2010 RLG Partnership European Meeting, Oxford, October 2010.

How Bad is "Good Enough"?
Mass Digitization of Photographic Archives

James Eason
The Bancroft Library
University of California at Berkeley

The following text was prepared and delivered, in abbreviated form, as part of the "Doing More For Less" panel at the OCLC-Research conference **Moving the Past into the Future: Special Collections in a Digital Age (**12 October 2010.)

- **Context / Cold Room slide**

The Bancroft Library, U.C. Berkeley, is a large collecting institution. As the primary Special Collections library of UCB, it holds rare books, manuscripts, literary papers, organizational records and personal papers with an emphasis on the American west. In addition, its collections contain an estimated eight million photographs, most of which are negatives, with most coming from professional archives and organizational records. In particular, the majority come from news "photo morgues;" the archives of local newspapers, consisting of millions of negatives by local photographers.

- **Preservation / negative slide**

Film-based media pose huge challenges, for access as well as for preservation. This is overwhelming on the scale that we face.

- **Is mass-digitization an answer?**

We do need to fully explore making archival primary materials fully available, without curatorial selection, on the internet.

We must evaluate the cost of mass digitization with **access** as our over-riding goal. Given the expense, secondary goals must be considered as well.

- **Special considerations for photographs**

For photographic collections in particular, which exist in great quantity, consist of fragile media, and require a high standard of quality to be useful, we must carefully think about price points, value, and secondary goals when we formulate digitization strategies.

There are numerous important factors that justify differences of approach when digitizing photographs as opposed to textual documents.

A variety of values are placed on photographs. They are documents; specifically they are carriers of visual information. But they also have **aesthetic** value. Aesthetic quality enhances an image's power to communicate, and therefore raises the bar in terms of reproduction quality needed. Of course, photographs may have artistic or other cultural value that lead us to approach them with even more care and concern for quality.

Photographs have particular preservation needs as well. They are especially subject to damage through handling, exposure to light, and poor environment. Film in particular does not have a long life expectancy. This is especially true of early film bases; flammable nitrate of the 1920s and 1930s and early acetate ("safety film") of the late 1930s to 1950s can not be expected to last for generations. High quality "preservation reformatting", a very expensive undertaking, has long been standard for preserving early film negatives.

Aside from preservation considerations, quality reproduction is a key step in the use of photographs. You can't quote a photograph, and researchers need to be able to publish them in order to use them. High quality accurate reproduction is therefore central to management of photographic archives.

Another factor that distinguishes photographs from textual archives is that we have a tendency to justify more detailed description of images than of individual archival documents. This adds significantly to costs of digital projects, of course, and I argue it is not always necessary.

I also want to introduce the problematic concept of "archival evidence" as it applies to photographic archives. For decades now, photo archivists and scholars have emphasized the case for photographs to be seen as historical documents in themselves. They are not just illustrations sought out immediately prior to publication of a text. However, photo **archives** are often not often maintained according to the same archival principles that govern management of records and papers. True photo archives are the byproducts of the work of an organization or individual, and the relationships among the material and their filing systems (their original order) constitute evidence of how the person or organization worked. Rich ground for study **across** images in a file such as a news photograph morgue is a possibility, particularly when you consider the importance of the evolution of visual communication in 20th century news media. The archive as a whole contains value beyond the often-mundane individual images it contains, and scholars should be enabled to explore such archives as a whole, not just locate images illustrative of a given topic or event.

With these distinguishing factors and considerations in mind, and with all the challenges of overwhelming volume and extremely challenging access factors, we have been experimenting with digitizing negatives, for access, as cheaply as possible. **This means scanning *everything* in a targeted series of negative files, with no curatorial selection.**

This is controversial. A grant proposal to continue our efforts had one anonymous reviewer, in particular, who highlighted the most controversial aspects of our approach, whom I will quote shortly.

I believe that in formulating a strategy to scan photographic archives, the fundamental question comes down to:

- **Do it Once, Do it Right? or Just Do It?**

One anonymous reviewers of a grant proposal to scan negative files cheaply and quickly wrote:

> "I have fundamental objections… to this methodology and I do not feel the approach is ultimately beneficial to the materials. While it will create access to the materials, they would have to be selectively re-scanned to make preservation masters [in future]. This process seems redundant, costly, and potentially damaging for the materials…. [T]he massive amount of time and resources spent on digitizing these materials not once, but twice[,] is harmful….
>
> "I feel that this blind approach to digitization without any curation will allow for a large percentage of historically irrelevant images to be scanned with wasted resources."

