**Towards Universal Access to All Knowledge**

Brewster Kahle, Digital Librarian
Jeff Ubois
Internet Archive

Abstract: Advances in computing and communications mean that we can cost-effectively store every book, sound recording, movie, software package, and public web page ever created, and provide access to these collections via the Internet to students and adults all over the world. By mostly using existing institutions and funding sources, we can build this as well as compensate authors within what is current worldwide library budget.

This article offers an update on the current state of progress towards that ideal, which would allow us to bequeath an accessible record of our cultural heritage to our descendants.

## Introduction

The Library of Alexandria was founded with goal of obtaining a copy of every published work in the world to become center for scholarship in 288 BC. It was based on a technical advance, or rather, a change in the medium of recording to papyrus from clay tablets.

Technological advances, for the first time since the loss of the Library of Alexandria, may allow us to collect all published knowledge in a similar way. But now we can take the original goal another step further to make all the published works of humankind accessible to everyone, no matter where they are in the world.

Thomas Jefferson statement that "All that is necessary for a student is access to a library"[1] may be an exaggeration, but access to information is a key ingredient to education and an open society.

While difficult to prove, it seems safe to say that the creation and dissemination of knowledge is important for building societies that grow and prosper.

Many in the library community agree that universal access to all knowledge could stand as one of the greatest achievements of humankind, up in the pantheon of the Library of Alexandria or landing a man on the moon.

For that to happen however, we need to answer three questions:

- Can we?
- May we?
- Will we?

**Can We?**

This is essentially a question of technical and financial feasibility.

To assess that, it is necessary to estimate how much published information there is in variety of different media: text, audio, software, moving images, and the web. From this, we can estimate how difficult the digitization and storage challenges would be if we were to attempt a comprehensive collection.

Text

How many published writings have there been, and how hard would it be to put it all online?

The largest print library in the world, the Library of Congress, has about 29 million volumes.[2] A book is about a megabyte when stored as ASCII text. So the words in the Library of Congress amount to about 29 million megabytes, or 29 terabytes, which fits on a bookshelf-sized rack of computer servers that costs about $60,000. Therefore, storing books in ASCII is not a technical challenge.

Scanned images of each page with the text available to aid searching requires 100 to 1000 times more storage, but even that is still affordable by our larger libraries, or by consortia willing and able to share facilities.

Based on the experience of the scanning projects at the Library of Alexandria, in Egypt, and in the Million Books Project in India[3], we can say that scanning books costs between $5 and $20 per book to produce images of the pages at 300 dots per inch. This resolution is sufficient to allow the book to be re-printed or read on a screen. Therefore the technology of scanning and storage are affordable with current technology.

Transmitting digitized books can be done using Internet connections commonly available in schools, libraries, and Internet cafes worldwide. But for those who cannot afford these facilities, print-on-demand bookmobiles are being tested in India, the US, Egypt, and Uganda.[4] The paper, printing, and binding costs for a 100 page black-and-white book is about one dollar. Therefore access to a multi-million-book library using today's infrastructure is possible.

Organizing the collection to be usable will require ongoing help from teachers, librarians, and parents, but the new technologies make some parts of this problem easier even though the collection is much larger than those typically housed in physical libraries. Searching for words inside books can help augment the catalogs, for instance. Other techniques, such as those employed by the web search engines, which have helped millions of users navigate through billions of objects, could be employed on these collections as well.

Books could then be widely digitized and made available offering an unprecedented opportunity for students, scholars, and lifetime learners around the world.

And if we want to be really comprehensive, the How Much Information study done at UC Berkeley[5] estimated that the total production of text on paper produced worldwide in 2002 would take 1.6 petabytes to store in uncompressed form. That's about $3 million worth of storage -- for everything printed on paper in the world every year.

Music

The total of all sound recordings that have been made for the last 150 years, amount to approximately 2 to 3 million.[6]

To store digitized versions of these, at 500 megabytes each, would take 1000 to 1500 terabytes of storage, which would cost two to three million dollars with current technology.

While we are unaware of large-scale digitization efforts of these materials, commercial services now offer to digitize personal collections of CDs for $1 or less per disk, and an LP for 12 euros.[7]

Therefore, the storage and the digitization costs would be in the tens of millions of dollars. Transmitting these materials over the Internet requires broadband connections, so centralized hubs, such as schools and libraries, will be needed to serve a broad population. Luckily, over the last 10 years, there are many programs working to provide high-speed access via our schools and libraries. While these programs will take decades to complete, we can hope that much of the world's population to have these facilities within 10 years.

