Looking Back: Observations on Digital Initiatives from the 1980s to the Present
For the National Archives Preservation Conference, March 26, 2009
By Carl Fleischhauer, version for the Web, March 31, 2009

Introduction

My first job at the Library of Congress was in the American Folklife Center. I worked with folklorists. We took cameras and tape recorders and looked for grandmas who knew their neighborhoods or old guys with long beards to tell us about how it used to be. Today, I'm the old guy with the beard, looking back at twenty-five years of digitization.

If our topic was *imaging for preservation and access*, we would have to reach back to the 1920s and 1930s when preservation microfilming got rolling in libraries and archives. (My beard isn't that long.) From the start, microfilming was partly about making safety copies (what if the original was lost?) and also about dissemination. These copies might be called *virtual replicas*: no one will mistake them for the original paper items but they provide good service to researchers. In 1957, for example, the Library received funding from Congress to microfilm our presidential papers and place copies in major research libraries around the US. Preservation *and* access. But there was third factor that became more prominent in the 1960s and 1970s that had to do with saving space: discard certain types of printed matter after filming. That goal has not been prominent in digital reformatting.

My simplified three-phase story emphasizes technical matters and the federal agency role. There are many important non-technical topics, some will be alluded to today. And I regret that I have time to say very little about the many terrific contributions from digitization programs at university libraries and archives, and say nothing of the excellent work carried out in other nations.

Phase one is about exploring technology in the 1980s. One of the questions turned out to be: "which technology?"

One paragraph in the Library of Congress annual report for 1981 features the word *preservation* and names two new projects: the *Optical Disk Pilot Project* and another devoted to mass deacidification. By 1982, two optical disk contracts had been awarded and a team of staffers-on-detail was being assembled--proof of the long beard, I was one of those staffers. The 1983 annual report states that the optical disk project will "help in preserving and improving access to the Library's vast collections." The overall effort was intended to adapt new technologies to the needs of libraries and archives. The project's name celebrated *one* technology: the *optical disk*. But the exploration of another technology proved to be more important: *digital imaging*.

The part of the project focused on printed matter borrowed much from the emerging field of digital document imaging. We were in step with the industry. Here's a hint: in July 1983, the National Micrographic Association was renamed the Association for Information and Image Management.

The optical disks that the Library used for images of printed matter were 12-inch write-once platters in a large jukebox. User access was provided by state of the art display screens and printers. (Pretty basic display devices by today's standards.) At first, the plan was to scan current periodicals--for example, *Time* magazine--and to make the imagery available in several reading rooms, reducing wear and tear and theft of the originals. There was a strong local-access emphasis.

Compared to what we can do today, the imaging left a lot to be desired: bitonal images (pure black and white, like a FAX), typical for document imaging at that time. Dithering was applied to the pictures in the magazines but the results were just OK.

Meanwhile, in what we called *the non-print project*, we explored different technology, optical disks of the replicated, mass-produced variety. Laser videodiscs, which carry an *analog* video signal, predated our project. But we got started in almost exactly the same year that Philips and Sony first introduced the audio compact disk: 1982. (CD-ROMs were not introduced until 1985.)

We put some moving image content on videodiscs. One set of early silent movies documented the circumstances surrounding President William McKinley's assassination in 1901. There was also a videodisc with a short segment of Martin Scorsese's *Taxi Driver*. This was when Scorsese and others in Hollywood were on a campaign against fading film stocks and our little experiment was a gesture toward the retention of color information by electronic means, imperfectly realized given the very low levels of resolution offered by broadcast video. For sound, we put two concert recordings on compact disks.

Most interesting was the work with still images. A single videodisc side can carry 50,000 video frames, each with a new picture. We put 25,000 images of glass plate negatives from the Detroit Publishing Company--one of their main businesses was postcards--on one disk side and yoked it up to a searchable database with bibliographic records. This was very instructive in terms of user service: a researcher could examine hundred of pictures very quickly and, with the accompanying metadata, make some good choices.

