

Kris Kiesling
UT Harry Ransom Research Center

It is the best of times, it is the worst of times: Issues in Digitization

With apologies to Charles Dickens:

It is the best of times, it is the worst of times,
it is the age of wisdom, it is the age of foolishness,
it is the epoch of belief, it is the epoch of incredulity,
it is the season of Light, it is the season of Darkness,
it is the spring of hope, it is the winter of despair,
we have everything before us, we have nothing before us,
we are all going direct to Heaven, we are all going direct the other way....

I believe I'll stop before reaching Dickens' comparison between England and France (or, in our case here today, between the United States and France) and their respective leaders and political, religious, and cultural climates. And my point here is not to draw comparisons between our two countries but to make an analogy between Dickens' words and the digital epoch in which we find ourselves.

Is it the *best* of times or is it the *worst* of times?

It may be both. For libraries, archives, and museums, everything we do that is digital is an add-on to the responsibilities we have always carried out in the analog world. We cannot abandon any of the acquisition, preservation, storage, description, or public service activities that are necessary for the physical objects in our care. We also cannot ignore the demand for digital access to those objects as well as to the born-digital objects that increasingly are finding their way into our collections. There is constant pressure on us to do more with fewer and fewer resources. The notion that digitization will make our lives easier or that we will become a paperless society is utter nonsense. These days, if you don't have a web presence, you don't exist, so in a sense we're trapped. We must digitize.

A year ago two professors at the University of California at Berkeley School of Information Management and Systems released a study entitled "How Much Information? 2003."¹ In it they report that in 2002 five exabytes of information were created in four physical media--print, film, magnetic, and optical. I didn't even know what an exabyte was, and in case you don't either, let me help a bit. A gigabyte is a billion bytes or a thousand megabytes (notice that we barely speak of kilobytes any more). If my math is correct, an exabyte is a billion gigabytes. 92% of the new information was stored on magnetic media, primarily hard disks; .01% was stored on paper. While the amount of paper is still increasing, most new information on paper is created by individuals in office documents and as postal mail, and not as books,

newspapers, and magazines. Two and a half times more information was created in 2002 than in 1999. Since these statistics are two years old, imagine where we are in 2004. Probably seven or eight or maybe even nine exabytes. Are you worried yet?

The Association of Research Libraries endorsed digitization as an acceptable preservation reformatting option a few months ago. Preservation librarians have looked upon digitization with some skepticism, but the ease of migration and good backup systems have allayed fears that digital data is somehow too fleeting for preservation purposes, so this endorsement by ARL is a big step. And it's a good thing, too, since virtually all of the new information being created is digital. One might logically assume that we should begin to convert all existing information to digital formats. But should we?

The digital world can be a lonely place, designed primarily, though not exclusively, for individual participation rather than group interaction. People go to the theater, concerts, and art galleries at least in part for the social experience, even though they are experiencing the event on a personal level. Seeing a photograph of Monet's "Water Lilies" in a book or as a digital image on a computer screen cannot compare with the experience of seeing one or more of the canvases displayed in a gallery. Listening to even the most technically perfect recording of a Yo-Yo Ma performance cannot compare with the experience of sitting in the fourth row at one of his concerts, hearing him suck air through his teeth while he plays and watching him be physically transported by the music. Seeing an image of a handwritten letter on a computer screen, while it conveys the textual information, cannot compare with the tactile experience of holding it in your hand or catching a hint of perfume from the page.

Yet there are many things one can do with a digital image that cannot be done with the original. We can make the cultural materials in our care available to a broader audience—to people who might never have an opportunity to see them otherwise. We can engender an appreciation of primary research materials in children while they are learning history, art, or literature in grade school. We can enhance the resolution of faded inks or photographic images to make them readable or viewable. We can provide transcriptions of difficult handwriting as the mouse rolls over the text on the screen. It's magic! And now we have born-digital art, music, and literature that is intended to be experienced in an online environment, as this conference will explore.

In September the University of Southern California Annenberg School's Center for the Digital Future issued its fourth Surveying the Digital Future report entitled "Ten Years, Ten Trends,"² which highlights the impact of the Internet's first decade on Americans. Lest you have any doubts about the impact of the Internet, I'd like to share with you some of their findings:

- Approximately three-quarters of Americans now go online, either at home, at work, in school, in libraries, or some other location. In fact, use of the Internet is so pervasive that the stereotyped perception of users as "geeks" or "nerds" alienated from mainstream society is now dead.
- The number of hours spent online continues to increase and is now at an average of 12.5 hours per week.