Moving Images

Moving images require considerably more storage and transmission bandwidth, but again, the numbers are technically achievable.

Rick Prelinger, founder of a film collection known as the Prelinger Archives, has estimated that the total number of theatrical releases of movies was between 100,000 and 200,000. At DVD quality, this equates to less than 1000 terabytes, which would cost about $2 million today. Therefore, for a year's budget of a single city library, a complete collection of all theatrical movies could be stored and accessible.

Other films, not meant for theatrical release such as government, educational, and promotional films, have been estimated at a couple of million items. While these may be difficult or impossible to obtain, they still do not pose a large technical or cost challenge. Digitizing movies and transforming it into different formats for different users is now done for $15 per hour of video and $120 per hour of film.[8]

Television also is archivable. An independent library, the Television Archive, archives 20 channels (including Russian, Japanese, western, and Arab channels) 24 hours a day with DVD

quality. Broadcasts are stored on off-line hard-drives. The recording takes around 20 terabytes of storage space a month, that is, one channel requires about 12 terabytes of storage per year. Where 20 channels are a small fraction of the total number of channels, if duplicates were eliminated from the materials recorded, then the costs could very well stay affordable with current technologies.

## Software

Based on the size of the collections at Stanford University and the Internet Archive, we estimate that there have been roughly 50,000 packaged software titles that were published since the dawn of the PC era. Again, the storage requirements for this are relatively modest -- even at 100 megabytes per title that is only 5 terabytes.

As with all the media we have discussed so far, there is an issue of accessing this data at a later time because the hardware they were designed for become obsolete. Fortunately, volunteers have made "emulators" of the common computer platforms so that programs for old hardware can be run on new computer systems by having the new computers execute the instructions for the older machines.

## The Web

The World Wide Web has quickly evolved into a publication system for tens of millions of people, many of who would not have gotten the opportunity in the printed world. This flourishing of publishing now has more websites "in print" than there are total books, both in- and out-of-print books, in the Library of Congress.

The Internet Archive has archived the web since 1996, taking a snapshot of the publicly available web pages, all the text and all the pictures, every two months. It currently takes about 20 terabytes storage space each month.

A collection of this size is affordable and even accessible with current technology.

At over 400 terabytes in size, this collection is publicly available through the Internet Archive site, where it is visited by 150,000 people a day, and it gets 8 million hits a day, making it one of the top 200 most popular websites, according to the rankings published by Alexa Internet.[9]

We believe this is one of the largest databases in existence; it is hosted on almost 1,000 Linux machines.

## And Now the Good News

The really good news is that storage on this scale is not only currently available, but affordable, and the price keeps dropping.

While the total one-time cost of digitization are in the few hundred million dollars, and ongoing preservation and access costs will be tens of millions a year, it may be useful to compare this to what we currently spend to maintain our paper libraries. The cumulative yearly budget of the libraries worldwide is $31 billion every year,[10] and a Carnegie Mellon survey put the United States spending at $12 billion a year.[11]

Second, there's another factor work, which is the cost of storage, computation, and sometimes communications has been falling with Moore's Law. Moore's Law, formulated in 1965, states that the number of transistors on a chip doubles every 18 months. Today, it implies a lot more; for purposes of this discussion, it is more useful to say that Moore's Law means that our buying power increases ten-fold every five years. That is to say, the capacity of computing systems, including disk storage, improves by a factor of ten every five years, while prices stay flat.

Therefore, the 2 million dollars that would buy a petabyte (i.e. 1000 terabytes) of storage today will buy 100 petabytes in 10 years.

Compensating authors and publishers is important to maintain a healthy information environment, and this too built into the current library system. An estimate of between one-quarter and one-third of the average library's budget now goes to compensate publishers. This would be between $3 billion and $8 billion revenue for publishers worldwide. One could expect that a similar compensation balance would work in the digital age for all the same reasons it works in the paper era.

Therefore, it is plausible that our societies spending on libraries could bring us universal access to all knowledge.

**May We?**

Will we allow ourselves to re-invent our concept of libraries to expand to use the new technologies is fundamentally a societal and policy issue. These issues are reflected in our governments spending priorities, and in law.

In many ways, the change has already happened-- students are looking to the Internet for information that a generation before required research in physical libraries. In 2001, over 70% of students, aged 12-17, used the Internet as their primary information resource on their last major project, and this figure may be rising.[12] Therefore the Internet, for better and for worse, *is* the library. The good part of this is that students and adults are using the Internet for finding information much more frequently than those in the past visiting physical libraries. The bad news is that most of our cultural works are not on the Internet yet.