But the non-print project didn't teach us much about digital imaging. We saw the need: these analog videodisc images could not be networked. You had to be there, next to the disk player looking at a TV set, to see the pictures. The levels of resolution were just OK. If you found something promising, you had to ask to see the original in order to discern the details.

One of the biggest lessons from the Library's optical disk project had to do with copyright. As you might expect, it was challenging to get an agreement from the owners to include even a segment of *Taxi Driver*--we had to omit the sound track--and the print publishers spoke to us in forceful terms about scanning current periodicals. "You can't do it without the owner's permission," they said. Since the project was already rolling, there was not time to work through the needed negotiations, so we put the document scanners and the disk system to work on printed

matter in use by the Congressional Research Service and the members of congress they serve. The Library of Congress Optical Disk Pilot Project wrapped up in 1987.

At the National Archives, similar technologies were featured in the Optical Digital Image Storage System project (ODISS), which ran from 1986-1988. The ODISS planners were able to take advantage of some of the lessons we had learned at the Library. The project name still spotlights technology, using the two terms o*ptical storage* and *digital imaging*. Some of the ODISS images were scanned from microfilm. In the years that followed, many others have scanned from microfilm; the ODISS project was a good trail blazer for us.

The ODISS project put two hundred thousand Confederate Civil War Muster Records from Tennessee on 12-inch Optical Disks stored in a big jukebox. (Copyright is not a big an issue for holdings like these.) One description of the ODISS project hints at motives that pertain to both access and preservation: it was a "program that tested the feasibility of substituting digital images for physical records and microform." The images were described as being *online* which, at that time, more or less meant "in a local area network." The Internet existed in 1987 but the first Web browser was still five years in the future.

Phase 2 of my story is about "access projects" in the 1990s. Access was in the front room but there was also a back room, for technology development, some of which was relevant to preservation.

In 1987, James Billington was installed as the new Librarian of Congress and, within a year, he proposed a new project: *American Memory*. This name celebrates a body of content rather than a technology. The word *memory* meant that the content would be historical, thereby nourishing the field of education and also permitting us to minimize the inclusion of materials protected by copyright.

For American Memory's five year pilot--1990 to 1994--we re-used the technology playbook from the optical disk project: laser videodiscs, this time together with CD-ROMs. We presented multiple original formats from text to sound to movies, with an item-level bibliographic database that ran on a Macintosh. We carried some old content forward, for example, the Detroit Publishing Company photos and the McKinley assassination movies, still on laser videodiscs.

From American Memory's start, one fascinating conundrum was "how shall we reproduce a book in digital form?" There were a number of ideas abroad. Some were influenced by the traditional models of microfilming and preservation-via-high-end-photocopying. The first undertakings of the important book scanning project at Cornell University, for example, gave us a great model for printing the digital images back to paper and rebinding a fresh copy of the book. This approach (like preservation photocopying) produces a *physical replica* to put back in the stacks. At the other end of the spectrum were the academics associated with the Text Encoding Initiative (TEI), launched in 1987. The *summum bonum* for the TEI practitioners was a perfect text transcription, marked-up with SGML (the parent of HTML and XML).

In the end, most of us found a middle ground, combining online images and machine-readable texts. Like the microfilm mentioned earlier, this produces a *virtual replica*, this time with greater ease of navigation and featuring searchable texts. One of the most influential models was the Making of America project, carried out at Cornell and the University of Michigan beginning in the late 1990s. The outcome of that project, as well as the results of our smaller-scale efforts in American Memory, demonstrated that this approach did not merely reproduce the book but transformed it into a new resource, more potent for certain types of research and even for plain reading, if you are a Kindle fan. Users can search for words or mine the text as they pursue new research agendas. The combined approach has been widely adopted and the details are perpetually being refined and re-refined, with a considerable back and forth among practitioners about the levels of quality to be sought for image and/or text.

