The beginning of the 21st century, several trends in the evolution of libraries had emerged—collaboration was a key to survival; technology would play an integral role; library as “place” would supersede a warehouse function; and digitization would prevail.

In this article I want to explore two experiments that represent the perfect interweaving of these trends—HathiTrust (hathitrust.org) and the Digital Public Library of America (DPLA; dpl.a). These experiments in shared systems, metadata, and digitized content represent projects of a grand and grander scale. While there is no guarantee that either of these projects will be around, at least in current manifestation, it is almost certain that within 15 years their models will provide guidance for any large-scale library ventures of the future.

GOOGLE BOOK SCANNING

When Google began scanning book collections in 2004, it woke up many of us. Suddenly, we realized it was possible to scan millions of pages in a ridiculously short period of time, making the possibility of digitizing large collections a reality.

The internet and other large networks gave us the infrastructure to share such collections. The mindset of collaboration gave us the will and desire. And open source technology gave us, and continues to give us, the mechanisms for blending and integrating mammoth collections, varied formats, and conflicting metadata to create digital asset management systems unlike anything we have had before. Never has history been so open to projects like this.

We cannot expect Google, as a public corporation, to have the same values as libraries, nor should it. To assume
that a corporation is acting altruistically, and without the need for profit, is foolish. That’s not the kind of beast it is. Selling books online, selling advertising, developing digitizing processes—all these fall into the realm of what a corporation will do. Preservation and access may be peripheral collateral, but they are not the intended target. Google continues to digitize books in libraries—it has expanded to Europe and Japan and has 20-plus libraries as partners—but libraries should not look to Google to do what libraries do best.

**SHARED AND UNSHARED GOALS**

The HathiTrust and the Digital Public Library of America are two major projects inspired and assisted by Google's digitization projects. While there are critical differences, they share two common goals: first, the desire to preserve, digitally, the great and rare collections housed in libraries, museums, and other cultural institutions; and second, the desire to allow free access to these collections. This article will explore where these projects overlap and where they differ.

The stated goal of Google Books digitization project is "to create a comprehensive, searchable, virtual card catalog of all books in all languages that helps users discover new books and publishers discover new readers" (books.google.com). There are several notable and critical elements contained within this goal. The first is that of a discovery tool (searchable virtual card catalog); the second is scope (comprehensive, all books in all languages); the third is user-based (helps users discover new books); and the fourth is commercial (publishers discover new readers).

Several of these goals overlap those of the HathiTrust and DPLA (discovery tool, scope, user-based), but the commercial element is solely Google's. Libraries are not in the business of selling books or other materials. Furthermore, both HathiTrust and DPLA have goals that are unique—namely, preservation, perpetual free access, and a collection that contains nonbook, as well as text-based, material.

**HATHITRUST**

At its website, HathiTrust says its mission "is to contribute to the common good by collecting, organizing, preserving, communicating, and sharing the record of human knowledge." According to Heather Christenson, HathiTrust is a "shared secure digital repository owned and operated by a partnership of major research libraries" ("HathiTrust: A Research Library at Web Scale." Library Resources and Technical Services v. 55 n. 2: pp. 93-95). It is not a legal trust, but a trust in the sense of shared partnership and preservation. These are the paraphrased primary HathiTrust goals:

- To build a comprehensive digital archive co-owned and managed by academic (at least for now) institutions
- To create a new, open technical framework that will improve access to these materials
- To preserve these materials

Here are what I would call secondary (although this is debatable) goals:

- To share storage strategies among member libraries
- To mitigate the problem of "free-riders" (Free-riders are institutions that contribute little or nothing but take a lot.)

There is also a set of Functional Objectives, which largely have to do with technology (hathitrust.org/objectives).

**MEMBERSHIP**

HathiTrust was created in October 2008 by the 12 universities of the Committee on Institutional Cooperation and the University of California library consortium. HathiTrust is an extremely vibrant project, having grown to 60-plus libraries today. The concepts behind it are obviously very attractive to libraries. HathiTrust is a membership organization. Membership was originally dependent on the holdings/content a member library could contribute to the whole, but luckily, this model has been relaxed. Currently, membership is open to libraries that have desired content as well as those willing to pay. Thus far, HathiTrust members are academic libraries, although The New York Public Library, the Library of Congress, and The Getty Research Institute also belong. Jeremy York, project librarian for HathiTrust, told me, "We would love to see partners of all kinds join HathiTrust, and realize that institutions will see the value of membership in different ways." This seems to leave future membership fairly open.