There is reason for optimism in the long term for getting the bulk of published materials within reach of our children since the business opportunities and education necessities are starting to become more widely understood. New sets of companies, and many older companies, have

learned to make money in the networked environments. Some innovative universities are making teaching materials openly available to benefit their own students and others worldwide.[13]

Yet even as the technology to implement these goals has improved, the legal climate has worsened. These are not disconnected phenomena: the ability to store and share music has been used to argue for further copyright restrictions. And so far, our governmental institutions have been responsive to these private interests, while seemingly leaving aside a public that is expecting new ways to grow knowledge and culture. Legislative pressure from established media companies have expanded the realm of copyright over the last 50 years to an extent that the equivalent library services that many of us grew up with do not exist, and may never exist, in the digital era.

For instance, most works in physical libraries are out-of-print but through the expansion of copyright, most of those works are still under copyright. In the physical library, these works are available to patrons, but it appears that these books may not be loaned out over the Internet without finding rights holders and negotiating permission. In practice, this proves very expensive or impossible. With most books falling out-of-print within months of publication, and the current laws that limit digital support of these materials making them unavailable for almost 100 years. We have radically changed the role of libraries at precisely the time we could be enjoying the benefits of our digital investments. To attempt to change this in the United States, legislation has been proposed[14] and a lawsuit filed by the Internet Archive and the Prelinger Archive and litigated by the Stanford Law School Center for Internet and Society.[15]

Without some way of obtaining the necessary permissions, most of the works in our physical libraries will not be available to the over 70% of our students that use the Internet for their research.

As another example, when the U.S. copyright law was changed in 1998 with the Digital Millennium Copyright Act, it became illegal to migrate most commercial software from old floppy disks to newer storage technologies thus imperiling the creative works of thousands of software authors. The Internet Archive petitioned for an exemption, which was thankfully granted, but it cost over $20,000 in legal fees and unfortunately only lasts 3 years, when it can be taken away again**.** [16]

Expansive copyright legislation, often promoted by US media companies, is being woven into trade negotiations and lobbying efforts at World Information Property Organization (WIPO).[17] For example, the United States Federal Communications Commission has adopted one proposal that could impede current libraries video archiving efforts, called the Broadcast Flag. Now, WIPO is looking to impose this requirement in the new broadcast treaty. According the Union for the Public Domain, the broadcast treaty "would extend the power that broadcasters have to control how we use and record images and sounds, including material in the public domain."[18]

In the U.S., the broadcast flag mandate, and a recent bill, the INDUCE Act, may make it illegal under copyright law to archive television by making off-air recording impossible without circumventing copyright protection mechanisms.[19] [20] As the Christian Science Monitor noted, "The Supreme Court has ruled it is legal to tape broadcast TV shows, but new HDTV standards

will make it illegal to copy a digital broadcast without the permission of the TV station." [21]

Still, there are reasons for long-term optimism. It makes business sense and societal sense to make information products available to markets of billions of people. While incumbent companies are usually resistant to change, open governments keep markets open for new ideas to blossom and take root. After the change is recognized, even the old companies can often profit from the expanded market, but that change rarely comes from within. For instance, when video recorders first came on the market, the Motion Picture Association of America opposed them.[22] Not quite twenty years later, video rentals generated $8.4 billion in revenue for video stores,[23] much of which went back to MPAA members in the form of tape purchases.

We hope legal environment evolves to strengthen our libraries rather than hollow them out, but this is not the current direction.

To leverage the opportunities these technologies afford innovative businesses and the broad public, we will need public policies that are balanced, and that recognize public as well as private interests.

**Will We?**

We have the ingredients to do something great: the storage technology, the communications technology, and most importantly the political will to live in an open society. Indeed, universal access to all knowledge is within our grasp.

But to make it happen, and happen in Europe, we must take some deliberate steps:

First: we need to bring the idea of Universal Access to All Knowledge to the foreground with a series of meetings of stakeholders culminating in the European plan.

Second: we need to create the European implementations of the inexpensive digitization techniques in use elsewhere by encouraging existing vendors or creating new vendors that will deliver quality work for similar prices.

Third: we must build a European Archive or archives that would serve as technology partners and trusted digital repositories for the European libraries and archives involved in this transformation.

We have the beginnings of a European Archive in Amsterdam. Expanding that Archive into a center of technology and policy that would partner with existing libraries and archives could move us concretely forward. With startup funding of fewer than 20 million euros and steady funding of 10 million euros a year, such an institution could serve the European community and anchor a city as a hub in the evolving knowledge economy.

Furthermore, this Archive could work with other International Libraries on large-scale data swaps to bring remote collections to Europe and bring European collections to the world.

