Bringing Your Physical Books to Digital Learners via the Open Library Project

Brewster Kahle
Internet Archive, brewster@archive.org

Brewster Kahle, "Bringing Your Physical Books to Digital Learners via the Open Library Project" (2017). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/10.5703/1288284316710

Bringing Your Physical Books to Digital Learners via the Open Library Project

Presented by Brewster Kahle, Internet Archive
Moderated by Ann Okerson, CRL

The following is a transcription of a live presentation at the 2017 Charleston Conference.

Ann Okerson: Good morning. Having entreated Brewster to attend the Charleston Conference, I have the opportunity and the privilege to introduce him for this morning's plenary keynote session. Brewster is a passionate advocate for public Internet access and a successful entrepreneur. He's spent his career on a singular focus, which is providing universal access to all knowledge. He is the founder and digital librarian of the Internet Archive, one of the largest digital libraries in the world, which serves three to four million patrons each day. With 170 staff members in the U.S., Canada, England, and China, and digitization centers at LC, Princeton, the University of Toronto, and the Boston Public Library, the Internet Archive works with more than 500 libraries and university partners to create a free digital library accessible to everyone. You're going to hear later about the Internet Archive/MIT project here at this conference. There's not sufficient time in this introduction to describe the extent of Brewster's activities and contributions, so look up his bio, which is easily available online. Meanwhile, here's a story that Wendy Hanamura, from Brewster's staff, had sent to me and I liked it so much that I'll use it.

Back when he was an undergraduate at MIT, he was crossing one of the bridges over the Charles River with a friend. The friend asked him, "Brewster, do you believe in utopias?" "Yes, well, I do," Brewster replied. "Are you a technologist?" his friend asked. "Of course," Brewster said. "Then why aren't you using technology to build ideal societies?" And that got him to thinking. What could I do, he thought? Well, I could build the Library of Alexandria for a digital age, but that's so obvious everybody's going to do that. So he went off and worked on privacy instead. But after a while, he noticed that nobody was building that digital library. So he started developing the tools and technology that he would need. Supercomputers and thinking machines, a Web publishing platform, Web crawling and archiving technology at Alexa, which he eventually sold to Jeff Bezos of Amazon. That gave him the capital to start one of the world's largest nonprofit digital libraries: the Internet Archive. For the last 21 years, Brewster Kahle has focused with single-minded attention to building that Library of Alexandria 2. He has preserved 300 billion webpages in the Wayback Machine. But there is a missing piece of the 20th century that keeps him awake at night, and that's why he's here at Charleston to talk to us about the urgency and importance of that missing piece. So we're honored and really pleased to have you here. Thank you for coming.

Brewster Kahle: Thank you, Ann. Actually, the way I think of that is we're going to have to build this darned thing together. The idea of going and having a guy stand up here and saying "I'm going to go and deliver the Library of Alexandria Version 2" is just not the way that this works. We're going to have to pull together, and I think that we've got a solution to one of the things that has really been a problem and trying to figure out how do we get there? How do we have a complete library of everything when there are costs, when there are legal issues, when there are roles and responsibilities, and how do we make it so that there are lots of winners? At the end of the day we were lots of libraries. We want lots of publishers. We want many, many more authors and everyone a reader. If we do our jobs right, we will get that kind of solution. What I would like to talk about today is how we can all become open libraries. What's the method of doing that? And I'll try to define that and the steps that we could take.

First, a little bit about who the Internet Archive is. Our motto is "universal access to all knowledge." We have this awesome building in San Francisco. Please visit. And we've collected a lot of material so there are just blithering numbers, just big numbers of software titles, moving images, going and digitizing these, uploading them, and having lots of audio recordings being done at scale. Recording television, off-air television, 60 channels of television, 24 hours a day in DVD quality in 25 countries. It's doable within even a relatively small budget to be able to do this. And we've been working with 500 libraries, we've digitized about 3 million books and put them into e-book form and tried to make it so that people can have access to them. But most of these have been before 1923. We are probably best known for Web, archiving the World Wide Web. It's big and it's very popular and it's become a very important

https://doi.org/10.5703/1288284316710

resource for many for personal use, for institutional use, even people are finding that as administrations change in the United States and elsewhere, things go up and down. Our whole newspapers blink off-line and since we're now—we've been so oriented into pouring our lives into the Web, the average life of a webpage is only 100 days before it's changed or deleted, so we have to really be on our game to be able to do this. And we do this now with 500 other institutions, so we have 500 book partners and 500 Web partners where these libraries basically build subject-based collections to make sure that those things are done and put together well. One of the great things about librarians is they are picky. I know. I am one. Where it's like "Ahhhh! It's gotta be right!" So, these collections are done very well out of a medium, the Web, that was not really designed to be archived. All in all, it's a lot of bytes. So, it's an enormous collection.

