HathiTrust Constitutional Convention Opening Presentation
HathiTrust's Past, Present and Future
John Wilkin, HathiTrust Executive Director

Think back to 2004 and the conversations going on in our community around digitization and the challenge of making big things happen at the intersection of our institutions. Digitization on a grand scale was 10,000 volumes, and we rejected any notion of digitizing a large corpus of materials like US federal government documents for countless reasons. In the years since our 2005 announcement that we were undertaking digitization on a large-scale, our community, in collaboration with Google and the Internet Archive, has digitized over half of the collective holdings of ARL libraries. Three years later, we launched HathiTrust, an organization that facilitates collective action on a grand scale. Seldom has so much in our world changed in such a short time. Together, we have utterly transformed parts of the library landscape.

My plan today is to talk about HathiTrust's past, present and future. Don't worry—I won't do a history of HathiTrust. My discussion of the "past" will be primarily about the organization's early accomplishments, and begins with a review of our Short- and Long-Term Functional Objects. I'll then talk briefly about a few things in the HathiTrust pipeline, and finally conclude with an overview of some of the larger changes that have taken place since 2008. A point I'd like to emphasize now and throughout is that this is a "libraries writ large" success story. What has happened is something that *we accomplished collectively*. This is not a story of an external organization—Google, a government agency, or some external champion—doing something for us. This is our story, and one that we need to understand and celebrate.

Short- and Long-Term Functional Objectives

In those early, heady days of HathiTrust, the first partners established a list of Short- and Long-Term Functional Objective. These objectives were not meant to encompass all of HathiTrust development, but were a vehicle to articulate goals for a quickly emerging organization, a way to give some initial direction until other mechanisms could create a more nuanced roadmap. We needed to define goals in order to test *responsiveness* for this new organization.

Short-term

1. Page turner mechanism
2. Branding (overall initiative; individual libraries)
3. Format validation, migration and error-checking
4. Development of APIs that will allow partner libraries to access information and integrate it into local systems individually
5. Access mechanisms for persons with disabilities
6. Public 'Discovery' Interface for HathiTrust
7. Ability to publish virtual collections

8. Mechanism for direct ingest of non-Google content

Long-term

1. Compliance with required elements in the Trustworthy Repositories Audit and Certification (TRAC) criteria and checklist
2. Robust discovery mechanisms like full-text cross-repository searching
3. Development of an open service definition to make it possible for partner libraries to develop other secure access mechanisms and discovery tools
4. Support for formats beyond books and journals
5. Development of data mining tools for HathiTrust, and use by HathiTrust of analysis tools from other sources

For every one of these functional objectives, HathiTrust has delivered something meaningful to the partnership. It's worth noting that some of these objectives were monumentally difficult and there was absolutely no certainty that we would succeed in all of them. In the end, what we accomplished was the creation of a rich, open system with a nuanced understanding of rights and the ability to deliver various forms of content to different audiences in different ways. All of the content in HathiTrust is discoverable with a superb balance of precision and recall, and the services we offer around the preservation of the content are without peer.

Although I won't cover the Functional Objectives in detail, I would like to highlight three of the more ambitious accomplishments: our TRAC certification, the full-text cross-repository searching, and the creation of a research center.

HathiTrust is only the second repository (after Portico) to receive certification by CRL. HathiTrust's process for certification involved countless hours of staff work developing processes and products, and creating and providing documentation. And that is as it should be. Certification is all about accountability and openness, and we can take pride in obtaining it. We are a distinctive type of organization, not analogous to OCLC or Portico, and our organizational distinctiveness tends to confound those who want to see a central office and central staff. It was important to document for CRL the large commitment of staffing across the partnership to help them understand that HathiTrust is not *apart* ***from*** us, but rather a *part* ***of*** *us*—that HathiTrust is not separable from our institutions. We excelled in the technical components of the review, but CRL has lingering questions about the organization. I believe we too have lingering questions about the organization. We want this effort to be part of us and not separate, and there are few models of how to make that work. This tension between something central and something that we are all a part of will, I believe, be a leitmotif in our meeting over the next three days. I think we've made good progress and that we've created a productive and healthy tension. {At this juncture, I'd like to pause to introduce Heather Christenson, the chair of the HathiTrust

