The Other Side of the Computer: Spending a Summer with Digital Collections.

DANIEL DAVIS

During the summer of 2010 I was granted a sabbatical from my day-to-day duties as the photograph curator in Special Collections and Archives. My main goal was to create a complete catalog with visual representations of all the photographs taken by Andrew Joseph Russell of the Union Pacific Railroad in 1868 and 1869. You've probably never heard of Andrew Joseph Russell but I bet you've seen some of his classic images of the building of the first transcontinental railroad which have been reproduced in textbooks, coffee-table books, and documentaries about the American West. Russell's photographs are scattered from coast to coast. Most of the original negatives are held at the Oakland Museum of California, with major collections at the Union Pacific Museum in Council Bluffs, Iowa, the New York Public Library, the Special Collections at the University of Iowa, the National Archives and Library of Congress, and the Beinecke Library at Yale. As well, private individuals own significant collections of his images. My initial, and naïve, thought was to create a digital collection that would have links to digital representatives of all his images. Why this won't work could easily take 10 minutes or even an hour in and of itself but suffice it to say that fiscal restraints and archival territorialism won't allow it to happen.

Lacking the funds to travel to all of these places, I resolved to find what I could on the internet. Ironically, I went from a creator of digital exhibits to a user overnight. In fact I had spent the previous 12 years creating digital exhibits (As an aside I should note that those who create digital collections frequently don't use the collections or they don't do broad-based research across a variety of institutions or even sometimes don't double-check the exhibits they create after uploading the content). Now as someone on the other side of the computer I was constantly surprised at how small details could create roadblocks for me. So, because I have only 10 minutes I better get into my complaints.

1. Digital Collections that are deeply buried in a larger Website. You try to find Russell's Civil War photographs at the Library of Congress, or his stereoviews in the online Central Pacific Photography Museum. I know you can search individual websites in Google's advanced features, but I think that less sophisticated users need good pathways and linkage to the actual material in the institutional holdings.
2. Different digital collections with images by Russell within the same institution that have no cross-linkage. For example at the New York Public Library there are no cross-links between a collection of stereoviews by Russell and the book "The Great West Illustrated" which features his large-format photos and his text.
3. Verso-side of image not scanned - Stereoviews often have additional information on the verso. I can tell if an image is by Russell even if that information is not included on the front by the design on the verso.

4. The name of the photographer is not included in the metadata even when known. Or sometimes the name of the photographer is listed as O.C. Smith when O.C. Smith was just a publisher who used Russell's images without attribution.
5. How is it that Frank Leslie's Illustrated Newspaper, probably the 2nd most read periodical of the 1860s-1870s is not digitized, but some hopelessly obscure newspaper from Utah or New York is? This said I was pretty lucky because there are so many Russell images out there. They are a hot commodity for collectors so archivists have naturally come to the conclusion that images of the building of the first transcontinental RR are important enough to be digitized. Yet even here it is surprising because the two biggest collections of Russell images (both museums) have only a handful of images online.

But there's one source which falls into another category completely. At first this website seems hopelessly flawed to the point that it simply isn't useable. I'm speaking of fultonhistory.com which is a compilation of digital images and OCR'd newspapers from Upstate New York compiled by a very enthusiastic amateur historian who has a lot of time on his hands (As a second aside I should mention that Russell was from upstate New York and his local, gossipy hometown newspaper closely followed his western adventures). Where do you begin? First off the graphics are hideous. Second off the site says that 19,680,000 pages of newspapers can be searched, but the problem is that when you do a search it seems to bring 19,680,000 hits. Third off you can't search by a single newspaper, you can't search by date, you can't browse page by page through a single edition, and in some cases you don't even know what date the individual page is. Essentially each page was scanned and OCR'd as one pdf in a series based on the Newspaper title and a range of years. When you bring up the PDF you don't know the date unless it's printed on the newspaper but rather you know it's from a two or three year span (which is not nearly exact enough for picky researchers). When you type in A.J. Russell as an exact phrase you get 5000 hits (the maximum it will bring up) and when you put in Andrew Joseph Russell as an exact phrase you get 1 hit.

After quitting the site in disgust I said I would give it one more chance, and I'm really glad because the second time around I found a wealth of previously unpublished material on Russell. I described the method I finally used to successfully navigate the website to a colleague thusly, "I hopped on one leg going in counterclockwise circles while patting my head."
One problem was that references to Russell might be under a phrase like, "the photographer Russell," or "Captain Russell," or "Capt. Russell" or "The Union Pacific photographer," or "Russell's images." So I started to string together words like Nunda, and Russell, combined with words like photograph, stereograph, images, etc. and add years. Of course the OCR was inaccurate. I calculated that it only got about 1 in 3 words correct, but by putting a string of words together I was more likely to get a hit. Then I would create a big list and scan through individual listings and select those that I thought likely had something, doing a "Find in page" search on the actual pdf. Also the search display showed the text where Russell appears in the pdf and I could tell at a glance if it was my Russell. Once I figured out that the file names, which seemed at first were random numbers, had been sequential scanned by date (rather than putting the actual dates in the file names - that would have been silly and helpful!) I was able to use a file name in the search box and figure out a date for those

pages that didn't include a date. Also, where I knew there was something about Russell on a certain date I could "guesstimate" about where it would be and type in different numbers until I found the right date.

Before going further, I should posit the idea that perhaps I'm not the intended target of a digital collection. After all I'm a very specialized researcher who is both an archivist and a historian. But now I'm wondering if that's not the right way to look at it. The central question I'm dealing with is whether the time-intensive digital collections I used and created have inherent flaws that can readily be improved upon, or whether mass-digitization projects that stress quantity over quality provide a ready answer for both advanced users like myself as well as first-time and casual users.

Ricky Erway wrote that, "Mostly we have been creating portals that lead to other portals that eventually lead to deep collections. Each collection has to be discovered and searched individually. How many of them do how many users ever find?" So even a site so highly flawed as fultonhistory.com can reveal hidden treasures because of its 19,000,000 pdfs. In fact I've had hits on this site while doing other google searches relating to my research (As a third aside I should mention that I have never actually "googled" Russell's name). Perhaps a better approach is to digitize "for the masses" and let the user sort it out for him or herself.

I'm coming to a conclusion which has surprised even me. That with digital projects, quantity matters more than quality as long as you have broad search ability. Although I've preached digitizing only a selection now I'm rethinking this approach. In the book I'm publishing about Russell [Did I not mention the book? As a 4th aside I should mention that I have a book contract with the University of Utah Press. The book will be part biography and part catalog] I'm emphasizing using Russell's entire body of work, over 1,000 images, as opposed to the 4-5 images that are used over and over again.