(I want to pause for here and point out the phrase "historically irrelevant images". Where is the concept of archival evidence? I think this reflects our tendency to value individual images over the archive and favors a traditional approach to "cherry pick" historical high-points and justify high investment in their reformatting, to the detriment of the archive as a whole.)

Another reviewer, whom I shall call Reviewer Two, also raised valid concerns, writing:

> "This project has the potential to create a vast collection of mediocre images that users will find cumbersome to navigate in order to find the gems…. Particularly when the metadata is similar."

Reviewer Two continued, also acknowledging

> "I can also see the flip side where the continuity of the collection is maintained in the creation of a complete archive."

(*That* is what I like to hear! This reviewer did, by the way, rate our proposal very highly is spite of the concerns stated.)

- **Low-cost approach**

The approach we've been testing consists of the following cost-saving characteristics.

No curatorial selection.
Filenames and very basic metadata recorded in a simple spreadsheet.
Batch processing.
No manual adjustments to individual scans.
Trust the vendor / operator to count items in each sleeve (no item inventory precedes shipment: log of the sleeves only.)
Only unadjusted tiff files are delivered, with no derivative jpegs. (Derivates created later, through in-house batch processes.)

The **omission of individual file adjustments** (to brightness, contrast, etc.) is the most important cost-saving factor.

➢ Cost / Quality Compromise

We did, however, find that with a modest price increase over the bare minimum, we could generate more useful image files, adequate for most expected uses, by maintaining fairly high resolution and bit depth (approximately 800ppi, 16 bit grayscale, for 4x5 film transmissive originals). There are costs basic to carrying out a scanning workflow, and it made sense to add 30 or 50 cents an image to yield a much more versatile image file. Although files were not adjusted and optimized, such work can be done selectively, in future, if an image is requested for publication.

800ppi meets the minimal standards for digital imaging for content submitted to the California Digital Library[1]. 4x5 negatives and transparencies are at the cusp between a recommended 800ppi and 1200ppi. Highly detailed views, such as aerial photographs, should be scanned at 1600ppi[2], in general. 800ppi "master files" must be understood as "production masters" rather than "preservation masters", as the quality and detail of the originals can not be fully replicated at 800ppi, yet the scans can yield an acceptable print, in moderate size, for most purposes.

➢ Results

In less than a year of part-time effort, testing a vendor workflow and a workflow using student assistants, we were able to scan over 21,000 negatives and their annotated negative sleeves.

The costs, depending on the workflow, came in at approximately $1.50 per image (for scanning by student employees) to just under $3.00 per images (for outsourced scanning by a vendor).

Case Study Details

➢ SF Examiner numbers / negatives

Given in 2006, the San Francisco Examiner Newspaper Photograph Archive more than doubled Bancroft Library's pictorial holdings. The archive consists of some 3.6 million negatives, 70,000 of which are on flammable, hazardous, and strictly regulated cellulose nitrate film stock. This massive and problematic acquisition forced creative thinking and action.

➢ Priorities for Examiner

[1] See: CDL Guidelines for Digital Images. http://www.cdlib.org/services/dsc/tools/docs/cdl_gdi_v2.pdf
[2] See: Federal Agencies Digitization Guidelines Initiative (FADGI). http://www.digitizationguidelines.gov/stillimages/documents/Technical.html

A triage approach to preservation had to be our first priority. Nitrate had to be identified and separated from safety negatives and glass plates, which were interfiled. Film degraded beyond usefulness had to be removed and disposed of in accordance with hazardous waste disposal regulations.

After preservation through film separation and cold storage, the next priorities were access (as there was no index or catalog of the archive) and, concurrently, formulation of a long-term preservation plan. (As alluded to earlier, cold storage buys time, but is not a permanent solution, Preservation reformatting needs to be part of film preservation planning.)

- **Images of cold room and freezers**

Initial efforts were supported by grants from the National Endowment for the Humanities and, more recently, from the Save America's Treasures program of the National Park Service.

The most daunting and most immediate challenges concerned **nitrate storage, management, and retrieval**. This material, by necessity, was segregated, packaged, and put in freezer storage in a facility ten miles from campus. We can not serve these to researchers, and access to any box of negatives requires hours of staff time. Due to time considerations and safety requirements, there could be no appraisal or description prior to freezing. The access concerns alone make this material a high priority for scanning. But can the cost of scanning be justified? And how can *affordable* scanning fit into a preservation strategy, given that quality is essential in duplicating photographic images?

- **What IS "Doing it Right"?**

More broadly speaking, we are at a point of technological change in which there is no agreement on what it means to "do it right." Digital imaging standards are generally geared to producing "production masters" not "preservation masters". There is no agreement on digital as a preservation media[3]. Therefore standards for preservation reformatting of photographs are in flux, and ill-defined. Film-to-film analog duplication is largely a thing of the past.

