Symposium: Digital Archives: Navigating the Legal Shoals

Archives in the Digital Age

Ricky Erway*

I understand my role is to convey the scope of "digital archives," so that we're all on the same page, talking about the entire range of what's included in "digital archives."

And first, I'd like to say a little something about archives in relation to special collections. Archives can refer to things like official records of a corporation or institution, personal papers, literary manuscripts, but they also contain photo collections, recorded sound, clippings and, often, combinations of these things. Special collections can refer to these same things, and sometimes archives units are actually part of special collections divisions.

While in content there are many similarities, in practice there are some real differences. Archivists typically think in terms of collections, devoting much of their descriptive efforts to providing context at the collection level. Organization of the collection into series and other hierarchical levels is key, because they seldom describe at the item level. Special collections curators come from the library tradition of describing at the item level. However, there is more and more pressure to disclose otherwise hidden collections, so they're beginning to think a lot more like archivists. Throughout this talk, I'll end up using "archives" and "special collections" interchangeably, and I think there's good reason to do that.

Online Computer Library Center (OCLC) Research undertook a survey to update and expand on information gleaned in a similar survey of ARL (Association of Research Libraries) members in 1998.[1] The recent survey extended the

* Ricky Erway is a senior program officer at Online Computer Library Center, Inc. (OCLC), the world's largest library cooperative. She works on topics related to digitization, including rights issues, mass digitization and increasing the scale of digitization of special collections. Prior to July 2006, Ms. Erway was manager of digital resources at Research Libraries Group (RLG), a library consortium that combined with OCLC. At RLG, Ms. Erway was responsible for the Catalog of Art Museum Images Online (CAMIO) and RLG Cultural Materials, which digitized special collections from a variety of libraries, archives and museums. She was also a key player in the development of ArchiveGrid, a service that aggregates bibliographic records and finding aids describing archival and special collections. Before joining RLG, Ms. Erway worked at the Library of Congress for nine years, the last five as associate coordinator of the American Memory program, aimed at significantly increasing public access to the special collections of the Library of Congress. Ms. Erway has a Master of Library Science degree from the University of Wisconsin.

1. JACKIE M. DOOELY & KATHERINE LUCE, TAKING OUR PULSE: THE OCLC RESEARCH SURVEY OF SPECIAL COLLECTIONS AND ARCHIVES (forthcoming Oct. 2010); JUDITH M. PANITCH,

population to include a variety of other institutions with special collections.[2] There were 164 respondents, mostly from college and university libraries.[3] I'm previewing some of the early findings with you; the survey covered a lot more than the digital topics that I'm highlighting today.

The respondents were asked to describe the nature of their digitization program for special collections materials; and in this case they could check all that applied. Seventy-eight percent of them said they have done one or more digitization projects; fifty-two percent said they have an active program within special collections; fifty percent said that they have an active program library-wide; twenty-two percent said that they could only undertake digitization projects with special funding; and just a tiny three percent said they have no projects yet. We asked about large-scale digitization, which we defined as systematic conversion of entire collections using streamlined methods. Thirty-eight percent said they've already done large-scale digitization projects, and thirty-six said they intend to; only eighteen percent said they had no plans. We think that thirty-eight percent is one of the more surprising answers in the entire survey when you think about what it means to digitize an entire archival collection. In a separate question, we found another somewhat surprising answer in that twenty-six percent said that they had entered into licensing contracts with commercial firms to do the digitization and to sell access to the content.

We asked a number of questions about born digital materials. Seventy-one percent said they have born digital material ranging from under a gigabyte to 89,000 gigabytes, although there was a lot of uncertainty evident in the ways that question was answered. Forty-five percent said that use has increased for born digital materials since the year 2000. I suspect that means that they're now getting *some* use, because there probably wasn't *any* use in 2000. Forty-four percent said that they permit use of unprocessed born digital materials. This is a significantly lower percentage than those responding for other formats that archives collect. Of the majority who don't let unprocessed born digital materials be used, the most frequent reason was that it was insufficiently processed to be usable. Other reasons given were privacy, confidentiality, preservation and security. And in an additional question we asked, "Where within your institution is responsibility assigned for management and preservation of born digital archival materials?" Quite tellingly, the most prevalent answer was, "Responsibility has not been formally determined."

