PERSPECTIVE

Universal Access to All Knowledge

Brewster Kahle

I like the idea that there's a God somewhere looking after us. But sometimes I think it's a wrathful God. Who are the right gods? Janus sounds nice, but the idea of Sisyphus having to roll a rock continuously up a hill strikes me as more appropriate. When tending files on spinning magnetic storage systems and trying to preserve them for centuries, I think of Damocles' sword hanging over us. One major slip and it could be a very, very bad day at the Archive.

I thought I'd go over some of the Internet Archive's experience working with digital materials in large numbers from an archival point of view. What are the technological issues? What are the access and use issues? What are the institutional problems and issues? What are the copyright issues? How should this all work?

I'm happy to say that after ten years of running the Internet Archive, it's all going very well.[1] In many ways, we're *cool*. The idea of archives being popular and in demand is happening in a big way. I leave from here to go to Boston to give a keynote at Wikimania. How cool is this? That we archivists have our day? This is our day—certainly for access. And, in this time of digital transition, we will discover new forms of preservation, even if not in exactly the same way we've always thought of it, as well as radically opening up our archives. This is the librarians' and the archivists' day, if we step forward and take it.

I'd like to give you an idea of some of the gritty stuff. How much does it cost? What's it like working with different organizations? Which organizations are going to do what? What about the skills required? What can we outsource, what

Brewster Kahle was invited to give a keynote address at the Society of American Archivists Annual Meeting, Washington, D.C., 4 August 2006, as part of the plenary session on technology following Richard Pearce-Moses' Presidential Address. The presentations complemented each other, with Pearce-Moses touching on the challenges of the digital era and Kahle describing efforts to address those challenges at the Internet Archive.

[1] Visit the Internet Archive at www.archive.org.

can we not? What do we take in ourselves? Do we want to create a repository in one place or in lots of places? At the Internet Archive we're wrestling with these questions by just *trying* things.

We could actually make the dream of the Library of Alexandria a reality—the dream of having it all. The idea of having all published—and I'd even suggest the bulk of unpublished—things be universally accessible. In general, archivists are used to being overwhelmed. We're trying to deal with the small. What can we do a little bit more of next year? We should rethink the question. Why don't we put it *all* on-line? Sometimes it's a lot less expensive, because often the selection process is more expensive than the digitization process. And sometimes technology can help us find what we're seeking in a way that prohibitively expensive record-by-record cataloging cannot. Let me suggest some radical ways of thinking that could put the goal of universal access to all knowledge within our grasp.

The goal will not be achieved by a bunch of folks out in California. It's got to be a large-scale societal effort. Let me try to emphasize that it's actually possible. I'll be emphasizing published materials, even though I realize I'm in a room full of archivists.

Let's start with text. How hard is it to bring printed materials, whether bound or unbound, on-line on a large scale? Well, we started to look around. How big is the task? Take the largest library collection of published materials, the Library of Congress, which I understand holds between 26 and 28 million bound volumes in its collections. If you take the text of a book and digitize it in Microsoft Word format, it's about a megabyte. If a book is about a megabyte, 26 million books is 26 million megabytes. The units go mega, giga, tera, so it would take 26 terabytes to store all the words in the Library of Congress. In current terms, that's a computer that's about the size of this podium and costs about $60,000. So for $60,000 you could buy a computer system that could store all the words in all the books of the Library of Congress. Pretty cool. So for much less than the cost of a house, you can have the Library of Congress. Or in California, you could have a garage, or a really nice rose garden, or a parking space. But it's within our grasp to talk about having all the words in the Library of Congress on spinning storage that could be accessible from a large number of places. So then, the question is, "How do we get it there?"

It is a little bit more difficult to go through and image every one of those pages and then run them through optical character recognition (OCR), but even that process is getting much, much cheaper. If we add OCR, we end up with books that are searchable so that you can go and find things within a book, rather than just finding the catalog record or a picture of a page. Books are starting to appear on the Net with little tabs: If you search for "Boston" in a given book, these little tabs say "look here" and highlight the text.

After digitizing, we can do other things that we couldn't do before. One thing is the ability to deliver paper materials to people remotely. Take the

archives' and the libraries' stuff, digitize it and put it on the Net, and then turn it back into paper. Maybe I'm a little old- fashioned, but I like books. The idea of actually delivering books to people is, I think, a good thing. It isn't that hard to build a whole print-on-demand bookmobile. You can actually do it in a little van with a satellite dish, a printer, a binder, and a book cutter. Kids can actually print their own books. Turns out if you're printing a hundred-page book—okay that's not a very big book—but a hundred-page book costs a buck. A dollar a book to print and bind for the paper, toner, and cover. That's not the capital costs and that's not the labor, but for a buck a book you can afford to give away books. Which is kind of nice in the sense that you don't have to give out the only copy that you have. I think of the librarian's dilemma, wondering every time somebody walks out the door with a book, "Is it going to come back?" If we have the ability to print on demand, we can get around that particular problem.