HathiTrust is governed by a 12-member board of governors, composed of six elected members and six members appointed by the founding HathiTrust institutions. The board is responsible for appointing workgroups and committees. Current workgroups include Usability, Communications, User Support, Collections, Surrogates, Quality and Error Rate, and Full-Text Search.

**PRICING AND PRIVELEGES**

HathiTrust’s 2013 pricing formula is complicated (hathitrust.org/documents/hathitrust-cost-rationale-2013.pdf). It allows for different degrees of access and gives non-content-contributing members participation in the curation and management of the collection. The budget currently resides within the University of Michigan’s budget system, which in a sense gives the University of Michigan a solid hand on the wheel.

Although nonmembers are welcome to use the system, membership has certain privileges. Members can download full PDFs of all public domain works and those made available under Creative Commons licenses. There are restrictions for nonmembers: Some materials, typically those with third-party agreements, are unavailable. Members can create collections using a local institutional login, but nonmembers can’t. There is also an ongoing pilot program through the University of Michigan to access the full text of copyrighted...
works given certain conditions. These conditions fall under Section 107 (Fair Use) and Section 108 (Libraries, Archives & Disabilities) of the U.S. Copyright Law.

One word should be said about the quality of text. Books that were scanned by Google using OCR (optical character recognition), particularly early scans, were notorious for textual errors. HathiTrust is committed to rescanning these retroactively and improving quality.

COLLECTIONS
As of December 2012, HathiTrust contained 10,599,355 total volumes, comprising 5,574,128 individual book titles and 276,192 individual serial titles for a total of 3,709,774,250 pages. That's 475TB, in case anyone's counting. Approximately 31% of this total is in the public domain (or has a Creative Commons license) and can be viewed and/or downloaded by anyone. The remainder is subject to member privilege.

HathiTrust's most prominent subjects are Language and Literature, followed by History, Sociology, and Business & Economics. The primary language of the collection is English (48%), followed by German (9%), French (7%), Spanish (4.5%), Chinese (3.9%), Russian (3.8%), Japanese (3%), Italian (2.5%), Arabic (1.9%), and Latin (1.3%). The overwhelming majority of materials are from the 20th century, but nearly 2% of all materials are pre-1700s.

The overall collection contains many smaller collections, created by members and nonmembers (nonmembers have to obtain a University of Michigan Friend Account to set up collections; itcs.umich.edu/itcsdocs/s4316), which may contain more than a thousand titles or as few as one. The ability to create these smaller collections represents a true source of strength for HathiTrust. Smaller collections number 1,063 and are particularly strong in the Humanities and Social Sciences.

These smaller collections should be of great interest to librarians and faculty alike. Interested librarians and faculty can set up unique collections in subjects as diverse as Human Sexuality, Folklore, Papyrology, and Apiology. These collections can be used for course reserves or to provide a unique research library. Once a collection is created, the contents can be subjected to full-text searches, mining the collections for untold riches.

DIGITAL PUBLIC LIBRARY OF AMERICA
Although it arose out of the same fertile digitized soil as HathiTrust, DPLA is far different. This is because nobody really knows what the end product should or will be. Despite this, and the fact that John Palfrey, one of the primary architects, recently called it "completely ambitious and almost crazy" (acrila.org/techconnect/?p=2098), the project's leaders are still calling for a beta start in April 2013.

DPLA, which launched in 2010, is the brainchild of the Berkman Center for Internet & Society (cyber.law.harvard.edu) at Harvard University. With initial funding from the Alfred P. Sloan Foundation, the Berkman Center convened an initial group of diverse stakeholders to discuss the project, exploring its concept, scope, architecture, membership, cost, administration, and logistics. By late 2011, the group had expanded. It received additional money from the Sloan Foundation and the Arcadia Fund. On July 26, 2012, the National Endowment for the Humanities awarded the project another million.

As a concept, DPLA is far more complex than HathiTrust. First, the term "public" appears in DPLA's name (although it must be confessed that there has been controversy about keeping that word), which means that conceptually the model should include a discussion, and ultimate adoption, of what that notion means. Public libraries are many things to many people, and a large percentage of these are social. Library as space has always been a critical function of a public library, as is its role within a community. Public libraries host numerous events, discussions, and training. In addition, they are gathering spots, nodes of connectivity, and (in some cases) refuges. They are arbiters of censorship and privacy. How these attributes will be woven into the fabric of DPLA, if at all, is yet to be seen.