Archive.org; you can use it. It's free. In general, we have been paid to give things away and that approach works well on the Internet, so it's how libraries work. We buy books, we acquire materials, and then we make it publicly available. But as Ann said, we're making progress in a lot of ways but there is a problem. And I'm going to actually start this about a story about Wendy Hanamura, director at one of our partnerships, who's here. She is Japanese American and there was a book that was very important to her growing up. It was *Executive Order 9066*. And when she got it as a middle school student in Oakland, California, it was the first time that she knew then what her parents and her grandparents were describing as "camp." It had barbed wire. They were in a Japanese internment camp and this was a picture book that was very important to her for understanding her past and she got it out of her library. But at this point, her son, who is now in college, was cramming for an exam or doing a paper on race and ancestry and this would've been a perfect book for him. But if you were to go to Hathi Trust or the Google Books Project you would only get a snippet of it. The only place was not in those other repositories and if it's not digital, it doesn't exist. He really wasn't going, even though it was probably in Widener somewhere, he wasn't going and getting it out of the blue. It had to be available online. And it turns out that we had digitized it and it's available on the Open Library website so one reader at a time can borrow this book, and he was able to see this picture book to get sort of a relationship to a book that was important to his mother and it was making it forward into a new century, so the lessons of the last century were making it forward. I would say that is a good story, but we're surrounded by bad stories.

This is the Internet Archive's book collection by publication date. We're doing pretty well up to 1923. You guys have participated in getting it there. And we're doing pretty well when it comes to the 21st century, but there is this missing century of books that are not there and it's also not there in Amazon. This is a study of books that are in their collections so there is this gap. It's an important gap and if we go and bring up this next generation on works that are before 1923 and completely current, we're going to get the generation we deserve. They're going to come up without knowing a lot about the 20th century and that can be a problem.

So, I would suggest we need to find a way to fix that and fortunately we do. What we're proposing is working together to build free digital access to 4 million more books, and then that would give us a Yale, a Princeton, or a Boston Public Library–class library available to all. But how do we then do that in a decentralized way? And how do we do that legally? So, the approach here is to do what we have always done. We buy and lend; and so what we could do with e-books is we buy and we lend them. It's what we do. But, what about all those things from that missing century? Well, a lot of those aren't available in e-pub form, so we buy what we can and digitize what we have to and we then lend it one reader at a time electronically. You say, "Gosh! That's pretty lame. I mean, here we are in the digital age. Anybody can have access to it." It's like, yeah, but there's this copyright thing and trying to be respectful to publishers and authors and trying to move our collections forward. This format shifting, if you will, works. And we've been doing it now for six years on the Open Library site with 500,000 books with a large number of libraries including the Boston Public Library digitizing their in copyright non-rights-cleared books and lending them.

What would that then mean? What we want is all libraries to go this way. So what we would like to do is wave a wand over all libraries and turn all of your collections digital. Then your patrons would have a choice between taking out the physical book or the electronic book but, again, under the same kind of restrictions of one reader at a time. If we could do that, then our friends at Red Hook, New York, where they now have 40,000 physical books, they would have 40,000 physical books at the end of the day and 40,000 e-books, and again only one would be

circulating, so the idea is to use the collections that we've already used, that we've built and invested in and our communities are based around, and just offer something a bit more appropriate for the current generation. We are working with the Digital Library Federation, the DPLA, and the ALA to help basically guide this to go and make sure that a great set of books are done, so it's a diverse set of books, and it's representing our communities as they are now, not actually how they might've been 50 years ago, and there are studies that say, "Well, our collections aren't as diverse as you'd kind of imagine," so maybe we can make some changes to that as we're going through and picking these 4 million books. And 4 million is a lot, so the idea of picking the highest ones that overlap with all of our collections and finding the ones that are in current curriculum and those sorts of things to make those books much more available. Again, we buy the e-pubs when we can, but we digitize the rest to make sure that we've got a complete digital library. We are working with MIT Press, and we'll be talking about that soon after this session where they've actually explicitly signed on saying yes, this digitize and lend works for them, and they're bringing their backlist to a scanning center and we are digitizing that backlist and it is working for them. Houghton Mifflin Harcourt has their archives at the Boston Public Library and we're now digitizing their backlist for lending. So this idea of digitize and lend is working for publishers, and I'm hoping it can work for you guys as well, not just the libraries, but the authors and publishers can bring forward their last century of wonderful materials to a broader use.

