Transforming Our *Libraries* from Analog to *Digital:* A 2020 *Vision*

Brewster Kahle

ILLUSTRATION BY DUNG HOANG, © 2017

Finding the Future in the Past

Today, people get their information online—often filtered through for-profit platforms. If a book isn't online, it's as if it doesn't exist. Yet much of modern knowledge still exists only on the printed page, stored in libraries. Libraries haven't met this digital demand, stymied by costs, e-book restrictions, policy risks, and missing infrastructure. We now have the technology and legal frameworks to transform our library system by 2020. The Internet Archive, working with library partners, proposes bringing millions of books online, through purchase or digitization, starting with the books most widely held and used in libraries and classrooms. Our vision includes at-scale circulation of these e-books, enabling libraries owning the physical works to substitute them with lendable digital copies. By 2020, we can build a collaborative digital library collection and circulation system in which thousands of libraries unlock their analog collections for a new generation of learners, enabling free, long-term, public access to knowledge.

The Problem

We all want to see the modern-day Library of Alexandria, a digital library where the published works of humankind—all the books, music, video, webpages, and software—are available to anyone curious enough to want to access them. I believe now is the time to build it.

The technology and costs to achieve this vision are now understood, and in fact, various projects are proving that it can be done. Three major entities have digitized modern materials at scale: Google, Amazon, and the Internet Archive, probably in that order of magnitude. Google's goal was to digitize texts to aid user search and its own artificial intelligence projects. Amazon's book-digitization program helps customers browse books before purchasing them; Amazon is quiet about the number of books it has scanned and any future plans for them. The Internet Archive has digitized more than 2.5 million public domain (pre-1923) books and made them fully downloadable and 500,000+ modern (post-1923) books and made them available to the blind and dyslexic and through its lending system on its Open Library site.

Yet bringing universal access to all books has not been achieved. Why? There are the commonly understood challenges: money, technology, and legal clarity. Our community has been fractured by disagreement about the path forward, with ongoing resistance to some approaches that strike many as monopolistic. Indeed, the library community seems to be holding out for a healthy system that engages authors, publishers, libraries, and most importantly, the readers and future readers.

I suggest that by working together, we can efficiently achieve our goal. This will require the library community working with philanthropists, booksellers, and publishers to unleash the full value of our existing and future collections by offering them digitally.

For the books we cannot buy in electronic form, I am proposing a collaborative effort to select and digitize the most widely held and used books of the 20th and 21st centuries, and to build a robust system to circulate the resulting e-books to millions and eventually billions of people.

Mike Lesk, considered by many to be the father of digital libraries, once said that he was worried about the books of the 20th century and noted that we haven't figured out "institutional responsibility" in our digital world.[1] He believed that the materials up to the 19th century would be digitized and available and that the 21st-century materials, since they were born-digital, were going to be circulated effectively. But the 20th-century materials, he thought, would be caught in machinations of copyright law—most remaining out-of-print, and all seemingly locked up by late-20th-century laws that appeared to make digitization risky.

As we shift from the analog to the digital era, Lesk's comment about "institutional responsibility" is also apt. Today, public, university, and national library leaders are not clear how best to perform their preservation and access roles, at a time when subscribing to remote databases is increasingly common and when publishers are trying to adapt to a world

By working together, we can efficiently achieve our goal. This will require the library community working with philanthropists, booksellers, and publishers to unleash the full value of our existing and future collections by offering them digitally.