We asked the respondents what types of born digital content they were currently collecting or managing. And photos, audio and video were among the most prevalent responses, but institutional archival records was third. Each was mentioned by over fifty percent of the respondents. Also frequently mentioned were other types of archives and manuscripts. Websites and data sets were mentioned by twenty-seven percent and eleven percent, respectively. Twenty-one percent said they were not collecting any born digital materials. While most of the

SPECIAL COLLECTIONS IN ARL LIBRARIES: RESULTS OF THE 1998 SURVEY (2001).

2. DOOELY & LUCE, *supra* note 1.

3. *Id.*

respondents selected more than one type of format in the response, none reported collecting across all of the formats.

When asked about impediments to born digital activity, predictably, lack of funding, lack of time, lack of expertise were at the top, with over fifty percent each, and next came lack of administrative support. Many of those who selected "other" and provided comments indicated surprising levels of guilt. Some were almost apologetic: "We know we should be doing this, but we really just can't get a handle on it."

And we asked them about challenges. In a question about training needs, eighty-three percent said that they need training in born digital. Of sixteen different choices, the other two in the top three were information technology and intellectual property. We asked about the most challenging issues, and while space is always right up there at the top, born digital and digitization were the next two most frequent responses.

So, that's what archives *say*. Let's look at what they *do*. I'll try to identify many of the aspects of digital activity at archives: boutique digitization, digitizing whole collections, digitizing at scale, public/private digitization agreements, scanning on demand, massively massive digitization, audio visual digitization, born digital, web harvesting and data curation. And since I can't be all-inclusive, I'll show an example of each one and give you a sense of the whole gamut of what archives are doing in the digital realm. It would've been much easier for me to give you examples of ambitious, laudable plans, proposed practices and projects, intentions and aspirations—because there's a lot more of those than there are good, solid examples. But I did find some examples and there are, of course, many, many others.

First: A Boutique Collection. This Chicago Historical Society collection is lovingly curated with a narrative, a chronology, a bibliography, an essay on historical evidence.[4] Additionally, they reached out to Northwestern University to provide an interpretive website.[5] And here is an example of one of the artifacts.[6] These materials are highly selective and are provided in a customized portal: The Boutique Collection. A lot of early digitization activities were of this nature.

Next we have: The Whole Collection. The International School of Information Science research institute is part of the new Library of Alexandria. They've digitized the entire collection donated by the Nasser Foundation.[7] It consists of about 1,300 speeches provided in text, audio and video; over 51,000 photos and portraits; almost 2,000 documentary movies; tens of thousands of documents,

4. *The Haymarket Affair Digital Collection*, CHIC. HIST. SOC'Y., http://www.chicagohs.org/hadc/ (last visited Sept. 19, 2010).

5. Carl Smith, *The Dramas of Haymarket: Introduction*, CHI. HIST. SOC'Y., http://www.chicagohistory.org/dramas/overview/over.htm (last visited Sept. 19, 2010).

6. *Scene of the Chicago Bomb Throwing and Vicinity: Together with Portraits of Persons Convicted of Complicity Therewith, May 14th 1886*, CHI. HIST. SOC'Y., http://www.chicagohs.org/hadc/visuals/38V0410.htm (last visited Sept. 19, 2010).

7. *Gamel Abdel Nasser Digital Archive*, INT'L. SCH. INFO. SCI., http://www.bibalex.org/isis/frontend/Projects/ProjectDetails.aspx?id=cqO46kXxQR6Rzv9PVI1Cjg== (last updated Dec. 23, 2007).

including decrees, press clippings, minutes and hand-written documents; and a large number of cultural items including poems, national songs, books, coins, caricatures and stamps.[8] And it's all freely accessible.[9] And here's an example of an item from the collection.[10] Let's look an institution that's really digitizing at scale. The New York Public Library has digitized over 700,000 images, including manuscripts, maps, posters, prints and photographs.[11] When a collection is this vast, you've got to provide various ways to access it, and they do.[12] Here are some images from studies of a 1940s manufacturing site.[13]

The National Archives has set a good example for how to work with commercial partners. They identified their needs and in which ways they are willing to compromise, in advance of negotiating with a private partner, and then vetted it with the public, and now they are getting a lot of content digitized.[14] This process, however, was put in place after they had entered into a hotly debated agreement. Some have criticized this agreement with iArchives, which has an embargo, limiting access to NARA's in-archives visitors for five years.[15] So, they can provide access only to people in the reading room for five years, and the private entity has a monopoly on provision of online access to some of our nation's archives.[16] It sounds bad, but it will ultimately be free to anyone, anywhere, and, as we've all been debating this, three years have passed. So, we've only got two years before that material is freely available.