Even in the era of film a darkroom photography that ended some 10 years ago, such duplication was expensive, just as top quality digital imaging is technically demanding and expensive. When undertaking such necessarily high-cost work (digital or analog), selection is implied. This adds yet more expense. With a photographic archive like the Examiner, such cherry-picking is not desirable nor is it possible.

- **Mass Digitization as a Tool**

All of these factors -- the lack of clarity about *how* to preserve; the fact that quality preservation reformatting is very expensive and selection is necessary; the prohibitive

[3] See: Frey, Franziska S. and Reilly, James M. *Digital Imaging for Photographic Collections: Foundations for Technical Standards.* http://www.imagepermanenceinstitute.org/shtml_sub/digibook.pdf

volume we face; the difficulty we face in physically examining the material -- make total digitization an appealing scenario for the Examiner nitrate negatives. In addition to these factors (and perhaps most significantly), my misgivings over aggressively weeding or selecting among this archive and destroying its archival coherence and value convinced me to embrace a mass digitization approach to the nitrate film, with the possibility of extending, in future, into the acetate files.

- **Two work flows**

The two workflows tested yielded scans for more than 21,000 negatives and annotated sleeve exteriors.

The costs, depending on the workflow, came in at approximately $1.50 per image (for students using an office-quality Epson 750 flatbed scanner), to just under $3.00 per images (for outsourced scanning by a vendor). Vendor charges were under $2.50 per image, and in-house staff time added a bit under 50 cents per image. Vendor work took place over the course of about five months (with down-periods), and yielded about 11,000 scans. Although more expensive, the vendor approach yielded higher quantity (and somewhat better quality) over a much shorter period. With cash for a vendor in hand, we could theoretically scale up to scan about 3,000 images per month. Scaling up our student workflow would be challenging, and probably require dedicating a staff position to the effort, increasing costs significantly.

With both workflows, **the most significant savings were realized by omitting individual image file adjustments.** Unadjusted scans, with images simply made positive, were the product, with no adjustments to tone, contrast, brightness, or other characteristics.

- **Spreadsheet illustrating minimal data entry**

Data entry was minimal, also resulting in high volume at low cost. Original paper negative sleeves generally contain multiple negatives. (The average is three, but there may be several dozen negatives on a single news story found in a single sleeve.) Sleeves have a typed filing code and a few identifying words. Some have much richer information, with names of all subjects, a date, the photographer's name, and an accounting of film and flash bulbs used. Using a spreadsheet or database table, students keyed the filing codes and paraphrased the sleeve data for each sleeve, but made no attempt to provide full descriptions by transcribing all the information or enhancing it. Filenames were derived from filing codes according to pre-determined conventions, and the data entry table provided them based on the sleeve code. Negatives inherited filing codes, roots of filenames, and all descriptive data from the data entered for the sleeve. (In other words, a sleeve-level record functioned as a "parent record", and each negative within the sleeve was a linked "child record" and inherited most of its metadata from the parent.)

In short, although metadata was required at the item level, most of it was derived from sleeve-level data, and very little was manually keyed.

- **What did we get? (How Good?)**

As on-screen examples go, jpegs derived from our 800ppi scans yield results that are more than satisfactory for on-screen viewing. While they may not be the best possible rendering that could be derived from the original negative, they are detailed, reasonably sharp, and I believe they will meet the vast majority of user needs of this material. Note that the great majority of news photographs from this era are portraits and groups of people. Highly detailed city views, street scenes, and crowd shots are fairly uncommon. Even in examples of large crowds in a march on city streets, such as an example from the 1934 Longshoremen's Strike, our scans yield sufficient detail to enlarge images and read business signs, car license plates, and view other small features.

So, we have certainly been successful in producing good quality access scans.

- **Is that all? (No!)**

But, is that all? No.

I believe these scans will serve as more-than-adequate production masters, yielding good quality prints up to 8x10 inches and perhaps a bit larger. (If an image is selected by a user, the master can be adjusted at the stage of delivery in order to yield the best possible results.)

The complete body of scans provides us a curatorial review tool, and the foundation of a strategic preservation strategy. This inaccessible material can not easily be viewed, even by staff, and it is very hard to assess. What percentage has a high level of visual detail that would benefit from higher quality scanning? How many are redundant near-duplicates? How many are just poor photographs that could be weeded? We can assess and rate the material (or a significant sample of the files), establish selection criteria for higher-quality reformatting work, and make better-informed decisions about weeding redundant or insignificant originals.