Communications Working Group. We owe this group a great debt of gratitude for showcasing our successes, but this group also highlights the value of the inter-institutional work *and the tension it creates*.}

A second grand accomplishment I'll highlight is the creation of a viable full-text search mechanism that works with all of the content in the repository. I hope no one here is so jaded as to think that full-text searching across millions of volumes is a slam-dunk. Many were skeptical, and I can't tell you how many calls I fielded from vendors telling me that what we were attempting was impossible—or at least impossible without their help. The effort required a large amount of research and testing, and what we learned required deep collaboration with the broader community of developers working on the Apache Solr search engine project. The resulting service is sensitive to the amount of content—unparalleled in size—to the hundreds of languages and character sets, and to requirements like phrase searching that reflect the distinctive ways users approach a vast and diverse *library* collection. Our users can now search over 3 billion words and get results in a split second. Collective work in the partnership has produced faceted results in our full text, and ranking that takes bibliographic information in the full text into account. The functionality that we have today is tremendous, and it provides a foundation for a next generation of search that gives our users access to bibliographic information where needed, and full text where desired.

The creation of a research center is a very different kind of example and helps underline the value of collective action. Indiana University and the University of Illinois assembled the cyberinfrastructure resources to create a research center supporting uses of the HathiTrust collection. The consolidation of collections and institutional focus made HathiTrust a valuable partner for researchers at those two institutions. It was so valuable that they *redirected* institutional resources to create the infrastructure and leadership needed for this initiative—they created the research center at little or no cost to us. How much more compelling it is that the research center comes from faculty leadership (from those who would *do* the research), drawn to use of this immense library, rather than from us in support of those faculty. Indeed, because of their commitment and credibility, the research center has attracted significant funding from Sloan to deal with problems like security in use of the in-copyright materials, and I think we can expect them to be a magnet for other funding in the future. The research center will soon offer a platform for uses we could imagine but could not otherwise support. We're accomplishing the functional objective of support for research uses of the data in a number of ways, including by distributing public domain data, but the creation of a research center was a significant win for all of us and comes as a result of our working together to create a compelling library resource.

Other accomplishments

Holdings and the New Cost Model

Our accomplishments in other areas are equally impressive, and equally reflective of HathiTrust's role as a community resource. I hope that all of you are familiar with the work done by OCLC Research and Constance Malpas showing how HathiTrust's collection overlaps with those of our libraries. The first results of that work show a median ARL overlap of 19% in June 2009 and 31% in June 2010. The overlap rate was remarkably constant from big to small ARL. That is, by June 2010, nearly every ARL library could depend on finding approximately 31% of its collection online in HathiTrust. The rate of overlap continued to grow; by June 2011 *I estimate* the overlap rate to have hit a median of about 45%, and will reach something like 50% overlap early next year. Remarkably, the numbers for non-ARL institutions and particularly the Oberlin Group libraries are even greater. Materials not ingested—materials from partners like Harvard, Virginia, the CIC and Stanford, and from non-partners like Texas—could increase that number to more than 75%. The breadth of our holdings is so significant that HathiTrust is being used as one of the key resources for the just announced (Oct. 3, 2011) European serials preservation registry, The Keepers.

That any one of our libraries could find more than 50% of its collection digitized and online in HathiTrust creates real possibilities, and in this regard HathiTrust's leadership shows vision and commitment. The new cost model, which is based on overlap, is designed to share the burden of archiving in ways that are reflective of the value we derive from the collection. Our institutions share the cost of in-copyright volumes where we hold corresponding print volumes; all members of the partnership share the cost of public domain materials evenly. In order to make that cost model work, we needed a holdings database and are very close to unveiling the first examples of calculations that result from that system.