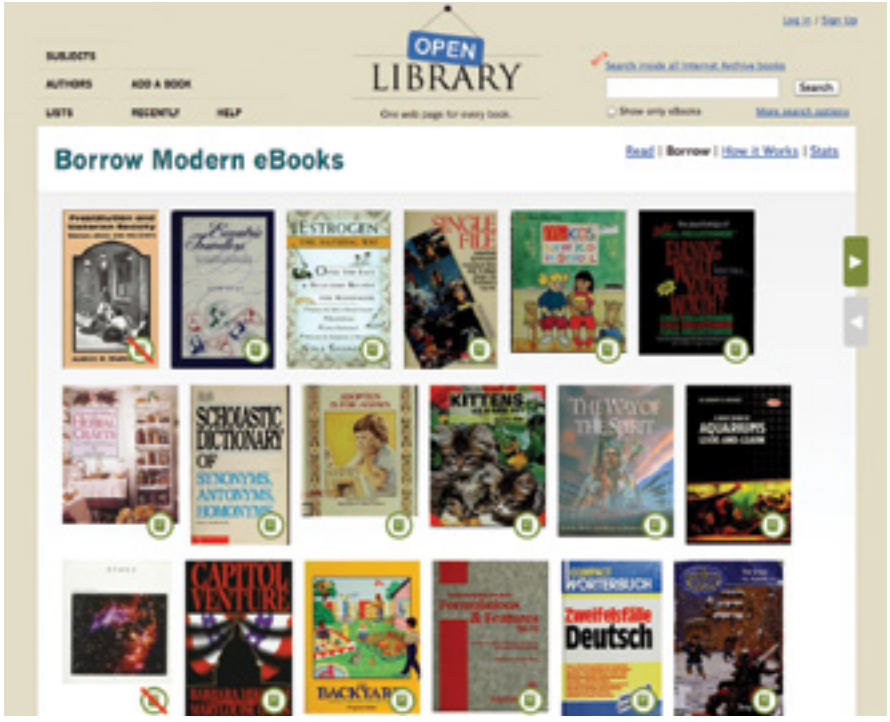
in which distribution is increasingly consolidated among a few powerhouses. If we are to have healthy publishing and library ecosystems, we need many winners and not just a few dominant players. But how do we achieve that?

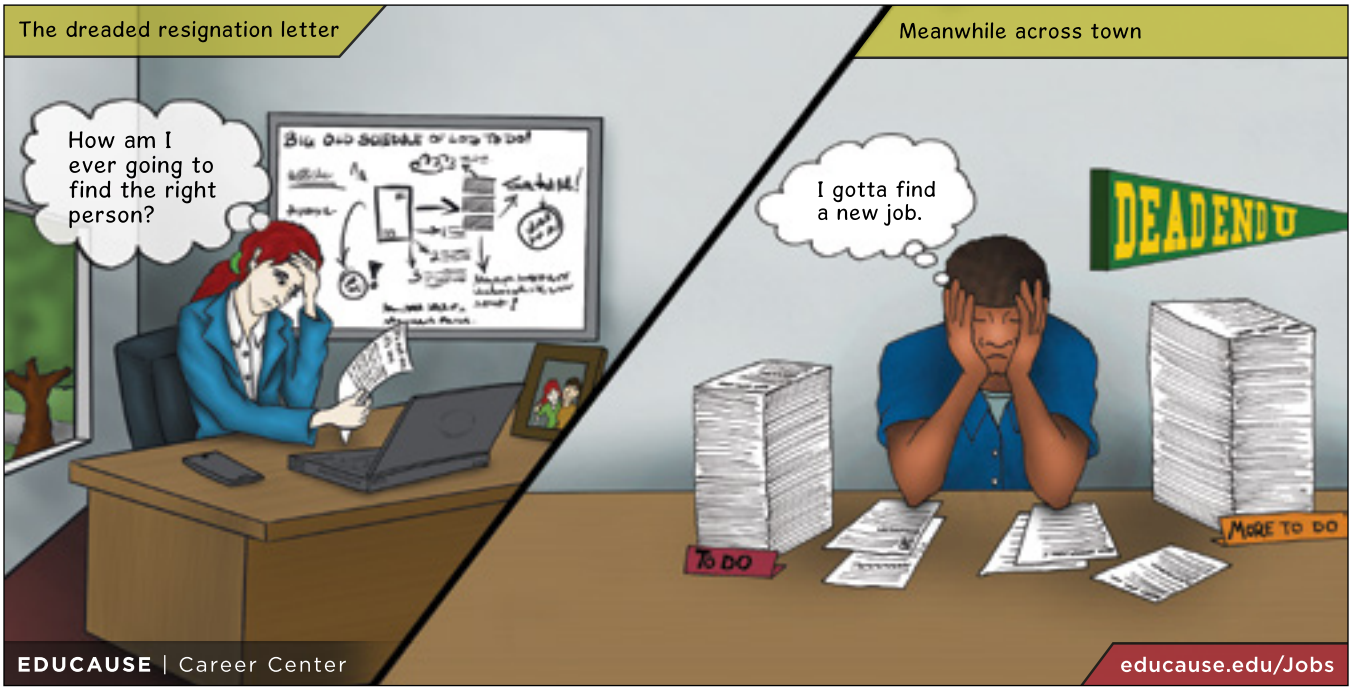
A step forward would be for libraries to buy e-books when they can, but also to transform efficiently the books currently on our physical shelves to sit on our digital shelves as well. Patrons could then easily borrow either the physical books or the electronic versions.

