68 (2005): 251.

[6] Shan C. Sutton, "Balancing Boutique-Level Quality and Large-Scale Production: The Impact of 'More Product, Less Process' on Digitization in Archives and Special Collections," *RBM: A Journal of Rare Books, Manuscripts, and Cultural Heritage* 13, no. 1 (2012): 50–63.

[7] Jody L. DeRidder, Amanda Axley Presnell, and Keven Walker, "Leveraging Encoded Archival Description for Access to Digital Content: A Cost and Usability Analysis," *The American Archivist* 75 (2012): 143–70.

[8] Ricky Erway, *Rapid Capture: Faster Throughput in Digitization of Special Collections* (Dublin, Ohio: OCLC Research, 2011), http://www.oclc.org/content/dam/research/publications/library/2011/2011-04.pdf.

[9] Erica Boudreau, "Digitization and EAD at the JFK: A Marriage of Opportunity" (paper presented at the Description Section program of the annual meeting of the Society of American Archivists, Chicago, Illinois, August 25, 2011), http://www2.archivists.org/sites/all/files/desc_sec_presentation_2011_boudreau.pdf; Barbara Aikens, "Going with the Flow: Mass-tering Digitization at the Collection Level: Workflow at the Archives of American Art" (paper presented at the annual meeting of the Society of American Archivists, Austin, Texas, August 15, 2009), http://files.archivists.org/conference/austin2009/Session601_Aikens.ppt; Erway, *Rapid Capture*, 6; Dan Jones, "Mass Digitization of Historical Records for Access and Preservation," *Serials: The Journal for the Serials Community* 21, no. 2 (2008): 98–101.

[10] Southern Historical Collection, University Library, University of North Carolina at Chapel Hill, "Extending the Reach of Southern Sources: Proceeding to Large-Scale Digitization of Manuscript Collections" (final grant report prepared for the Andrew W. Mellon Foundation, June 2009), 67, http://www.lib.unc.edu/mss/archivalmassdigitization/download/extending_the_reach.pdf; UNC University Library, From Investigation to Implementation: Building a Program for the Large-Scale Digitization of Manuscripts, "Digitization" (December 2009), http://www.lib.unc.edu/mss/archivalmassdigitization/index.html?section=digaccess&page=digitization.

[11] Marie Allen, "Optical Character Recognition: Technology with New Relevance for Archival Automation Projects," *The American Archivist* 50 (Winter 1987): 88–99.

[12] John H. Whaley Jr., "Digitizing History," *The American Archivist* 57 (Fall 1994): 660–72.

[13] Andrea Watson, with P. Toby Graham, "*CSS* Alabama *Digital Collection*: A Special Collections Digitization Project," *The American Archivist* 61 (Spring 1998): 124–34.

[14] Kenneth Arlitsch and John Herbert, "Microfilm, Paper, and OCR: Issues in Newspaper Digitization," *Microform and Imaging Review* 33, no. 2 (2004): 59–67.

[15] Simon Tanner, Trevor Muñoz, and Pich Hemy Ros, "Measuring Mass Text Digitization Quality and Usefulness: Lessons Learned from Assessing the OCR Accuracy of the British Library's 19th

Century Online Newspaper Archive," *D-Lib Magazine* 15, nos. 7–8 (2009), http://www.dlib.org/dlib/july09/munoz/07munoz.html.

[16] Rose Holley, "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs," *D-Lib Magazine* 15, nos. 3–4 (2009), http://www.dlib.org/dlib/march09/holley/03holley.html.

[17] Martin Reyneart, "Non-Interactive OCR Post-Correction for Giga-Scale Digitization Projects," ACM Digital Library, *CICLing'08 Proceedings of the 9th International Conference on Computational Linguistics and Intelligent Text Processing* (Heidelberg, Ger., 2008), 617–30, http://dl.acm.org/citation.cfm?id=1787647.

[18] Tobias Blanke, Michael Bryant, and Mark Hedges, "Ocropodium: Open Source OCR for Small-Scale Historical Archives," *Journal of Information Science* 38, no. 1 (2012): 76–86.

[19] Beinecke Rare Book and Manuscript Library, Yale University, "Table of Estimated Processing Time," Appendix A, *Archival Processing Manual*, http://www.library.yale.edu/beinecke/manuscript/process/appA.html#I.5.%20Table.

[20] Greene and Meissner, "More Product, Less Process," 228–29, 253.