In addition, we have digitally recorded the **context** for ***all*** of the images and annotations in the files. Regardless of whether all the originals are useful and worthy of future publication, relationships among material can be seen in a way never before possible, and these relationships have been digitally preserved. If our future plan results in weeding of inferior near-duplicates or otherwise reducing the quantity of flammable originals we are storing, the digital files will preserve the complete context to a reasonable level of quality.

- **Screen shots of review tool**

An in-house curatorial review interface has been developed that allows staff to view a sleeve exterior and its keyed data, and thumbnails of all the negatives it contains. Each negative can then be enlarged, viewed on screen, and rated for image quality, historical interest, and redundancy. Images can also be flagged for privacy concerns. (Some images, such as victims of domestic violence, sexual assault, or attempted suicides, should not be published on the internet. Catching these instances and reviewing them in light of our web policies is necessary.) Ratings can be exported to spreadsheet form and analyzed.

- **Preservation strategy**

Review of about 2,000 of the scanned images suggests that a small percentage are candidates for investment in higher quality reformatting. I'm projecting that about 2%, and perhaps fewer, will merit re-scanning.

The question of "What is to be preserved, and to what quality level?" is an important one. I believe we will settle on three or more quality tiers, depending on our goals. The low-cost initial scans serve to preserve context (and permit access). An adjusted and optimized 1200ppi scan will be adequate to serve as a preservation master for a small percentage of selected images of obvious historic interest but little visual detail (portraits versus crowds or scenes). Optimized 1600ppi scans may be appropriate for the tiny minority of items with historic interest and great visual detail (cityscapes, crowd scenes, aerial views).

Our initial low-cost digitization of the all of the negatives presents us with another preservation opportunity. We are considering the value of analog output of the entirety of the files, as a more stable record of the collection, in the form of 35mm output from unadjusted scans.

- **Comparison of reduced film output**

Now that technology has brought us to a largely post-film era, traditional dark room duplication of negatives is nearly obsolete. It is possible, however, to create film output from digital files, and produce analog copies that can be considered "preservation masters" with a digital master file intermediary. Such a costly process should, of course, be done to the highest standards if a preservation master is the goal. (4x5 copy negatives from digital files, from an LVT film recorder, can cost as much as $50.00 each; some savings can be expected on large-scale projects.)

However, my interest in preserving the context of the entirety of the Examiner nitrate files led me to wonder what might be achieved through low-cost small format (35mm) film output from our unadjusted low-cost scans. (Admittedly, there is a lot wrong with this idea: reducing size to 35mm reduces quality; why lock inferior, unadjusted images back into analog form without adjusting them to improve their quality as much as possible?) Still, I was curious enough to try a few tests by having about a dozen frames (from the 800ppi unadjusted scans our vendor created) recorded to 35mm film. To continue the test, our in-house imaging lab re-scanned this 35mm film to test just what sort of image quality we would get. (Think of the worst-case scenario in which the digital files are no longer available **and** the original has not survived. If all we had were this 35mm shadow of what the original had been, would it be at all useful?)

The results, while less than ideal (as expected) were actually surprisingly good. We have certainly lost aesthetic value. In a detailed street and crowd view (the 1934 labor march viewed previously), the image is less sharp, the highlights are a bit dingy; when compared closely, the 3rd generation copy is clearly inferior. Yet, the image at first glance, on its own merits, is acceptable. When zooming in you can still see a remarkable amount of detail. Signs on buildings are still legible, and small faces in the crowd can be seen clearly.

So, we are left with a cost/benefit question that hinges partly on our confidence in digital as a preservation medium. Is it worth doing complete 35mm film output (perhaps at a vendor cost of one to two dollars per frame at a bulk rate) from inferior and unadjusted, unoptimized scans in order to provide an analog “security blanket” that records archival context and provides marginally useful images? I’m not sure of the answer. While $70,000 is a relatively small investment for analog copies of 70,000 negatives, it is still a lot of money that we don’t have, and our staff time to manage the material and link reel and frame numbers in our metadata also must be added to the total. Still, the possibility is somehow compelling.

➢ Reiteration of approaches

To summarize, our low-cost approach reduced expenses through batch processing, minimal description (at the sleeve level), and leaving image files unadjusted (with no improvements by the scanning technician).

A typical project done to higher standards would have these characteristics:

- High res (1200 ppi +)
- High bit depth
- Huge file sizes
- Manually adjusted
- Quality control
- $12-$20 / image vendor cost

Our approach, by comparison:

- High-ish res (800 ppi)
- High bit depth (16 bit)
- Large-ish files (22 MB)
- Batch processed
- Batch validation
- $1.50-$3 / image

➢ What’s next?

Our next steps involve:

- User interface
    - Assess impact and search / navigation issues
- More funding to continue scanning
- Further assessment of film output approaches