Open Library: Building on a Six-Year Pilot

Since 2010, the Internet Archive's Open Library has been piloting collaborative collection and lending of 20th-century books contributed by dozens of libraries (see figure 1).[2] For six years, we have been buying e-books or digitizing physical books to lend. We now lend more than 500,000 post-1923 digital volumes to one reader at a time via the Open Library website (https://openlibrary.org/borrow). This digital circulation mechanism employs the same protection technologies that publishers use for their in-print e-books distributed by commercial operations such as OverDrive (https://www.overdrive.com/) and Google Books (https://books.google.com/). Watching Open Library being used by millions over the years, we have found this approach to work. The time is ripe to go much further!

FIGURE 1. THE INTERNET ARCHIVE'S OPEN LIBRARY

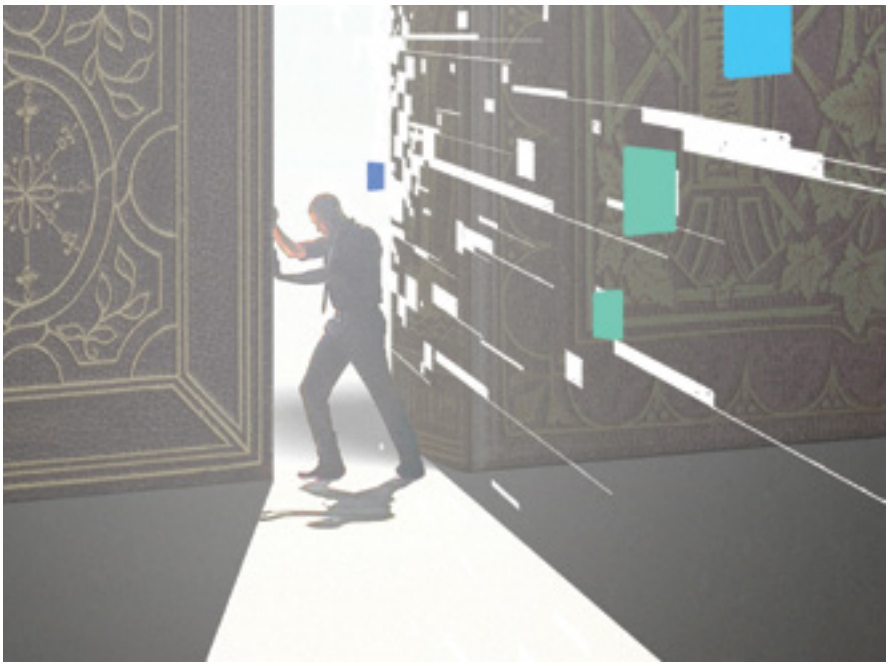




Using the Open Library approach as a foundation, we can expand to bring all interested libraries digital by 2020. By building upon the collection of 2.5 million public domain e-books that so many libraries have collaboratively digitized with the Internet Archive, we can bring the full breadth of books, both past and present, to millions of readers on portable devices, at websites, and through online library catalogs. With its extensive collections and strong public service mission, the library community can be central to this endeavor.

For instance, in each library's online card catalog, when a digital version of a book exists, we can include a web link on the record for the physical book, giving readers the ability to browse the book on screen or to borrow it from the convenience of their homes. In this way, we can smoothly enhance a library's collection, from analog to digital, at scale, by coordinating through the library catalog cloud-based vendors. We would also collectively work with publishers to purchase as many books as possible for library lending.