We want 119,000 libraries to do this. We don't want one library, even if it were, don't elect me king, either, I would say. We want it such that it's a decentralized, distributed approach with different collecting priorities and different service models. These libraries, they are in every town in the United States. They're not necessarily in every town around the world, but the United States has a huge investment in these. But there are people who can't get to them because of physical barriers. There are issues with the reader privacy when it comes to reading online that we would like to go and be our traditional selves and doing a better job of protecting than maybe going and buying books from Amazon or even offering other ways.

Long-term public access to knowledge. That's what we've always been about but in fact, based on this format shift, this technology transition, we libraries have had a hard time getting there. It's not profitable enough for a lot of the publishers to go and bring a lot of these things back into print just for these small cases, so we have to do our job to bring these forward. Buy or digitize, we want that century of missing knowledge either way. I think we're going to end up doing a lot of digitization because a lot of it is not commercially viable.

There are a couple of ways you could do this. We have funding to be able to digitize all the materials that are donated to us, so we're making the commitment, based on funding and future funding, to go and digitize everything that we physically own, so people are deaccessioning to us. We want one copy of every physical book, music video, webpage, whatever. For instance, the Boston Public Library, our friends there, is now in partnership where they've shifted their sound archives, their 78s, their LPs, that were actually sitting in the basement. This isn't their circulating set. This is their "downstairs set," and I bet a bunch of you guys have a "downstairs set" of a lot of materials, and they said it's better to go and digitize these things and so we're digitizing them and giving back the digital versions, and they also are posting them as publicly on the Internet as we can. So, you can deaccession to us or you can use local scanning centers. They're all over the place and it costs about $30 a book. It's basically labor. If you have mechanisms of sourcing the labor, then that can make it so that is even less expensive or free within your particular world. So, that's sort of the cost of getting over this hump.

What we are also working on is bringing the user experience to be as good as what the publishers are doing. Maybe we could even do better. By going and weaving these materials into our phones, tablets—scanned materials work better on tablets or on screens, but can we go and make it so that it's easy to read, to bookmark, to share, all of these materials? The blind and dyslexic benefit greatly from this, and we are finding that there are a lot more people with print disabilities than I thought. It's often dyslexia or older people that have trouble reading, so they want large-print editions that aren't fully blind, and for digitized books that works particularly well because the optical character recognition when we're reading the words, that sometimes has some faults to it. But the images of the pages you can do a lot with for many communities. So this is something we get as a by-product out of going and bringing the whole library up for lending, for everybody, we get steady open access for the blind and dyslexic. And

this requires work on all of our parts to go and make these books better in terms of accessibility.

A big headache we're trying to figure out now is how to integrate with your systems. So, we want to use the same OPACs, the same discovery systems that people have always used to try to find books, and this is going to be quite a challenge, and we're going to need your help. So, when people come to your OPACs to go and find books, let's have them help find these books as well. We also want to weave it into the Web itself so that people that are using Google that want to find a fact on Japanese internment would find that book. That if you read the Wikipedia article, we want to turn all of the footnotes in Wikipedia blue. Wouldn't it be great? You just go down and it turns blue and it turns into a link so you go and click on it and you open to the right page of the right book. That would be really dragging you in and into the library. We can build this. We also want to make it so that all journalists going forward can go and reference these materials and not just get whatever is on Wikipedia or some news site of questionable origin, but if we're going to be talking about the history of healthcare, let's talk about the history of healthcare with the best we have to offer. We have it in our libraries. Let's see if we can get it to those that are looking for answers out there.

It also makes dollars and cents reasons to do this. There is a study that says it costs $20 to get a book back from off-site storage; $20 and about 24 hours. Wouldn't that be great if that were instantaneous? Interlibrary loan, a study said, is about $35 a book or a total of $300 million a year, mostly going to UPS, shifting books around. What if those interlibrary loan requests, or a lot of them, were done electronically? We'd have happier patrons, it would be instantaneous, and that $300 million might be better used for something else: for acquiring better books, by paying for the digitization, whatever it is. So there is money in dollars and cents to do this.