Secondly, while HathiTrust is an organization of academic libraries, DPLA would be far more diverse. The Library of Congress, the National Archives and Records Administration, the Smithsonian Institution, the U.S. Copyright Office, the Internet Archive, the San Francisco Public Library, Apple, and BioOne have already signed on as partners. It is projected that other partners will be academic libraries, public libraries, archives, special collections, museums, nonprofits, and other institutions integral to the nation's heritage. This list should expand over time. This is an extremely ambitious project.
MODELS

DPLA certainly did not materialize out of thin air. Europe has had a number of digital national libraries, most notably in the Scandinavian countries. Europeana, launched in 2008 and funded by the European Union with the goal of making Europe's cultural and scientific heritage accessible to the public, has been the most ambitious (europeana.eu). Europeana is based at the Koninklijke Bibliotheek (National Library of the Netherlands) and contains more than 10 million items in its collection.

With more than 180 organizations involved, and a related mission, Europeana is the real precursor for DPLA. In fact, DPLA announced that it will design its technical infrastructure to provide interoperability with Europeana. Users will have access to both systems—an aggregation of millions of books, pamphlets, newspapers, manuscripts, images, recordings, videos, and other materials. There is enormous potential for further worldwide partners and networks. In fact, in this case, the sky may not be a limit at all.

THE GUTS OF DPLA

An array of open source code, linked data, open metadata, content digitization, new content forms, and tools and services for participants/audiences is at the core of DPLA. Workgroups called “workstreams” deal with how to go about these things, as well as other nontechnical, conceptual issues.

DPLA currently has six distinct workstreams, which are groups that explore and discuss identified components of DPLA, host conferences, disseminate information, and share information with other workstreams. The Audience and Participation workstream, for example, identifies potential users and stakeholders. Some of the questions it is wrestling with have to do with user assessment, interactivity, user-generated construction of digital resources, building communities, and community-generated policy. There are also workstreams for Content and Scope, Financial/Business Models, Governance, Legal Issues, and Technical Aspects.

Following a May 2011 Beta Sprint, in September DPLA announced six proposals that were going forward (cyber.law.harvard.edu/node/7115). They include an array of search tools, architecture, coding, digitizing (notably of government documents), metadata interoperability, and social user-based tools. What is perhaps as striking as the net result of Beta Sprint is how well it worked procedurally. In fact, it is rather amazing not only how well but also how quickly the entire plan is moving along, given its short life span.

CONTENT HUBS

While I have already mentioned discrete content from Google Books, the Internet Archive, the Library of Congress, and other organizations, perhaps the major content thrust will be achieved in the creation of what DPLA is calling Digital Hubs Pilot Project, which launched in 2012 (dp.la/about/digital-hubs-pilot-project). With this project, DPLA will attempt to establish a national network of large content repositories, bringing together a diverse collection of digitized content into a single portal.

The project anticipates the formation of Service Hubs (states or regions) and Content Hubs (content providers such as HathiTrust). Service Hubs would strengthen existing infrastructures by offering existing institutions and willing players a set of standardized digitization, metadata, data aggregation, and storage services. These would be available to any organization/institution that has content to offer and would offer in return an “on-ramp for every institution in a pilot state or region to participate in the DPLA network.” Seven of these have already been identified.

Content Hubs are large existing digital libraries, such as the National Archives, that will work directly with DPLA to identify and prepare their collections for inclusion in DPLA. Existing consortiums, such as the Texas Digital Library (tdl.org), may also play a role.

Emily Gore, currently the director for content for DPLA, came with the concept of Scannebagos—mobile scan centers that would enable the rapid digitization of collections from small, local cultural heritage institutions that may not otherwise have access to these services. This would enable DPLA to capture many unique discrete collections scattered about the country. These collections would then be added to the whole.

FUTURE DIRECTIONS

One can't help being deeply excited by the prospects these directions herald for the library and information field. Never before have collaborations of such enormous scope and complexity been possible. Whether they will succeed or not can only be measured in the future. How they will be measured is also up in the air.

Some would already call both projects a success, as they have identified the possible realms for future digital libraries. But one thing is certain: The collaborative and technical structures created by both of these projects will have major ramifications on how future generations do research. And that's a very good thing.

Paul S. Piper (paul.piper@wwu.edu) is a librarian at Western Washington University.

Comments? Email the editor-in-chief (marydee@xmission.com).