Other people are doing this. Eric Eldred, the guy who[2] got to the Supreme Court challenging the copyright term extension, has his own bookmobile. There are bookmobiles in India, Egypt, and Uganda. We found out that this technology worked quite well for printing and binding in the field.

We also found out that there weren't enough good books. The real challenge is to go back and scan these materials so we can make them available to be distributed this way. We started with out-of-copyright materials to assess the institutional responsibilities, such as, "Who can do what?" And "Who *should* be doing what?" And "How much does it *really* cost?" What we found is that if you ship your books to India, it costs about ten dollars to scan one book. That's not too bad.

Unfortunately, a lot of people don't like sending their books to India. We sent 100,000 books to India, and we haven't gotten them back yet. It's not a problem with the Indians, it's this whole coordination thing. People generally don't want to send their books out of house. The Library of Alexandria in Egypt is starting to do in-library scanning. They've scanned about 20,000 books. The idea of scanning inside libraries made a lot of sense, so we tried out some robotic technologies. We found out they weren't reliable. And they were quite expensive. The experience foretells good things for these advanced technologies in the future, but we found that going back and doing a simpler system—manual page turning—was easier. So we developed our own little book scanner to get the cost per page—if you were to scan inside the United States—down to ten cents. It holds the books in a cradle and is nondestructive. It uses glass to flatten the page to get really good images, 300 dots per inch at a minimum, but more likely 500 or 600 dpi with the current digital cameras. So, at ten cents a page, if it's a three-hundred page book, it's about $30 a book. If you're trying to do a million books—which is a lot—it's $30 million. Which is more than I happen to have at the moment. But that figure is not that large by the total economics of

[2] 537 US Supreme Court 186, *Eldred et al. v. John D. Ashcroft, Attorney General*, 15 January 2003.

what our institutions spend. In the United States, libraries and archives are a $12-billion-a-year industry. So, if it's $30 million to scan a million books, you can imagine starting to think large.

We've started to build these scanning stations and put them into partner libraries, including the University of Toronto and the University of California. We'll be going into the Boston Library Consortium, the Boston Public Library, and a number of other places as well. We'll be like an in-house service bureau. They could have done it themselves, but they actually wanted an in-house service bureau at ten cents a page. That's been working out very well, and we've been getting sponsorship from Yahoo! and Microsoft to digitize these materials in such a way that they're openly accessible, as opposed to digitization that goes into just a single commercial company's offerings as we've seen in some other projects. I think we've got some opportunities to get sponsorship or support for the early days, but in general it's going to be our responsibility to bring some of these collections forward. I would suggest that the cost of digitizing some 26 million books, at $30 a book—even if it were all done in the United States—is $750 million. It doesn't have to be done all in one year, but in time we could get all the books of the Library of Congress scanned and put on-line. Pretty neat. It's doable.

We started doing some work with the University of North Carolina, testing digitizing single pages, such as loose-leaf papers. We're still in the early stages to see if we can get that cost down to ten cents a page and do really high-quality imaging. In some ways, this project is reminiscent of microfilm, but with color, better resolution, and better accessibility.

So, audio. If we're trying to do the same thing with our audio collections, how much is there? As best we can tell, two to three million disks, 78s, long-playing records, and CDs have been published. Again, a couple of these podiums full of computers would do it; it's not that much stuff. The question is, "*Can* we do these sorts of things?"

We're working in a fairly litigious environment, especially in the published music area. However, we've found that a lot of communities aren't served very well by the current publishing industry. Niches of people who live by their music—people who aren't Madonna or Nirvana—have an interest in getting their music up and out. We made an offer to musical communities to host their materials. We provide unlimited storage, unlimited bandwidth, forever, for free, for those who want to share materials that, broadly defined, belong in a library. Lots of people are taking us up on that offer. They like the longevity of being archived. They also like the savings in bandwidth bills.

In this country, it generally doesn't cost you to give something away. If you give something to the public or to a charity, not only do you not get taxed for it, you get a pat on the back and a tax deduction. Except on the Internet. If you put something really popular up on the Internet, and it gets blogged, and somehow

it becomes very popular, your bandwidth bill from your ISP can go through the roof. And you could go broke. This makes no sense. The idea that it costs people to give things away makes no sense at a societal level. So, we in the library and archives world, I think, have a role to play. Let's provide room on our shelves for those who want to provide these things. They'll do a whole heck of a lot of work cataloging and doing all sorts of interesting things to make these works available. They'll do a lot of work for us if we do this *quid pro quo*.