The ancient Greeks knew the power and value in ideas and, with the Egyptians, built the great Library of Alexandria. Today, the Universal Declaration of Human Rights, Article 19 states "Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media regardless of frontiers."[24]

Let us take advantage of our new technologies and our open societies to make a Universal Library again, and go the next step and make all knowledge easily available to every man woman and child around the world.

---

[1] Jefferson was comparing the process of apprenticeship to the value of a good library. See Thomas Jefferson to John Garland Jefferson, June 11, 1790 http://memory.loc.gov/master/mss/mtj/mtj1/012/0500/0540.jpg

[2] See the Library of Congress's own description of its collections at http://www.loc.gov/homepage/fascinate.html.

[3] The Million Books Project is described at http://www.ulib.org/html/index.html

[4] United States Internet Archive Bookmobile and development is described at http://www.archive.org/texts/bookmobile.php. The Worldbank Uganda project is at http://www.anywherebooks.org/home.php . The Indian Bookmobile is at http://mobilelibrary.cdacnoida.com/.

[5] The How Much Information estimates the total amount of information in a variety of media and formats. The section about information printed on paper is at http://www.sims.berkeley.edu/research/projects/how-much-info-2003/execsum.htm#paper.

[6] Personal conversation with Bob George, the Director of the Archive of Contemporary Music in New York City.

[7] LP digitization offer on the Internet http://www.lpacd.com/Ang_LPaCD.htm (we have no experience with this vendor, so we are not endorsing their services)

[8] Video and Film digitization service http://www.avgeeks.com/. The Internet Archive has used this vendor extensively for video digitization.

[9] Alexa Internet provides traffic rankings based on the usage trails of its toolbar users http://www.alexa.com/data/details/traffic_details?q=&url=archive.org

[10] See The OCLC's report on worldwide library spending at http://www.oclc.org/membership/escan/economic/educationlibraryspending.htm.

[11] Personal communications with Raj Reddy, Professor at Carnegie Mellon University.

[12] See http://www.pewinternet.org/PPF/r/39/report_display.asp. "The Project surveyed 754 online youth ages 12-17 and their parents. Teens and parents report that Internet is vital to completing school projects and has effectively replaced the library for a large number of online

[13] MIT's open courseware project is inspiring in this area http://ocw.mit.edu/.

[14] The Public Domain Enhancement Act http://eldred.cc/

[15] A description of this case: http://cyberlaw.stanford.edu/about/cases/kahle_v_ashcroft.shtml.

[16] The Copyright Office ruled in late October 2003 that four exemptions should be added to the anti-circumvention clause of the DMCA, to be valid until the next Copyright Office rulemaking in 2006, including two that are related to the Internet Archive's original comments:

- Computer programs protected by dongles that prevents access due to malfunction or damage and which are obsolete.
- Computer programs and video games distributed in formats that have become obsolete and which require the original media or hardware as a condition of access.

[17] United States wraps copyright into individual country trade negotiations. http://www.ala.org/PrinterTemplate.cfm?Section=intlcopyright

[18] See http://www.public-domain.org/?q=taxonomy/page/or/1 . Further, David Tannenbaum of the Union for the Public Domain noted that the treaty "stands to give broadcasters (not creators) the power to regulate copying, reproduction, distribution and right of transmission of their broadcasts. It would extend the length of these powers from 20 to 50 years, and some versions expand the powers to webcasting. The treaty would also make it illegal to circumvent technological protection measures like broadcast flags. All of this even if the broadcast is of a public domain work." See http://www.boalt.org/biplog/archive/000596.html.

[19] See http://thomas.loc.gov/cgi-bin/query/z?c108:S.2560: for the INDUCE Act.

[20] See http://www.eff.org/IP/Video/HDTV/Final_Rule_FCC-03-273A1.pdf for the relevant FCC document. See http://www.eff.org/IP/Video/HDTV/Final_Rule_FCC-03-273A1.pdf for the relevant FCC document.

[21] See http://www.csmonitor.com/2002/0412/p13s02-almo.html.

[22] In 1982, in testimony opposing the VCR before the House of Representatives, MPAA president Jack Valenti said "the VCR is to the American film producer and the American public as the Boston strangler is to the woman home alone.' The complete record of the testimony is at http://cryptome.org/hrcw-hear.htm.

[23] According to data from the Video Software Dealers Association. See http://www.vsda.org/Resource.phx/public/press/october2002/oct28-02.htx.

[24] UN Universal Declaration of Human Rights at http://www.un.org/Overview/rights.html.