Collection overlap is an interesting phenomenon, with the various collections showing both important similarities and important differences. Focusing again on ARL institutions as the exemplar, you'll see in the scatter gram that we look remarkably similar in the rate of our overlap. However, as one might expect, the overlap profile for a collection like Harvard's and a collection like Lafayette's are so different that they will mirror each other, with Harvard holding more print corresponding to HathiTrust volumes uniquely, and Lafayette holding more volumes in common with other institutions, with a smaller number of unique volumes. These are the extremes, but all institutions will have distinctive overlap profiles. Here are just a few examples: [SLIDES]. What this means, then, is that each institution's cost will vary a great deal by size, of course, *and* by the nature of the collection. We're at a point where I can give you a preview of what that will look like.

Costs are attributed to three elements of our preservation work: the public domain; in-copyright books; and serials. Keep in mind that all partners share the cost of the public domain equally. As of the end of September, we have 2.6m public domain and 62 partners; thus, the cost of the public domain and open materials comes out to $9,300 per partner. Based on our overlap data to date, the cost for in-copyright books ranges from a low of less than $1,000 per year to a high of about $75,000 per year. I've masked the institutional names in the data here because it's still a bit early, but these numbers are largely right, and entirely based on holdings data. The high number is Michigan because Michigan's collection is the source for so much digitized content. Institutions with low costs would be institutions like Merced and Lafayette, with smaller collections and sometimes less overlap. Finally, the cost of serials is preliminarily based on holdings at the title level, rather than the volume level. Here are the same institutions arrayed along an X-axis with costs for serials on the Y-axis. The sum of these three costs gives us a low cost of less than $15,000 per year and a high cost of roughly $200,000 per year. Bear in mind that this is a likely reflection of the general shape of 2013 costs, with the bulk of the institutions paying much less than $50,000 per year. As more content comes in, costs go up; as more institutions come on board, costs go down; and as time passes, many elements of cost go down because of declining costs in the technology. So far, this has created a fairly flat picture of cost year-to-year rather than a dramatically increasing cost.

What I'd like to emphasize here is not only a concrete sense of the costs for the partnership—what they'll be and how we calculate them—but that we're well down the path to having in place the infrastructure to do this work. That is, we have a collection that represents a broad, common set of needs—not just public domain works, but in-copyright works that aid us in managing our print collections. We have technology that understands questions of holdings and overlap, which can produce cost calculations and also serve as an access control tool. Although the technology and metadata will benefit from refinement (e.g., our individual serials data could use some work), the partnership now has a good start on something that has tremendous practical value for our institutions individually and collectively.

At this juncture, and before turning to other accomplishments, I'd like to pause to consider one of the bogeymen of the new cost model: some have wondered, "what if an institution joins HathiTrust and brings with it one million public domain volumes? Won't that dramatically increase costs in uncontrollable ways?" Keep in mind the effect of scale, both of preservation costs *and* of the number of institutions. The cost for adding one million public domain volumes increases each of our costs under $4,000 per year, with a corresponding benefit of access to a phenomenal amount of content. There's nothing in the e-book marketplace that compares to this.

Publisher relations and publishing work

Never once in conceiving HathiTrust did we see this enterprise as being solely about digitized content: we believed that the digitized version of the published record provided an excellent foundation on which to add newly published materials in their original digital formats. To that end, we have set in motion three distinct efforts related to publishing:

1. Making it possible for rights holders to open access to works.
2. Making it possible for publishers to deposit digital master files for archiving and open access.
3. Making it possible for publishers to publish directly into HathiTrust.

The second and third initiatives are in their infancy, but all deserve a quick review.

In the first case, authors and publishers have opened thousands of works in an effort to share them more widely. Several presses, including university presses, and associations like ARL, have already opened substantial bodies of work with no expectation for compensation. They have relied on already extant files in the repository and have granted permissions where possible. Duke University Press recently announced an agreement with HathiTrust and Google, and will apply Creative Commons licenses to its materials, receiving in return digital files (from HathiTrust and with Google's permission).

Using born-digital materials rather than digitized versions of the books can improve the quality of HathiTrust content and the user experience. One university press is already depositing PDFs of published content. We are in discussions with two academic presses regarding an agreement where, in return for open access to their materials, we will store and provide access to the archival version.

