Dan Cohen's Digital Humanities Blog » Blog Archive » Rough Start For Digital Preservation

How hard will it be to preserve today's digital record for tomorrow's historians, researchers, and students? Judging by the preliminary results of some attempts to save for the distant future the September 11 Digital Archive[1] (a project I co-directed), it won't be easy. While there are some bright spots to the reports in D-Lib Magazine last month[2] on the efforts of four groups to "ingest" (or digitally accession) the thousands of files from the 9/11 collection, the overall picture is a little bit sobering. And this is a fairly well-curated (though by no means perfect) collection. Just imagine what ingesting a messy digital collection, e.g., the hard drive of your average professor, would entail. Here are some of the important lessons from these early digital preservation attempts, as I see it.

But first, a quick briefing on the collection. The September 11 Digital Archive[3] is a joint project of the Center for History and New Media[4] at George Mason University[5] and the American Social History Project/Center for Media and Learning[6] at the Graduate Center of the City University of New York. From January 2002 to the present (though mostly in the first two years) it has collected via the Internet (and some analog means, later run through digitization processes) about 150,000 objects, ranging from emails and BlackBerry communications to voicemail and digital audio, to typed recollections, photographs, and art. I think it's a remarkable collection that will be extremely valuable to researchers in the future who wish to understand the attacks of 9/11 and their aftermath. In September 2003, the Library of Congress agreed to accession the collection, one of its first major digital accessions.

We started the project as swiftly as possible after 9/11, with the sense that we should do our best on the preservation front, but also with the understanding that we would probably have to cut some corners if we wanted to collect as much as we could. We couldn't deliberate for months about the perfect archival structure or information architecture or wait for the next release of DSpace[7]. Indeed, I wrote most of the code for the project in a week or so over the holiday break at the end of 2001. Not my best PHP programming effort ever, but it worked fine for the project. And as Clay Shirky points out in the D-Lib opening piece[8], this is likely to be the case for many projects—after all, projects that spend a lot of time and effort on correct metadata schemes and advanced hardware and software probably are going to be in the position to preserve their own materials anyway. The question is what will happen when more normal collections are passed from their holders to preservation outfits, such as the Library of Congress.

All four of the groups that did a test ingest of our 9/11 collection ran into some problems, though not necessarily at the points they expected. Harvard, Johns Hopkins, Old Dominion, and Stanford encountered some hurdles, beginning with my first point:

You can't trust anything, even simple things like file types. The D-Lib reports note that a very small but still significant percentage of files in the 9/11 collection seemed to not be the formats they presented themselves as. What amazes me reading this is that I wrote some code to validate file types as they were being uploaded by contributors onto our server, using some powerful file type assessment tools built into PHP and Apache (our web server software). Obviously these validations failed to work perfectly. When you consider handling billion-object collections, even a 1% (or .1%) error rate is a lot. Which leads me to point #2…

We may have to modify to preserve. Although for generations archival science has emphasized keeping objects in their original format, I wonder if it might have been better if (as we had thought about at first on the 9/11 project) we had converted files contributed by the general public into just a few standardized formats. For instance, we could have converted (using the powerful ImageMagick[9] server software) all of the photographs into one of the JPEG formats (yes, there are more than one, which turned out to be a pain). We would have "destroyed" the original photograph in the upload process—indeed, worse than that from a preservation perspective, we would have compressed it again, losing some information—but we could have presented the Library of Congress with a simplified set of files. That simplification process leads me to point #3…

Simple almost always beats complex when it comes to computer technology. I have incredible admiration for preservation software such as DSpace and Fedora, and I tend toward highly geeky solutions, but I'm much more pessimistic than those who believe that we are on the verge of preservation solutions that will keep digital files for centuries. Maybe it's the historian of the Victorian age in me, reminding myself of the fate of so many nineteenth-century books that were not acid-free and so are deteriorating slowly in libraries around the world. Anyway, it was nice to see Shirky conclude in a similar vein that it looks like digital preservation efforts will have to be "data-centric" rather than "tool-centric" or "process-centric." Specific tools will fade away over time, and so will ways of processing digital materials. Focusing on the data itself and keeping those files intact (and in use—that which is frequently used will be preserved) is critical. We'll hopefully be able to access those saved files in the future with a variety of tools and using a variety of processes that haven't even been invented yet.

This entry was posted on Monday, January 2nd, 2006 at 3:23 pm and is filed under Libraries[10], Preservation[11]. You can follow any responses to this entry through the RSS 2.0[12] feed. You can leave a response[13], or trackback[14] from your own site.

References

1. ^ September 11 Digital Archive (911digitalarchive.org)
2. ^ the reports in D-Lib Magazine last month (www.dlib.org)
3. ^ September 11 Digital Archive (911digitalarchive.org)
4. ^ Center for History and New Media (chnm.gmu.edu)
5. ^ George Mason University (www.gmu.edu)

6. ^ American Social History Project/Center for Media and Learning (www.ashp.cuny.edu)
7. ^ DSpace (dspace.org)
8. ^ Clay Shirky points out in the D-Lib opening piece (www.dlib.org)
9. ^ ImageMagick (www.imagemagick.org)
10. ^ View all posts in Libraries (www.dancohen.org)
11. ^ View all posts in Preservation (www.dancohen.org)
12. ^ RSS 2.0 (www.dancohen.org)
13. ^ leave a response (www.dancohen.org)
14. ^ trackback (www.dancohen.org)

Excerpted from *Dan Cohen's Digital Humanities Blog » Blog Archive » Rough Start for Digital Preservation*

http://www.dancohen.org/2006/01/02/rough-start-for-digital-preservation/