The community of rock-and-rollers really took us up on our offer. In the rock-and-roll world, the Grateful Dead started a tradition of allowing people to record their concerts and pass around their tapes as long as no one makes any money. The key is: No one makes any money. You're allowed to pass around the tapes. Lots of bands copied them. Thousands of bands copied them. There's a whole community of people who trade concert recordings—bootlegs. When the Internet came along, they started trading over the Internet. We went to this community and said, "Would you like unlimited storage, and unlimited bandwidth, forever, for free?" They wrote back, "We don't believe you." So we said, "Try us—we know how to do big servers." People have been uploading their musical recordings. We went back to the bands and asked permission. Because trading tapes is different from putting something on a Web site, we went back to the bands and asked, "Is it okay to make this available?" The key thing we heard back from them was, "Are you going to make any money off it?" We said, "No. We're a nonprofit; we have no ads, no nothing." And they said, "Okay. Good. Let's give this a shot." Now we have over two thousand bands and over thirty thousand concert recordings. We have everything the Grateful Dead ever did. All these recordings are available on the Internet Archive, and it's providing both preservation and access to a type of material that was very difficult to get to before.

So the audio collections are small enough that we can do the whole thing. Audio, I would suggest, is within our grasp. We're not technologically limited. It's finding the niches that really make sense at this particular time that's the key component. And then, there are the rights issues, but I won't go into that much.

Within your collections, I'll bet you have a lot of audio recordings. It costs about $10 to digitize a long-playing record if you do them in bulk. That's not too bad. It costs quite a bit more if you're dealing with old tapes. But if you're dealing with relatively modern tapes and you can go through them without restoration, you can do it in bulk. Look into making your audio collections available. It's worth putting them out there.

We've also found that if you put things out there in a nonprofit setting, it works for people in the sense that they don't gripe. The idea of opt-out as opposed to opt-in—putting it up and then if somebody complains, taking it down—works very well in these sorts of communities. I would suggest being a bit bold and making things available, as Richard Pearce-Moses is coaching us to do.

Now, moving images. How many are there? If you take the Hollywood films, theatrical releases, I'm told there are between one and two hundred thousand. That's the universe of works designed for theatrical release. These works are quite heavily mined and used, so the rights issues are pretty thorny. We could get about six hundred works up on the Web site because the copyright registration lapsed. They're in DVD format, as well as other lower-resolution formats. Lots of old Westerns and so on; they're quite popular.

What we've found, actually, is that the archival films are a really big boon for us. I was introduced to Rick Prelinger, who runs one of the largest private film libraries, and we did a collaboration where the Internet Archive paid for the digitization of his top thousand films and put them up on the Internet for free. He supported himself from his library, and he found his business did better. People continued to go back to Getty Images[3] and pay fees to be able to get to his works, even though these same works were available on the Net. He expanded the program to two thousand films.

We're finding that lots and lots of people, untraditional people, are using our archives in ways that we've never imagined before. This stuff is popular. We've gotten over two million downloads just this year from just the Prelinger collection. Two million! And they're using them in untraditional ways. They're making mash-ups and other fun things. They're making new music videos. They're learning from them. We're injecting the past into the present in an interesting and accessible way to millions of people.

So, it's all working, both from a business model sense and a cost sense. We've gotten digitization down to about $15 per video hour. We helped put a guy in business based on $15 an hour. Almost all of us have cupboards full of videotapes; if you just send them to this guy, he'll put them on-line for $15 per video hour. It's cheap. Right? You can get a thousand hours done for $15,000. And we'll do all the hosting. So the cost of doing bulk digitization of video is quite inexpensive. We've even done some with 35 mm films, and that's on the order of $100, $150 per hour. Again, doable. Is this perfect restoration? No. But it's great access.

Take television. We estimate about four hundred channels of television. We've been archiving twenty channels of television twenty-four hours a day at DVD quality. It is like a big TiVo box, if you will. We hit the record button and started recording twenty channels, Russian, Chinese, Japanese, Iraqi, Al Jazeera, BBC, NBC, CBS, ABC. For a couple of years, we've only made one week available, 11 September through 18 September 2001. It was the news from around the world, what the world saw. We put that out one month after September 11th. So the idea of taking our archives and injecting them back into the common discourse immediately is within our grasp because of this

[3] See http://gettyimages.mediaroom.com/index.php?s=company_overview, accessed 22 January 2007.

technological capability. Collecting on the large scale of television is also within our grasp, even in terms of the finances of small organizations such as ours.

We've found that there's a real growth in new types of movies. Because people have cell phones and video cameras, they're doing different kinds of movies. Again, we've offered free hosting for them. We've been getting all sorts of things. There's this whole Lego community (who would have thought?) that makes movies with Legos as actors. You know, still cameras and sound effects. Some of these are cool. Some are terrible. But they're small, and it's easy to help support that community, as well as a lot of political videos that are now being put up on the Net. We're starting to get large numbers of universities putting out their archives of lectures. What a great thing. So a kid in Uganda who's a real math whiz could start to learn from the great university lectures. Isn't that right? If it cost $15 a video hour, why not? So I would say that video is also doable.