[21] Indus, International, "News Release," January 3, 2011, http://www.indususa.com/?page_id=296; Atiz Innovation, "BookDrive Pro," http://pro.atiz.com/; "Welcome to DIY Book Scanner," http://www.diybookscanner.org/; Charlie Sorrel, "High-Speed Camera Scans Book in Seconds," Wired, "Gadget Lab," March 18, 2010, http://www.wired.com/gadgetlab/2010/03/high-speed-camera-scans-books-in-seconds/.

[22] Karen Coyle, "Mass Digitization of Books," *The Journal of Academic Librarianship* 32, no. 6 (2006): 641–42.

[23] Erway, *Rapid Capture*, 14, 18; Internet Archive, "Internet Archive Digitization Capabilities + Pricing Guidelines for 2011" and "Archival Materials," http://archive.org/~pnguyen/pricing/index.html and http://archive.org/~pnguyen/pricing/archival/index.html.

[24] Erway, *Rapid Capture*, 13, 16.

[25] Examples are Christopher J. Prom, "User Interactions with Electronic Finding Aids in a Controlled Setting," *The American Archivist* 67 (2004): 234–268; Wendy Scheir, "First Entry: Report on a Qualitative Exploratory Study of Novice User Experience with Online Finding Aids," *Journal of Archival Organization* 3, no. 4 (2005): 49–85; Morgan G. Daniels and Elizabeth Yakel, "Seek and You May Find: Successful Search in Online Finding Aid Systems," *The American Archivist* 73 (2010): 535–68.

[26] For a compilation of relevant studies with commentary, see Jennifer Schaffner, "Users Search by Subjects and Keywords," in *The Metadata* Is *the Interface: Better Description for Better Discovery of Archives and Special Collections, Synthesized from User Studies* (Dublin, Ohio: OCLC Research, 2009), 6–9, https://www.oclc.org/resources/research/publications/library/2009/2009-06.pdf. A few examples of such studies are Karen Collins, "Providing Subject Access to Images: A Study of User Queries," *The American Archivist* 61 (1998): 36–53; Christine L. Borgman, "Why Are Online Catalogs *Still* Hard to Use?," *Journal of the American Society for Information Science* 47, no. 7 (1996): 493–503; Wendy M. Duff and Catherine A. Johnson, "Accidentally Found on Purpose: Information-Seeking Behavior of Historians in Archives," *The Library Quarterly* 72, no. 4 (2002): 472–96.

[27] See, for example, Joshua Ranger, "More Bytes, Less Bite: Cutting Corners in Digitization" (paper presented at the annual meeting of the Society of American Archivists, San Francisco, California, August 2008), http://www.archivists.org/conference/sanfrancisco2008/docs/session701-ranger.pdf; Jane Lee, "OAC First Round Usability Test Findings" (September 11, 2008), http://www.cdlib.org/services/uxdesign/docs/2008/oac_usability_aug2008.pdf; Michelle Light, "The Endangerment of Trees" (paper presented at EAD @ 10 conference, San Francisco, California, August 31, 2008), http://www.archivists.org/publications/proceedings/EAD@10/Light-EAD@10.pdf.

[28] California Digital Library, *OAC Best Practice Guidelines for EAD*, version 2.0 (Berkeley, Calif.: Regents of the University of California, April 2005), 15, http://www.cdlib.org/services/dsc/contribute/docs/oacbpgead_v2-0.pdf.

[29] Coyle, "Mass Digitization of Books," 643; ABBYY, "ABBYY FineReader 11 Corporate Edition," http://finereader.abbyy.com/corporate/features/.

[30] Examples of such studies are Karen Markey, "Twenty-Five Years of End-User Searching, Part 1: Research Findings," *Journal of the American Society of Information Science and Technology* 58, no. 8

(2007): 1071–81; *Some Findings from WorldCat Local Usability Tests Prepared for ALA Annual, July 2009* (Dublin, Ohio: OCLC, 2009), http://www.oclc.org/content/dam/usability-labs/213941usf_some_findings_about_worldcat_local.pdf.

[31] This assertive approach is in the spirit of, but not specifically sanctioned by OCLC Research's "Well-intentioned practice for putting digitized collections of unpublished materials online," revised May 28, 2010, http://www.oclc.org/content/dam/research/activities/rights/practice.pdf.

[32] Special Collections and Archives, UCI Libraries, University of California, Irvine, "Rules of Use for the Virtual Reading Room in UCIspace @ the Libraries," http://special.lib.uci.edu/using/docs/rules-of-use-virtual-reading-room-ucispace.pdf.