Here's the access provided by Footnote.com; they're a subsidiary of iArchives.[17] Here is one of the subset of free documents that they provide access to.[18] Otherwise, it's sort of a pay-as-you-go arrangement.

We can see scan-on-demand in action at the Amsterdam City Archives.[19] You can use the archives database to search for and consult the vast majority of

8. *Id.*

9. *Gamel Abdel Nasser Digital Archive*, INT'L. SCH. INFO. SCI., http://nasser.bibalex.org/ (last visited Sept. 19, 2010).

10. *Gamel Abdel Nasser Digital Archive: Documents*, INT'L. SCH. INFO. SCI., http://nasser.bibalex.org/Publications/publications.aspx?x=5 (last visited Sept. 19, 2010).

11. *Digital Gallery*, N.Y. PUB. LIBR., http://digitalgallery.nypl.org/nypldigital/index.cfm (last visited Sept. 19, 2010).

12. *Industry & Technology*, N.Y. PUB. LIBR., http://digitalgallery.nypl.org/nypldigital/explore/dgexplore.cfm?topic=industry (last visited Oct. 12, 2010).

13. "C.F. Braun & Co." Digital Gallery Search Results, N.Y. PUB. LIBR., http://digitalgallery.nypl.org/nypldigital (search "C.F. Braun & Co.") (last visited Sept. 19, 2010).

14. *Draft NARA Digitizing Plan Available for Public Comment*, NAT'L. ARCHIVES, http://www.archives.gov/comment/digitizing-plan.html (Sept. 10, 2007).

15. *NARA-iArchives Digitization Agreement*, NAT'L. ARCHIVES, http://www.archives.gov/iarchives/iarchives-digitization-agreement.html (Jan. 10, 2007).

16. *Id.*

17. *Lincoln Assassination Papers*, FOOTNOTE.COM, http://go.footnote.com/lincoln (last visited Sept. 19, 2010).

18. *Lincoln Assassination Papers, Proceedings of the Court-Martial May 26-29, 1865*, FOOTNOTE.COM, http://www.footnote.com/image/#6390472 (last visited Sept. 19, 2010).

19. *Introduction*, GEMEENTE AMSTERDAM STADSARCHIEF, https://stadsarchief.amsterdam.nl/english/archives_database/introduction/index.en.html (last visited Sept. 19, 2010).

documents preserved by Amsterdam City Archives.[20] In the case of many of their inventory, scans can be examined online.[21] If you are not yet able to see the content of a particular inventory, you can ask them to scan it for you.[22] In this way, together, we can open up all of their thirty-two kilometers of archives having to do with Amsterdam.

Now the massively massive. Trove is a new discovery experience focusing on Australia and Australians.[23] It provides reliable information from Australia's memory institutions. It provides integrated access to over 45,000,000 items in a range of formats, and one of the things it includes is this archived website of an anti-Olympics alliance.[24]

There are many good examples of audio-visual archives. The Library of Congress's *American Memory* has an incredible variety of materials, and that includes sound and motion collections.[25] And, here you see a couple of snapshots from an 1894 film with Annie Oakley, *The "Little Sure Shot" of the "Wild West,"* filmed in Thomas Edison's Black Maria studio.[26]

On to the Born Digital. I'm showing a dark slide while I talk about the Digital Dark Age. We have paper collections going back in time. Then there were a couple of decades during which we were wringing our hands. Then there will be the time when we regularly ingest and care for digital records. It may already be too late to recover some of the material from the Digital Dark Age. But there are many signs of efforts to turn on the lights. Those promising signs consist primarily of awareness of the issues, several instances where born digital is being collected, some instances where born digital is being preserved and just a very few where the public can actually access born digital collections.

Included in this UC Irvine collection are the electronic word-processing files created between 1988 and 2003 and retrieved from 3.5-inch diskettes during the processing of Rorty's personal papers. They were converted to PDF for both preservation and access.[27] It appears the collection is in good order, but access is restricted. The gift agreement didn't give the library permission to put the materials on the web, so they decided to create a virtual reading room that provides access according to the established procedures for in-person use.

20. *Id.*

21. *How Does it Work*, GEMEENTE AMSTERDAM STADSARCHIEF, https://stadsarchief.amsterdam.nl/english/archives_database/how_does_it_work/index.en.html#1RHf (last visited Sept. 19, 2010).