To build this future, we will need the participation of multiple sectors to bring thousands of libraries digital. That is one of the essential differences from the 2004 Google Book Search project, an attempt by Google and several large research libraries to bring 20th-century books online in a centralized way. That path yielded, in 2008, the Google Books settlement proposing a central controlling authority, which the courts halted in 2011 as monopolistic.[3]



A System with Many Winners

I believe this time we can pursue a *decentralized* approach, one that leads to many publishers and many libraries interacting through the market rather than having a single controlling entity. While libraries today often license e-books with restrictive terms, libraries are better served if they purchase e-books with the same rights to lend and preserve that they are entitled to when they purchase physical books today. Hopefully, going forward, all books would be available to libraries in this way—providing revenue to ensure healthy author and publisher sectors that would garner their support. But what about books that are not available in this form—including most of the existing library collections and some books published today? For these texts, libraries can work together to digitize the materials efficiently while minimizing duplication and can lend the digital texts with the same limitations placed on physical books.

In this way, patrons could read past and present books on the screens of their choice; librarians would perform their traditional roles of purchasing, organizing, presenting, and preserving the great works of humankind; publishers would sell e-books at market-based rates; and authors could choose how to distribute their works, including through publishers for payment. This may sound old-fashioned and not particularly "disruptive," but it bears the advantage that each institution plays a role structurally similar to the role it has played historically.

While libraries today often license e-books with restrictive terms, libraries are better served if they purchase e-books with the same rights to lend and preserve that they are entitled to when they purchase physical books today.

Different Eras of Books: Different Solutions

To bring our libraries digital, let's first discuss ways that groups are digitizing books at scale and then address how they can be made maximally available. The historical core of a great library, often pre-1923 books, resides in the public domain and thus does not have rights issues to hamper distribution.

Libraries with their rich special collections must still catalog and digitize their books, and we continue to work with hundreds of libraries to bring their special collections digital. But the large swath of public domain works has largely been digitized twice in the last ten years: once by the libraries working with Google and once by the libraries collaborating with the Internet Archive. Google's project has been much more thorough in its scope, scanning an estimated 25 million books thus far, but unfortunately, access to these works is limited. Institutional subscribers can gain limited access to the Google books through HathiTrust (https://www.hathitrust.org/), and the public can download some public domain books, one at a time, through the Google Books website. The Internet Archive's digitized 2.5 million older books, on the other hand, are available in bulk and for free public access. Indeed, content specialists from genealogy to biodiversity researchers actively download public domain materials from the Internet Archive, fueling innovation, dissemination, and broad public good. While we still need to complete the digitization of special collections and government documents, the pre-1923 corpus of published books is largely online and available, albeit often with restrictions.

The 20th-century books, the era that worried Lesk, are also the books librarians fret about due to rights issues. In most of the developed world, an organization can digitize books for the blind and dyslexic, and through the Marrakesh Treaty (2013), signatory countries can share these books with other signatories at scale in a way that is explicitly legal.[4] In practice, this means Canada can now digitize and lend a book from any era for the reading disabled and can share those digital copies with libraries in Australia or more than two dozen other countries. Furthermore, the U.S. court's ruling in *Authors Guild v. Google* found the basic act of mass digitization of books, even by commercial entities, to be legal under the "fair use" doctrine in the United States. So the right to digitize has been settled in many countries. A remaining legal question is what access is allowed; this proposal will allow different libraries to make their own decisions.

I believe that building a major library at the scale of the Princeton University Library, the Yale University Library, or the Boston Public Library would require institutions to offer access to a curated digital collection of 10 million books, most of which are post-1923. Collaborators can prioritize subsets of books, such as the 1.2 million books most widely held by libraries according to OCLC or the almost 1 million books that appear on one or more syllabi as determined by the Open Syllabus Project.[5] A team of collaborators could volunteer to ensure full coverage in the major subject areas while building on the core collection. But for the purposes of argument, let's stipulate that 10 million books is the number we would need to support a broadly useful public digital library system.

Collaborating to Build a Digital Collection

Building a collaborative digital collection of 10 million books will require our libraries and our partners to efficiently perform three functions:

- Coordinate collection development to avoid duplicating effort
- Offer local and cloud access
- Provide distributed preservation

In very broad strokes, to build the collections, we need curators or curatorial approaches for selecting the most useful books, then a process to determine which books we already have digitized. We need institutions or vendors able to source the missing physical books to be digitized. The participating organizations would need to have the funding to staff these functions, based either on their internal budgets or on funds raised from philanthropic sources. Maybe we could start with some already funded projects, since they might help shape the rest of the system.