Finally, the University of Michigan's MPublishing unit is working on a mechanism to publish open access content directly via HathiTrust. By binding together a publishing process informed by archival needs and an access mechanism informed by audience needs, they hope to build a system that makes an archival commitment to readers and libraries without losing the functionality needed for a credible publication. They hope to have the first iteration of this system available next year and to begin sharing their specifications and development process with partners following that.

Uses of in-copyright materials

We have made tremendous strides in facilitating lawful uses of in-copyright materials. Particularly in US copyright law, there are clear provisions for uses of in-copyright materials, according to the law—that is, limitations on the exclusive rights of the owners of copyright. We have legal and moral obligations to our users to provide services for these materials. And there have been important, untested questions that we need to explore as a community. I would like to briefly list work we've done to support access to in-copyright materials:

1. We have laid the groundwork for access to in-copyright works by users with print disabilities. Our technology incorporates Shibboleth for inter-institutional authentication, the holdings database as a check of a partner library's purchases, and cooperation with campus offices that provide services to users with print disabilities. We are ready to launch this service, which will provide unparalleled access to millions of works by this small group of users at our institutions. Never before have persons with print disabilities had ready access to libraries of content this large. This will be one of our proudest accomplishments.
2. Again using the holdings database and Shibboleth, we will soon be able to provide access to works that meet Section 108 criteria (i.e., that the work is damaged, deteriorating, lost or stolen and is not available on the market at a reasonable price). At the very least, we can make it possible for partners to create print replacements; it is also the case that the DMCA gives us some leeway for digital access to these works. The infrastructure is in place and we will soon use Section 108 provisions in US copyright law to extend access.
3. And, famously, we will soon be testing the concept of Fair Use and our ability to serve the imperative of preserving the materials in HathiTrust. How could I give this talk without touching on the suit by the Author's Guild against HathiTrust and several of the partners? Despite the well-documented missteps in our first orphan works identification process, our ability to make these uses under Fair Use and our ability to store the digital copies as part of an overarching preservation strategy are two of the most important principles underlying the HathiTrust effort. The access mechanisms that we have developed (e.g., taking into account holdings of the partner institution and relying on authentication of users) are thoughtful and appropriately conservative. We have taken steps to define lawful uses without antagonism of or disregard for the interests of rights holders. This was an important step for the library community.

Big issues

Creating a 10 million volume digital repository in and of itself changes the library landscape, and these things I've just discussed do as well, in that they change our sense of who we are and what we're doing: we have had a positive impact on our institutions, our users and on the profession. Additionally, there are several other developments worth considering as we look back at the last several years.

- Our institutions are now **pooling resources** in ways we rarely saw in the past. We have pooled resources to solve the digital archiving problem, to address collection building, to perform collection analysis, and I hope we will soon do so to address print monograph storage issues. We have shifted our investments from funding spent in isolation to common pools of funds to solve common problems. Before someone accuses me of being historically myopic and draws the comparison to WorldCat, keep in mind that in HathiTrust our resource pooling **replaces** (rather than enabling) local work. WorldCat makes it possible to devote resources in our separate institutions more efficiently.

- We have begun to **mobilize resources and expertise from within the various partner institutions to deal with problems common to us all**, such as copyright determination, digitization of government documents, and the refining of bibliographic information. These problems can't all be met by pooling our resources; instead, we must rely upon our individual institutional resources and perspectives. The diversity of our resources and perspectives improves the quality of our work and so makes us all stronger. (Consider the example of the copyright expertise advisory group for the new grant, which has extraordinary talent, and talent that would not be assembled in one place.) The Copyright Review Management System is a good example of early collaboration, and now IMLS has funded us again for a much more ambitious effort to work on copyright determination for publications from around the world. We have used HathiTrust to galvanize the community to address problems collaboratively. If we can find a way to deal efficiently with metadata remediation—changes and improvements to our bibliographic records—this too will surely be done by working within various institutional contexts rather than by pooling resources.
- And, finally we have begun to **approach the question of fair use in a large and coordinated way**. For some time, libraries have recognized the need for coordinated action on best practices in order to bolster our use under this part of copyright law. A few of our institutions made bold and solitary moves, and the rest of us have tried to learn from the experience. Working together on this question of fair use does, I believe, position us to develop defensible best practices and establish a clear legal precedent. In the lawsuit brought by the Author's Guild, whether or not we win remains to be seen; that we undertook this work collectively is important and a big change.