The openlibrary.org website, I urge you to try it out. Try borrowing with a book! See how it works. And critique us. Be a little picky. It's like, "Hey! That didn't work quite right." We want to respond and we want this to be a community project to get going. We want your OPACs to have little e-book buttons next to every one of your books. We did this with the Mechanics Institute Library, a charming little library, and we are trying it out with some of the larger libraries so that there is a flag that says, "Hey! This is now available. Do you want it for free?" Borrow from Open Library; check out an awesome book, like this from our Mayflower ancestors from the Boston Public Library. If it's already been checked out, then you have to go and get on a waitlist and it will e-mail you when it comes back. So, that is the general idea.

Who would benefit out of turning all of our libraries into open libraries? Well, I think our patrons would. They would find new uses of all sorts of unbelievable materials. So, here is a charming little book that was at the University of California and it had been checked out three times since it was originally published many, many decades ago, but now that we've digitized it, it has been used about 1,400 times. So, this charming little book about dahlias, begonias, has actually found new life for a new readership, and some of those readers are robots, so these aren't just all humans reading these things. These are other things that are going and mining these books in new and different ways.

John Szabo, a librarian that runs the LA Public Library system, we ask him what would this do for you? And his answer was it would be more equitable access for his patrons. There are some people that use the libraries a lot and there are a lot who don't use it at all. Every kid in LA has got an LAPL library card, so if he could go and get his books more integrated into their lives, into their Google stream, into their Facebook stream, into their Wikipedia use, he would get more equitable use to his materials. He would also be able to potentially win back some of the space that he's using for closed stacks in downtown LA and be able to go and use that space for other reasons and he could save money from shuffling books around his many, many branches.

Michael Connolly Miskwish is a Native American and he is a Wikipedian and he has been participating in Wikipedia events to go and try to represent his tribes and what happened in the California tribes differently than how Wikipedia had been doing it before. This is a photograph of an American Indian shooting settlers in California, and he was able to go and find good solid references of how did the Native Americans impacted actually throughout California by bringing the library into what is currently being read in the number five website of Wikipedia. This is bringing new depth to a resource that is terrific but could be a whole lot better.

Josh Miele is blind, and so the idea of him getting access digitally as opposed to having books read to him is just a much more cost-effective mechanism to move forward.

Susan Steinway, the archivist for Houghton Mifflin, is thrilled that all of these wonderful books from a century of this venerable publisher is going to be able to see a new audience. Books that I would like to read; actually I got to see this at the Boston Public Library. It's up three stairs and it's actually in a fairly dark room, but they are all there in closed stacks, and that is pretty much the only copies of a lot of these materials from 100 years ago. They wrote in their contract that I thought was something really terrific, that they see that there is a way of balancing the interest, to going and making this out there, there is public access in a way that is respectful, supportive of authors and publishers going forward.

One of the good stories around our place is Eileen. She is the daughter of Roxanna, who works at the Internet Archive, and that family doesn't have very many books and she does a lot of her homework at the Internet Archive. But now she's got an ability to go and read large numbers of books and we've been using her to see does this work for her? And she is loving it! So, now she has access to tens of thousands, hundreds of thousands of books that she wouldn't have had in other ways.

So, who cares? There will be lots of uses as we bring our materials public. There's just been stumbling blocks, there's been costs, there's been copyright, there's technical issues about how do we do this in a decentralized way? But, I think we've got a method to go forward and bring 4 million e-books to our patrons.

What roles do we play? Publishers: let's digitize your backlist. Let's get the whole damn thing online. Let's at least do digitize and lend, and maybe you will find that some of these things are actually popular and you want to bring them back in print, but you'll get the statistics and be able to find out what deserves to be made more available.

ILS vendors: let's integrate this so that the OPACs surface these books smoothly and easily. Not easy to do, but we can get there. Writers and editors: let's go and get your books linked in to the Web. Let's make it so that these aren't just dumb old footnotes. Leverage the digital materials, link to it, and drag people into the rest of the library. Readers: what did you find important to you? What should be brought forward going forward?

In summation, we can all build these open libraries. We can build these open libraries and become open libraries together, coordinated collection development, this "digitize and lend" approach; we could use each other's technologies for distribution, in a respectful mechanism. But if we don't do this, we're going to have a generation that is going to grow up without actually a lot of access to the great materials that we have in our collections. And this is not just true for public libraries. It's true for academic libraries at every rank. The Hathi Trust is terrific, but it's not substituting for having books on our shelves. It is great for data mining, but it is not good enough. We together can build this.

Thank you very much.