[33] Steve Kolowich, "A Legal Sweep," *Inside Higher Ed* (October 12, 2012), http://www.insidehighered.com/news/2012/10/12/hathitrust-ruling-universities-fair-use-winning-streak, discussing *Authors Guild v. HathiTrust*.

[34] Southern Historical Collection, "Extending the Reach of Southern Sources."

[35] Collections in which privacy is protected by a use agreement rather than redaction include the Norfolk Public Schools desegregation papers at Old Dominion University, the Eesti NSV Riikliku Julgeoleku Komitee records and Hib al-Ba'th al-'Arab al-Ishtirk records at the Hoover Institution, and the Sovetskaja voennaja administratsii v Germanii (SVAG) digital archive, http://svag.unc.edu, hosted by the University of North Carolina.

[36] Dan Charles, "Google Book Tool Tracks Cultural Change with Words," *All Things Considered*, National Public Radio (December 16, 2010), http://www.npr.org/2010/12/16/132106374/google-book-tool-tracks-cultural-change-with-words.

[37] Steven Puglia, "The Costs of Digital Imaging Projects," *RLG DigiNews* 3, no. 5 (1999), http://chnm.gmu.edu/digitalhistory/links/cached/chapter3/link3.10b.digitalimagingcosts.html.

[38] Chip Walter, "Kryder's Law," *Scientific American*, July 25, 2005, http://www.scientificamerican.com/article.cfm?id=kryders-law.

[39] Dooley and Luce, *Taking Our Pulse*, 13, 25–26, 60–61.

[40] Bill Carter, "All the TV News Since 2009, on One Web Site," *New York Times*, September 17, 2012.

[41] Amanda Shuman, "Foreign Ministry Archives of the PRC," *Dissertation Reviews* (February 14, 2012), http://dissertationreviews.org/archives/936; Beijing Municipal Archives, "About Us," http://www.bjma.org.cn/eng/m_aboutus.ycs; Wei Dawei and Sun Yigang, "The National Digital Library Project," *D-Lib Magazine* 16, nos. 5–6 (2010), http://www.dlib.org/dlib/may10/dawei/05dawei.html.

[42] DamSmart, Innovative Media Migration, "SAMMA System," http://www.damsmart.com.au/products/samma; David Stewart, "ADAM Lends a Helping Hand," *TVBEurope* (September 2012), 44, http://content.yudu.com/A1y8nl/TVBESep2012/resources/44.htm; Martin Jacobson, "Migration of 1.2 Million Hours of Audio Material in a Three-Year Period" (paper presented at Unlocking Audio: Sharing Experience of Mass Digitisation, London, October 26–27, 2007), 8–9, http://www.bl.uk/reshelp/bldept/soundarch/unlockaudio/papers07/unlockingaudio.pdf.

[43] Nuance, "Dragon," http://www.nuance.com.

[44] Kenton McHenry, "Tools for Image-Based Search: Providing Search without Transcriptions" (paper presented at the annual meeting of the Society of American Archivists, San Diego, California, August 10, 2012).

[45] Rose Holley, *Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers* (National Library of Australia, 2009), http://www.nla.gov.au/ndp/project_details/documents/ANDP_ManyHands.pdf.

[46] FamilySearch, "Indexing Makes Records Free and Searchable," https://familysearch.org/volunteer/indexing.

[47] Luis von Ahn et al., "reCAPTCHA: Human-Based Character Recognition via Web Security Measures," *Science* 321, no. 5895 (published online August 14, 2008): 1465–68, http://www.sciencemag.org/content/321/5895/1465.abstract.

ABOUT THE AUTHOR

Larisa K. Miller has nearly twenty-five years of experience as an archivist. She is an associate archivist at the Hoover Institution at Stanford University, where her responsibilities include directing the digital program. Her earlier work at Hoover included supervising the processing of archival collections. Before joining Hoover in 2002, she was senior archivist at the National Archives regional facility in San Bruno, California. She started her career in 1989 as an archivist at the National Archives in Washington, D.C. Miller is the immediate past president of the Society of California Archivists and was the founder and first chair of SAA's Metadata and Digital Object Roundtable. She coauthored *"Capture and Release": Digital Cameras in the Reading Room* (OCLC Research, 2010). Miller has a BA in geography and an MA in American studies from the University of Minnesota and an MLIS from San José State University.