Curating a Collaborative Collection

Prioritizing the books is still an open question. One approach might be to break the collection into a widely-used core of books for K-16 learners and into important topical collections. The Internet Archive could focus on obtaining and scanning the core collection of perhaps 1–2 million volumes, and then partner libraries with strong specialties could develop and scan the subject-based collections. An engineering school might take on engineering books, and a law school could focus on law books.

We must continue to work with Google Books, HathiTrust, and Amazon to explore areas of alignment. No one in the library world wants to waste precious resources by digitizing a text more than once. It would be a public benefit if these large-scale digitizers would be willing to contribute to this collaborative effort.

We will also need to research which books are emerging from copyright protection and create a comprehensive list of all digitized works. These will be important areas of research to support.

Various Levels of Access

Once we have established the core collections, each library can determine its own approach to providing access to modern works. Some might want to start by giving full access to the blind and dyslexic, as the University of Toronto is doing through the Ontario Council of University Libraries (OCUL) and the Accessible Content E-Portal (http://guides.scholarsportal.info/aceportal). Others, such as the University of California, might want to create a preservation copy. Some, such as HathiTrust, might prepare datasets for nonconsumptive researcher access. And many others, including the Internet Archive, may choose to lend their copies while keeping the physical copy on the shelf. This flexibility in access models could be one of the great strengths of this overall approach to bringing 20th-century books online—different libraries in different countries can play varying roles as their environment permits.

Libraries can take a giant step forward in the digital era by lending purchased and digitized e-books. The Internet Archive digital e-book lending program mirrors traditional library practices: one reader at a time can borrow a book, and others must wait for that one to be returned manually; alternatively, after two weeks the book is automatically returned and is offered to any waiting patrons. The technical protection mechanisms used to ensure access to only one reader at a time are the same technologies used by publishers to protect their in-print e-books. In this way, the Open Library site is respectful of rights issues and can leverage some of the learning and tools used by the publishers. The California library consortium Califa (http://califa.org/) has set up its own lending server, and it makes purchased and digitized books available through its own infrastructure to California residents. We understand the Department of Education in China also loans books it owns to one reader at a time at a major Chinese university. We all learn and benefit when different organizations in different countries test a range of approaches to access, balancing convenience and rights issues.

How would we circulate the digital e-books? Some libraries are integrating links into their library catalogs, so information about the digital versions and physical copies are side by side in the same record. Libraries can always link to the copy in the Internet Archive's Open Library, but if this is a modern book, there may be only one copy available for the whole world. Libraries can also store their own digital copies and administer their own lending system, as Califa has done. Another alternative is that the Internet Archive could create a circulation system that would administer the lending for libraries. In effect, then, each library can choose from a variety of methods to lend digital versions of the physical books in its collection. This would keep the local libraries in control but leverage the convenience of a cloud-based system that others maintain and update.

Turning on the e-book links in a catalog might be very easy now that many libraries have their catalogs on cloud services from major catalog vendors. Persuading those providers to collaborate

Each library can choose from a variety of methods to lend digital versions of the physical books in its collection. This would keep the local libraries in control but leverage the convenience of a cloud-based system that others maintain and update.

If we are striving to build the modern-day Library of Alexandria, we should avoid the fate of the first Library of Alexandria. Our community should preserve multiple copies of the books that are bought and digitized.

with this community could help deliver e-books to millions of patrons with a flip of a digital switch.

Distributed Preservation

If we are striving to build the modern-day Library of Alexandria, we should avoid the fate of the first Library of Alexandria: burning. If the library had made another copy of each work and put them in India or China, we would have the complete works of Aristotle and the lost plays of Euripides. Our community should preserve multiple copies of the books that are bought and digitized. While many libraries may be content with access to the collection on a cloud-based server, we can empower and encourage a number of libraries to store local digital copies of their books.

Fortunately, digitized books are compact enough to be affordable for libraries to store. Digital books, even with high-resolution images and all the derivative formats, are often 500 megabytes in size, so 1 million books would be 500 terabytes, which is increasingly affordable.

Distributed preservation of both the purchased e-books and the digitized books can help ensure the longevity of the precious materials in our libraries.