- Although the Internet has become the most important source of current information, the level of credibility of that information is dropping. Further, more than 40% of users believe that only half the information on the Internet is accurate and reliable. The most trusted sites are ones that users visit regularly, and are usually web pages created by established media and the government, not web pages posted by individuals.
- Nearly 90% of Internet users feel that freedom of speech is important, but 72% think the government should not allow “undesirable” content on the web. There’s a serious disconnect here.
- Internet users are spending less time watching television. The study reports that time spent with friends and family is not decreasing, so people are buying their time to spend on the Internet by giving up TV. Of those surveyed, 14.8 % say their children spend too much time on the Internet, while 46.2 % say their children spend too much time watching television. Loss of online privileges is virtually on a par with loss of television privileges as a method of punishment.
- The “digital divide” is no longer between Americans who have access to the Internet and those who do not, but is now between those who have broadband and those who use a modem.
- Online buying is increasing dramatically, and while concern about security of personal information is still high among Internet shoppers at 46%, it is less of a concern than it was three years ago, when it was 66%, and concern about identity theft was preventing people from shopping online.
- Email is the single most important reason people go online, and yet it can be a source of great irritation. However, a new trend is that experienced email users do not answer messages as quickly or as often as new email users do. Clearly the immediacy or perceived urgency of email is wearing off.
- Broadband will change everything. Again. Broadband users spend more time online than modem users, so as broadband becomes more pervasive, one can assume that time spent online at home will increase.

Why are the results of this study significant for us? It’s obvious that the Internet has thoroughly penetrated the American psyche, culture, and economy. I heard on the news Tuesday evening that a woman in Miami, Florida, is auctioning on e-Bay a ten-year-old toasted cheese sandwich which she claims has an image of the face of the Virgin Mary on it. If this is what the Internet has done for us, I think we’re in trouble. I’m sure this impact of the Internet is felt in France as well. There is no going back and the technology is constantly changing, so we had better understand what we’re dealing with. Libraries, archives, and museums need to pay more attention to Internet trends and to their web sites as tools not only for the dissemination of information, but also for education and marketing. This is a significant change for us.

Archivists deal primarily with physical objects such as manuscripts, letters, diaries, photographs, maps, and sound recordings. That is, until recently. Archivists are faced increasingly with the storage, preservation, and description of born-digital materials as well as the materials we digitize ourselves. Standards for digital data capture are pretty well established for most formats. As with any new technology, numerous file formats

were developed early on, but the industry has settled on one or two standards for each type of data, including tiff, jpg, and mpeg. Even the metadata, or the information about those digital objects, has reached a state of standardization. The cost of storage decreases every year, which is fortunate, since clearly we're going to need a lot of it. So things are a lot better than they were even five years ago in terms of knowing which standards to use for digital objects. But there remain a number of issues that libraries, archives, and museums must grapple with in the digital realm. They include but certainly are not limited to copyright, cataloging, user expectations, resource allocation, and marketing.

I'm going to speak about copyright from the perspective of one whose mission it is to make cultural materials available for research use whether they are published or unpublished, copyrighted or not. This may be a very different perspective from those of you in the audience who are creating and copyrighting such materials. And I'm going to focus on copyright for unpublished materials because that's we deal with much of the time at the Ransom Center.

To say that U.S. copyright law is convoluted and confusing is a gross understatement. Copyright begins at the moment an original work of authorship is fixed in any tangible medium of expression, now known or later developed, from which it can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device.³ Copyright protection is automatic. If a work was created on or after January 1, 1978, copyright lasts for the life of the creator plus 70 years. For unpublished materials, if the work was created *before* January 1, 1978, copyright lasts for the life of the creator plus 70 years or until December 31, 2002, whichever is greater. If the work was then published between 1978 and December 31, 2002, the life plus 70 rule still applies. If a work has never been published, and the author died 70 years ago (in 1934 or before), the work is in the public domain. Still with me? And if we don't know who the creator of the work is and therefore cannot establish a death date, copyright protection lasts 120 years from the date of the work's creation. *American Libraries* columnist Walt Crawford refers to this as "extreme copyright."⁴ I won't even get into the special rules for corporate authorship, works for hire, or published works, because those are even more extreme.

Just as U.S. librarians and archivists were on the verge of seeing a lot of material go into the public domain, along came the Digital Millennium Copyright Act and the Sonny Bono Copyright Term Extension Act, the latter of which extended the old life plus 50 rule to the current life plus 70 rule. The upshot of all this copyright confusion is that library, archives, and museum holdings are being held hostage in the digital environment. Operating under the provisions of the fair use and library exemption sections of U.S. copyright law, we are authorized to make digital copies of both published and unpublished works for storage and retrieval purposes, but we cannot make the vast majority of our collection materials available on the Internet without first securing permission from the copyright holder. All this from a law that originally was intended to promote the growth of knowledge.

And yet, we digitize. Securing permissions to put images on our web site takes a significant amount of staff effort. Whenever possible, we use materials that are already in the public domain.

In the digital age archivists, librarians, and museum curators are learning to deal effectively with user expectations (or perhaps, following our Dickensian theme we could appropriately call them “Great Expectations”). During the last fiscal year at the Ransom Center, of the reference queries that were received as email or regular post, 96% were email, and there seems to be some expectation that these queries will be answered immediately. Several years ago we implemented a policy where emailed reference queries are fed into a queue along with those received by post and fax rather than being dealt with the moment they arrive. We also know that when we put descriptions of manuscript collections online some users expect to be able to click on a folder heading and have digitized images of the contents of the folder displayed for them. While that is a lovely idea and something we’d like to be able to do for our researchers, it is neither feasible nor necessary. We don’t have the resources to digitize all of our holdings, nor do many of them merit digitization for online access. And then there is that copyright problem.