Although American Memory is described as "an access project," emphasizing the goal of placing content in the hands of people around the country, we continued to explore new technology. What are the best ways to reproduce various forms of content? What would make digitization more effective and efficient? How might digital reformatting connect to the traditional work of library and archive preservation? What approaches to naming files and numbering items would serve the longer-term goals of *digital archiving*? (Videodisc frame numbers were our starting point. We put frame numbers in the bibliographic database to link the metadata to the pictures.)

Central to our explorations in the early 1990s were digital imaging, text conversion, digital audio, the new field of digital moving images, and a panoply of issues pertaining to search software and CD production. But another technology was about to have a big impact and would soon fuel the fires of digitization: the World Wide Web. The first Web browser to catch on--Mosaic--was released in 1993 and soon the Internet was prominent in the thinking of libraries and archives, to say nothing of the public.

Our timing at the Library of Congress was perfect. When the Web came along, we had content ready to go. Our 1995 move from the American Memory pilot to what we called the National Digital Library Program marked our jump from disk media to online access. We successfully re-presented our content on the Web, including the migration of our electronic-but-still-analog photographs and movies into digital form. (Digital versions are still online today.)

In Phase 3, access and technology development are *both* in the front room, and there is increasing interest in the implications for preservation.

By the year 2000, we were spending a lot of time discussing reproduction quality, for all types of content. And there is no better sign of the movement of this topic from the back room to the front than the 2004 publication of the excellent National Archives imaging guidelines, by Steve Puglia, Jeffrey Reed, and Erin Rhodes. The authors are very careful not to use the *p* word. Read the title with a lawyer's eye: *Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files - Raster Images*. The phrase *for electronic access* is intended to signal that no claim is made that the images described here are "for preservation." And the term *production master files* is also carefully chosen. The authors write, "In order to consider using digitization as a method of preservation reformatting it will be

necessary to specify much more about the characteristics and quality of the digital images than just specifying spatial resolution." (p. 66)

Well, true enough. There is more to image quality than spatial resolution and more to preservation than just reproduction quality. In fact, some things about those two "mores" are still being worked out, as I will discuss in a moment. But I thought the authors were too modest: simply following their 2004 recommendations will produce imagery that surpasses the quality of, say, *preservation* microfilms.

Digital quality improvements have proceeded category by category. Higher quality came to still imaging first. After a six to ten year lag, improved digital reproduction of sound recordings has also come along nicely. With moving image content, the serious investigation and experimentation with high resolution is just beginning.

How shall we assess quality? In the beginning, when scanning printed matter, we tried to adapt microfilming standards, where text legibility to the human eye is an important desired outcome. You can measure the size of the finest fine print and determine how many line pairs per millimeter--a measure of spatial resolution--your film must provide in order for 6 point type to be legible. Recently, however, some have suggested embracing a different metric when printed pages are digitally imaged, and when our post-processing includes Optical Character Recognition (OCR) to convert the typography into searchable text. OCR success can be measured by statistical means. Michael Stelmach--speaking later today --recently commented that text conversion today depends on more than just image quality. "For modern languages," he said, "I believe that syntactical information will trump most of the image quality issues." Google Books is an example: OK images, pretty good texts.

For pictorial content, there are other factors. For example, tonal range, aka bit depth, is as important as spatial resolution. Consider our 1862 glass negative of President Lincoln and General McClellan sitting in the front opening of a tent at the battlefield at Antietam. There is something--perhaps a blanket--hanging on a frame in the darkness at the back of the tent. Let's assume that the original negative offers some detail in the shadows. Now, if you open the digital copy in, say, PhotoShop, does it offer enough tonal range to permit a researcher to see what is hanging on the rack? Meanwhile, for other content, there is color reproduction . . . a thorny family of issues all on its own.