The Internet Archive's Funding and Technology

The Internet Archive has secured new funding to develop "super scanning centers" for the mass digitization of millions of books per year, at a significant cost savings. With the first funded super scanning center in Asia that we are now certifying for production, we anticipate being able to scan books for about one-third of the normal in-library rates achieved by the Internet Archive's twenty-eight Regional Scanning Centers. Through the Asian super scanning center, the Internet Archive can offer partners a cost savings of 50–60 percent for those willing to scan large quantities of books and have them out of circulation for several months. We are now talking with a large university research library about a plan to digitize 500,000 modern books using an Internet Archive super scanning center. This project offers the library new options in collection management, allowing it to provide digital access to books that are moving to an offsite repository. Librarians may find mass digitization at reduced cost to be a powerful tool for collection management.

In the past year, the Internet Archive has developed an in-library book-scanning system that integrates duplication detection, catalog lookup, digitization,

FIGURE 2. THE INTERNET ARCHIVE'S TABLE TOP SCRIBE, A PORTABLE, LOW-COST SCANNER

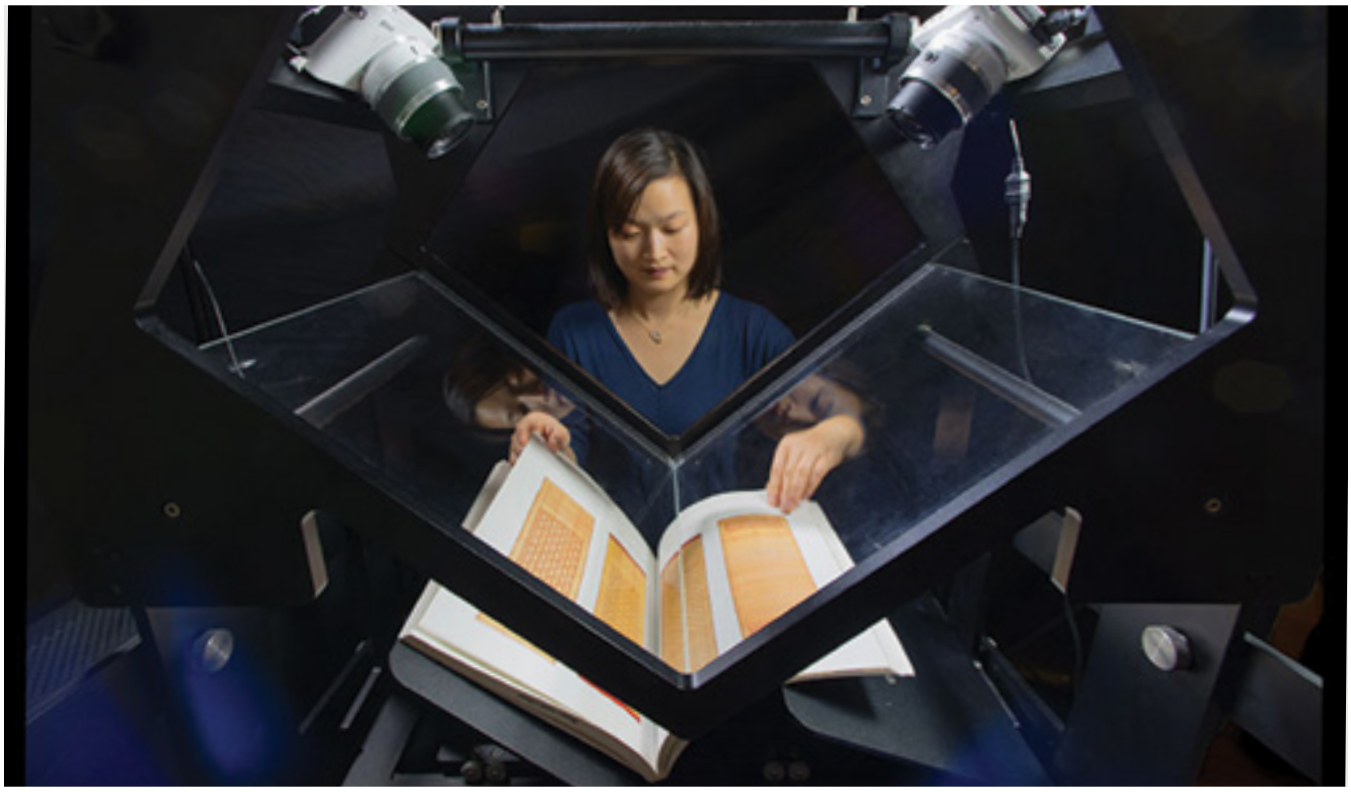


Photo Credit: David Rinehart

and integrated delivery. This can be useful for organizations that want to move through their collections, discover what has not been digitized either by themselves or by others, and digitize just these texts—while gaining access to the Internet Archive's digitized versions of all of their books, digitized from a large variety of source libraries.