We've been working on software as well. We believe there are only approximately thirty thousand titles. Most of the issues in their preservation pertain to law and policy. Technical issues include how to emulate the old platforms. Some of the best archival work is being done by underground communities that support gaming. We can help those guys by giving them an umbrella to work under.

I've tried to look at these different media and show that they can be brought on-line. We're probably best known for our Web collection. We have 55 billion Web pages. We take a snapshot of the Web every two months. It's a full public snapshot of the Web since 1996. We can show you what Yahoo! looked like in 1996. Or Pets.com. In fact, most people use our Wayback Machine to look at their old stuff. We get a couple hundred thousand users a day, which we're very proud of.

Preservation

How do we preserve *and* facilitate access to the materials? That is a real problem. If there's one lesson from the first library of Alexandra—which is probably best known for burning—it's "don't have just one copy." Living on the San Andreas Fault line in San Francisco, we're aware of what can happen. As a result, we've built relationships with other organizations and are giving copies of the things that are most precious to us to people who are as far away and as different from us as we could find. A copy of the Internet Archive is at the new Library of Alexandria in Egypt. We've given a copy of our collections to them, and they're giving a copy of their collections to us. We've also started the same kind of thing in Amsterdam.

So preservation means "make copies." It may be the best way to start. There's a lot of hand wringing. You're going to lose stuff. So go and put things in other people's hands who are friends enough that you can go to them and say, "Mind if I get a copy of that back?" These digital technologies erode very

quickly. The current digital technologies only last about three years. In the last ten years, we've moved—transitioned—our materials three times. It's painful. And lossy. It's very difficult. We're in the process of doing it again right now. So I would suggest you put it under different administrations so that they'll have different faults than you do. At some point, maybe there will be some digital technologies that work for the long term. But demands that technology companies think about what we archivists need are unlikely to be heeded. They're going for the commercial sectors. They don't really care about the longevity of this stuff. They just want to get their Microsoft Word documents fast. So we get to suffer and use these technologies as best we can. Copying things forward within our own organizations is key. The other is making copies in far away places.

Access

We've tried different access methods. I've said we've gotten a couple hundred thousand users a day, which is pretty good. It should be more. We've put the Wayback Machine up, and you can go and see old Web pages. We've tried some different search services that use time-graphs on top showing the frequency of use of particular words that you're using. I think time-based search is going to be pretty interesting as a mechanism to bring broader access to these materials.

Will we live up to this opportunity? I don't know. A bunch of us are really throwing ourselves at it. We're trying really hard. But it's got to be a much broader effort. We really need help. We want it to be in the public sector for all the reasons of longevity and openness. If you could help, that would be fantastic. If we can help you, that would be fantastic. It's not as hard as you might think. It's not as costly as you might think. One of the key areas is to be bold and try things.

What are some of the issues? The network layer is kind of working, and we have some issues at the software layer. We're facing unknowns at the content layer. Will the future of these libraries and archives be public or private? Is it going to be Elseviers, ProQuests, and Googles, or is it going to be the Library of Congress, NARA, the Internet Archive, and similar places that will really provide the service layers to get to collections? It's a big unknown right now. I think there's definitely a role for the commercial guys. But libraries and archives really have to keep our roles whole and moving forward because we have a very different point of view than the commercial guys. Will access be open or proprietary? Will it all go under digital rights management, or are we going to help push some of these systems to be more open to fulfill the democratic ideals that are baked into our profession? I'd say those are a couple of the big, open questions that we haven't figured out even as the technologies are moving along.

One organization that might help is the Open Content Alliance.[4] It's an organization of technology companies and libraries that is working on the issues of dealing with rights to our holdings. How do we build joint collections with joint service models that make sense so that people don't have to poke into every random archives server? We can have joint collections that users can search in groups so that people can get what they're looking for without having archives and libraries feel like they've sold out. How do we strike that balance? The Open Content Alliance is attempting to achieve it.

In conclusion, I argue for universal access to all knowledge. I argue that it is within our grasp financially. It's within our grasp technologically. It's within our grasp politically. It's a great project to work on. In many ways, we've been working on it for centuries. It's a project in many ways we're *all* working on. Technologies make it possible to do things that, with paper, were very, very difficult. Universal access to all knowledge is possible, and I'd say it could be measured as one of the great achievements of humankind, along with putting a man on the moon or assembling the Library of Alexandria. I think our generation could bring universal access to all knowledge, and that's something we'd be proud of for centuries.

[4] See http:// www.opencontentalliance.org/, accessed 22 January 2007.