22. *Id.*

23. TROVE, http://trove.nla.gov.au/ (last visited Sept. 19, 2010).

24. *Anti-Olympics Alliance*, NAT'L LIBR. AUSTL., http://pandora.nla.gov.au/nph-arch/2000/S2000-Sep-11/http://www.cat.org.au/aoa/ (last visited Sept. 19, 2010).

25. *Inventing Entertainment: The Motion Picture and Sound Recordings of the Edison Companies*, LIBR. CONGRESS, http://memory.loc.gov/ammem/edhtml/edhome.html (last visited Sept. 19, 2010).

26. *Annie Oakley / Thomas A. Edison, Inc.*, LIBR. CONGRESS, http://memory.loc.gov/ammem/index.html (search "Search All Collections" for "Annie Oakley"; then follow "Annie Oakley/" hyperlink) (last visited Sept. 19, 2010).

27. *Richard Rorty Born Digital Files, 1988–2003*, UCISPACE @ THE LIBRARIES, http://ucispace.lib.uci.edu/handle/10575/7 (last visited Sept. 19, 2010).

Web Harvesting. There are many efforts to preserve the content of websites. Most people know about the Internet Archive, which purports to archive the Web.[28] But we always have to ask, "How broadly, how deeply, how often?" It's impossible to capture every change on every site.

Some institutions are attempting to do more focused web archiving. For instance, the British Library archives all the .uk sites.[29] Indiana University Archive captures all the Indiana.edu sites.[30] NARA harvests federal documents.[31] Other web archives are subject-specific, like NYU's Anarchism Web Archive.[32] Sometimes a web archive is event specific; the Library of Virginia is preserving web documents relating to the tragedy at Virginia Tech.[33] And here's a page from April 23, 2007, the day that the classes resumed after the tragedy.[34]

Data curation is certainly a form of digital archiving, but it is one that may get addressed outside the library or archives. Here's an example from the National Institutes of Health, the Cancer Gene Index.[35] Here's one where Cornell Library teamed with the USDA to provide access to curate agricultural data.[36] The National Science Foundation DataNet is a multi-million dollar grant program, which teams up library, IT and domain expertise to create data curation centers for the sciences.[37] The humanities are not so well funded.

So, archivists are being pulled in many directions in the digital age. And, I wanted to leave you with this thought. As if they didn't already have enough to occupy themselves, archives have to deal with what are often unusually difficult issues of sensitivity and a raft of rights issues. When our mission is to improve access to archivable collections, there's an awful lot to be done and a lot of issues to be addressed. And that's why it's a really good thing that we're together to find some ways forward.

28. *Wayback Machine*, INTERNET ARCHIVE, http://www.archive.org/web/web.php (last visited Sept. 19, 2010).

29. *U.K. Web Archive*, BRIT. LIBR., http://www.webarchive.org.uk/ukwa/ (last visited Sept. 19, 2010).

30. *Web Archives at Indiana University*, IND. U. BLOOMINGTON LIBR., http://www.libraries.iub.edu/index.php?pageId=4302 (last visited Sept. 19, 2010).

31. *Federal Web Harvests*, NAT'L ARCHIVES, http://www.webharvest.gov/collections/ (last visited Sept. 19, 2010).

32. *The Tamiment Library and Robert F. Wagner Labor Archives: Anarchism Web Archive*, N.Y.U. LIBR., http://webarchives.cdlib.org/a/Anarchism (last visited Sept. 19, 2010).

33. *Tragedy at Virginia Tech Collection*, ARCHIVE-IT.ORG, http://www.archive-it.org/public/collection.html?id=649 (last visited Sept. 19, 2010).

34. Virgnia Tech Hokies United Archived Page Collected at the Request of Library of Virginia, ARCHIVE-IT.ORG, http://wayback.archive-it.org/649/20070423190954/ http://www.hokiesunited.org.vt.edu (last updated Apr. 23, 2007).

35. *Cancer Gene Index*, U.S. NAT'L. INST. HEALTH, https://cabig.nci.nih.gov/inventory/data-resources/cancer-gene-index (last visited Sept. 19, 2010).

36. *Economics, Statistics, and Market Information System*, U.S. DEP'T AGRIC., http://usda.mannlib.cornell.edu/MannUsda/homepage.do;jsessionid=828960AC0472316F9F5E5AD1BBC34566 (last visited Sept. 19, 2010).

37. *Sustainable Digital Data Preservation and Access Partners (DataNet)*, NAT'L. SCI. FOUND., http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503141&org=OCI&sel_org=OCI&from=fund (last visited Sept. 19, 2010).