Also, we now have a funding commitment to digitize millions of books and other materials that are donated to the Internet Archive. Through this initiative, the Internet Archive will seek to acquire and then digitize a core collection of books based on the recommendations of a curatorial team, while considering lists such as those compiled by OCLC and the Open Syllabus Project. This funding gives other organizations the option to donate appropriate physical books to the Internet Archive and receive a digital copy in return, at no cost to their institution.

In these ways, libraries can choose the most appropriate means of scanning their holdings. We now offer options ranging from the Table Top Scribe (see figure 2), where institutions purchase the hardware and supply their own staffing, to our regional centers in institutions such as the Boston Public Library, the University of Toronto, the Princeton Theological Seminary, and the Library of Congress. We offer lower costs for mass digitization at our Asian super scanning center and free digitization for appropriate materials donated to the Internet Archive. Our goal in offering this plethora of scanning options is to encourage all libraries to participate in the collaborative collection building in a paradigm that works for them.

Costs of Digitization

At the Internet Archive, the cost of digitization varies between $10 and $30 per book, depending on where the scanning occurs—offshore or in a library. Additional costs include acquisition, storage, and lifetime digital file management, which may come to be the predominant cost in the future.

Current in-print books are often available in e-book form, but there are few publishers willing to allow libraries to buy e-books with similar rights to the physical books they purchase. There is hope that if we coordinate our buying power, the book publishers will embrace selling e-books to libraries, much as the music publishers have come to embrace, or were forced to embrace, the selling of MP3s to services that provide broad access.[6] When available, the purchase price for these e-books tends to be approximately the same as the cost of the physical book.

Financial Stability

So far there has been little discussion of money changing hands or of any financial model to support maintaining and growing this system. If the libraries share the burden of the digitization and share the results, there would then be an incentive for some to "freeload" and wait until other libraries digitize the books and provide the services. If we want to

counter this, those libraries that did not contribute digitization or backend services could be charged for access to digitized books. And we could charge a one-time transfer fee to libraries that want to store their own local copies. But we should think carefully about financial models and avoid incentives leading to dominant systems that will limit innovation.

Conclusion

Each of our organizations has a role to play in building this collaborative digital library collection and circulation system. The Internet Archive is ready to contribute scanning technology, backend infrastructure, and philanthropic funding to digitize a core set of books that will serve K-16 learners. We are calling for partners who will help curate and source the best collections beyond what we can do, for vendors who will help circulate digital copies, and for leaders who are bold enough to push into new territory.

Because today's learners seek knowledge online, we must enable all library patrons to borrow e-books via their portable devices, by searching the web or by browsing online library catalogs. By working together, thousands of libraries can unlock analog collections for a new generation of learners, enabling digital access to millions of books now beyond their reach. The central goal—for future learners to have access to all books without physical constraints—could be realized for millions of people worldwide by the year 2020. ■

Notes

An earlier version of this article was published as the white paper "Transforming Our Libraries into Digital Libraries: A Digital Book for Every Physical Book in Our Libraries," Library Leaders Forum, Internet Archive, San Francisco, October 2016.

1. Mike Lesk, personal conversation with the author.
2. Geoffrey A. Fowler, "Libraries Have a Novel Idea," *Wall Street Journal*, June 29, 2010.
3. James Grimmelmann, "The Orphan Wars," *EDUCAUSE Review* 47, no. 1 (January/February 2012).
4. "Marrakesh Treaty to Facilitate Access to Published Works," World Intellectual Property Organization (WIPO), accessed February 4, 2017.
5. "OCLC Provides Downloadable Linked Data File for the 1 Million Most Widely Held Works in WorldCat," OCLC, news release, August 14, 2012; Open Syllabus lists 933,635 texts as of February 2017: http://explorer.opensyllabusproject.org/.
6. Steve Jobs, "Thoughts on Music," Apple (via Internet Archive Wayback Machine), February 6, 2007.

Brewster Kahle (brewster@archive.org) is Founder and Digital Librarian of the Internet Archive (https://archive.org/).