Quality assessment is an important topic for the cooperative Federal Agencies Digitization Initiative--to be covered in the next session. Tools are being developed to permit archives to measure factors like Spatial Frequency Response, Opto-Electronic Conversion Function, and noise. DPI (properly PPI for pixels per inch) in and of itself ought no longer be the prime metric. And we find ourselves sorting out what are called *image states*: is this version of the image "original referred" or "output referred?"

As a sidebar, during the late 1990s and early 2000s, we also explored access technologies. I won't delve into the many ways that the Web has changed but will mention the way in which we changed our online presentation of maps. The first breakthrough came from the private sector:

the proprietary MrSID format, with high quality compression and very flexible zooming. Within a decade, the same capabilities were offered by JPEG 2000. We jumped horses and migrated our access copies to this ISO/IEC standard format.

The trend to higher reproduction quality for our masters has been nudged along by one study or another. But another factor has also contributed. Crudely put, this is "what gear can I buy at the store?" Year by year, scanners and overhead cameras have kept improving and, sure enough, our images improved with them. The same thing happened with digital audio workstations, and the sampling frequency and bit depth of the files we make today have leapfrogged above the levels provided on audio compact disks. Our higher quality production has generally tracked the capabilities of the equipment on the market.

What has been the effect on the production of older preservation formats like microfilm, flat film, and analog audiotapes? We have seen two decades of incremental change. At the Library--my long beard--I can remember when the Geography and Map Division produced 105mm microfiche of maps: one map per fiche, some in color. This practice ceased during the 1990s, roughly at the same time that the division got their first large-sheet scanner. By the early 2000s, our Prints and Photographs Division produced more and more digital images and far fewer flat film copies.

There has been considerable movement away from the microfilming of books, although less so for newspapers, with their large pages. Book digitization was first led by universities like Cornell and Michigan. The more recent mass digitization efforts of Google Books and the Open Content Alliance have also advanced book scanning technology. In the first university projects, books were disbound and pages run across a flatbed scanner. Today, almost everyone scans volumes--still in their bindings--with overhead cameras. I just looked at Cornell's preservation-program Web page and see that it highlights the continuing value of microfilming, especially when done in a manner that supports subsequent scanning of the film. Meanwhile the Web site for Michigan's Digital Conversion Unit (formerly called *Preservation Reformatting Services*) says flatly, "digital imaging is now the library's preferred method."

It's not just digitization technology that has improved. Another critically important issue has also seen extensive work during this decade: ensuring the persistence of digital content over time. Although we are all anxious about long term management, having the years roll by helps. At the Library, our online offering today still includes some of the twenty-five-year-old fruits of our Optical Disk project. Another confidence-builder is the extensive, community-wide investigation of trusted digital repositories, including the valuable Electronic Records Archives (ERA) work here at the National Archives. Instead of asking, "Is this a *preservation copy*," we often find ourselves asking, "How are we to preserve our new digital resource?" That's a question with bonus value: answering it will contribute to our understanding of the preservation of born digital content: the e-books, e-mails, e-records, e-music, and e-videos coming our way.

Conclusions?

First, technology and our use of it will continue to evolve. We sometimes think of microfilming and other analog processes as having been fixed and wish for this state of fixity to return. (This memory is deceptive: specialists will remember the impact of the switch from acetate to polyester-based film.) In contrast, it is hard to see digital reformatting practices as being settled once and for all.

Second--and intimately related to the last point--there is a real synergy between what happens in industry, in the world at large, and what we do. We have seen this with document imaging, with FAX compression, with MrSID and JPEG 2000, with scanner improvements, with the Web, with XML, with the mass digitization projects. This synergistic process will continue.

Third, our analysis of reproduction quality issues for various categories of content (one size will not fit all!) will turn more and more on objectives and measurement, as in the example of General McClelland's tent or the success of OCR.

Fourth, the focus for the preservation work in institutions like the Library of Congress and the National Archives will not be limited to digital reformatting, but will also embrace the sustenance of the digital resources we have, no matter the source.

In short: there is plenty to do! Thank you.