In each of these cases, we can see new modes of collective action in libraries. Where it makes sense to pool resources, we do; where it makes sense to work together on common problems, we do; and where we need to act collectively to show a unified front, we do. These are important times.

Connecting the dots

Let's pause for a moment to put all of this together.

- Together, we have built a collection of nearly unparalleled size and richness. With our future work, it will only grow larger and richer.
- We are devoting collective resources to getting a bead on what we actually have here: rights determination is the big example, but we're beginning to see interest in bibliographic remediation, *at least* for things like government documents.
- We are working to create a record of contemporary publishing *within* this corpus by working with publishers, and in some cases those publishers that are our libraries, our organizations and our university presses. We are doing that by getting permissions from authors and publishers to open access to materials, by striking deals with presses like the one we just signed with Duke University Press, and, importantly, we will soon be publishing via the repository.

We have charted a path forward for an increasingly comprehensive shared collection, a collection that contains a vast body of open materials, a collection that facilitates lawful uses, and a collection that houses new publishing. This is a collection we can *use* for many things—to gain a better understanding of the shape of the published record and our collections, to shape shared storage strategies, to rationalize our collections and to serve our users.

What next for partnership?

The short answer to the question of where the partnership goes next is that it depends entirely on the discussions we will have over the next few days. In 2008 our intention was to get the effort off the ground, and then bring the community together in 2011 to plan next steps with a clearer understanding of what we might accomplish, and that's where we are today. I hope that we leave the Constitutional Convention with a course charted for clearer, more collective governance and strategies for defining future priorities. In the meantime we—i.e., HathiTrust, our community—will continue to move HathiTrust forward. We will continue to enhance the systems you see today, providing better full text searching, supporting more functionality through the APIs, and adding more content. The holdings database and the cost model will be fleshed out, and we will all have a better sense of what our costs will be in 2013. These are important things.

I'd like to use this bully pulpit to share my personal opinion, and declare that it's time to beef up the organization. We have made a good start in creating an organization that reflects our collective interests and I feel confident that with the right governance and leadership we can create a stronger HathiTrust *without* creating a new 501c3 or intensively consolidating staff. To create a large, centralized organization would be to create a HathiTrust divorced from our institutional contexts. This is also an opportunity for me to suggest that it's time for us to look for a full-time executive director. Although I've enjoyed this work immensely and feel proud of my accomplishments, I believe that a full-time, independent director, a visionary with strong organizational skills, will make it possible for us to build a stronger sense of community and more fruitfully talk to funding agencies, both things that can make HathiTrust all that much more durable. I'm not leaving this post today; however, I would like to urge the partnership to strengthen the core of HathiTrust by building a *small* central staff and hiring a director.

Closing

In closing, I'd like to return to this theme of the community and working collectively. As we know, so many of the challenges we face are shared challenges. Our metadata are not our metadata in isolation from each other; our collections are not our individual collections in isolation from each other; and many of the baseline services or capabilities we strive to offer are ones that all of us would like to offer in our institutions. The last several years have seen us move markedly in the direction of collective action on collective problems. Indeed, working collectively on collective problems makes it all the more feasible to create distinctive or tailored services for our individual campuses or communities. Whether we call it "group scale," as Lorcan Dempsey does, or "working globally so that we can better deliver services locally," HathiTrust is a remarkable example of collective action, of our community working together to solve a common problem. Although there are many rough edges and many things to work out, our first steps have been monumentally successful in beginning to change the work we do and the way we do it. This is a tribute to each of you and to your institutions: *we* did this as a community, and we did it because it made sense. I hope we'll reconvene every few years to ponder where we've come from and where we go next, and that we will look back on this moment as a powerful example of the changes we can affect for our users and for the profession.