Digitization of our holdings has an impact on descriptive practices. Archival description methods traditionally have focused on the forest rather than the trees. Unlike published materials, which are meant to stand alone, an individual’s papers are a continuum of related materials—the manuscript and its cover letter, or a series of exchanges between an author and his or her editor. So instead of describing each item in a collection, archivists characterize the relationships between the items and articulate the context of their creation.

When it comes to digital objects, however, our intention is to digitize an item only once, and to save the high resolution tiff image in perpetuity. This requires that we provide not only a description of what the item is, what collection it’s from, and who created it, but also information about how and when it was scanned, and at what resolution. What file format was it saved in? Was it captured from the original or another surrogate? Is it a single page or part of a multi-page object and, if the latter, where does it fit in the sequence? The file must be named and stored in such a way that we can find it again on the hard drive or server. Digitization requires us to provide an additional layer of description, or metadata, a layer that focuses on the tree, not the forest. While some of the information can be captured automatically as part of the scanning process, much must be supplied by staff if we are to be able to retrieve and reuse the image.

Reproducing the original as faithfully as possible also takes time. As an example, it took our staff only two days to scan the entire Gutenberg Bible, but even with the help of sophisticated software it has taken months to perform the post-processing necessary to eliminate page curvature and text distortion.

Another challenge for cultural heritage institutions in the digital age is resource allocation. Digitizing our collections for long-term storage and public access is not an

inexpensive proposition, as the Gutenberg example clearly illustrates. It's not the flatbed scanners and the cameras and the file storage that are costly (although they can be), but it's the staff resources to do the scanning, to create the metadata for each image, to design web pages through which the public can find the images, or to curate online exhibitions to explain their significance that up the ante. More than half the orders submitted by researchers for reproduction of Ransom Center collection materials are for digital images. All the work we do for in-house purposes, including publications, exhibitions, events, and press releases, is digital. There are a number of ways institutions can support these efforts—through external grant funds and gifts, by contracting with vendors, and by reallocating internal resources. Grants and gifts are wonderful for short-term, well-defined projects, and it's a pretty sure bet that when you receive a digital file from a vendor, you're going to have to tweak it. So, while useful, these strategies are not complete solutions. A robust digital program in any library, archive, or museum requires a dedicated staff, most likely a staff that will grow significantly in the years to come. A good web site, one that people will return to again and again, is dynamic and needs attention at least weekly, if not daily. This level of attention must be provided in-house. The more we digitize, the greater an issue this becomes.

The Ransom Center's web site has become one of our key outreach mechanisms. It averages about 1.5 million hits a month. It is our only public face to people who will never walk through our doors. It supports every aspect of our mission statement. In terms of marketing, it reaches our largest potential audience. Our web site tells newcomers who we are, it provides information about fellowships and volunteer opportunities, it announces new exhibitions in our galleries and other events, it provides access to descriptions of our collections and to databases such as the WATCH file, it contains online exhibitions, it makes available copies of our policies and application forms for use of the collections, it documented our recent building renovation, and it even has a mechanism to allow individuals to make monetary donations. In the coming months we will be creating online lesson plans for school teachers to use in their classrooms, enabling them to introduce students to the primary research materials that comprise our cultural heritage. We will also be adding a limited number of images to our manuscript collection descriptions—a photograph of the creator of the materials, a signature, a writing sample. These additions not only will make the finding aids more visually interesting, but also will assist researchers. Members of our Advisory Council think we should use our web site to attract new collections. What we don't do yet from our web site is sell ourselves literally—publications, reproductions of collection materials suitable for framing, greeting cards, T-shirts, the list is endless. But that is coming, too.

So is it the best of times or is it the worst of times? In the last decade digitization and the Internet have brought many changes and many challenges to cultural heritage institutions, but I think it is definitely the best of times. The changes and challenges are exciting ones, ones that we have grappled with more or less successfully. I'm quite certain that other challenges await us, but there is no turning back. So with further apologies to Mr. Dickens,

It is the best of times and the age of wisdom,

it is the epoch of belief, the season of Light, and the spring of hope,
we have everything before us, and we *are* all going direct to Heaven.

¹ <http://www.sims.berkeley.edu/research/projects/how-much-info-2003/>

² <http://www.digitalcenter.org/downloads/DigitalFutureReport-Year4-2004.pdf>

³ I highly recommend Georgia Harper's Copyright Crash Course at
<http://www.utsystem.edu/ogc/intellectualproperty/cprtindx.htm#top>

⁴ Crawford, Walt. "A Middle Ground on Copyright," in The Crawford Files column. *American Libraries*,
September 2004, p. 70.