Chapter Eight

Bigger, Better, Together
Building the Digital Library of the Future

Jeremy York

Whether we are sitting down to write a term paper, researching our family history, gathering information about an illness, or simply looking for a good book or movie to occupy our time, we are united by common desires in seeking information: we want it here, we want it free, we want it now. In the last several decades, there have been numerous developments that have served both to speed and to frustrate the fulfillment of these desires—from the development of computers and the Internet to movements for free and open software on one hand, to the monetization of digital access, aggressive licensing, and digital rights management on the other.

The cultural heritage community has not been removed from these developments and struggles. In recent years we have seen a redevelopment and retooling of cultural heritage institutions as they seek to fulfill their traditional missions to preserve and provide access to our collected heritage and knowledge. A major development in this remaking process has been the formation of large-scale collaborative initiatives—among libraries in particular, but among archives, museums, and other cultural heritage institutions as well. Some of the initiatives seek to aggregate large amounts of materials digitized from their institutional collections either for purposes of preservation or access or for both. Europeana, HathiTrust, and the Digital Public Library of America are some of the most prominent examples, and there are others as well. This chapter examines the work of large-scale digital initiatives today and explores the directions they might take over the next several years. Three key elements will guide the discussion: the

underlying data the various initiatives seek to preserve and make accessible; the types of collaboration involved; and the services they make available to end users.

THE DATA

I remember when I was young being captivated by the "Computer" in *Star Trek: The Next Generation*—how Captain Jean-Luc Picard could ask questions with equal ease about star maps and planetary systems, or passages of Shakespeare and literature from his distant past. What would it be like, I wondered, to have the resources of Computer at my disposal—and the capabilities of the character Data to read, digest, analyze, and understand galaxies' worth of information in a matter of microseconds? "Computer, give me the ten most authoritative analyses of Zora Neale Hurston's *Their Eyes Were Watching God*." "Computer. Give me everything you have on Teddy Roosevelt and the Spanish-American War." I would have aced my English and history exams (not so sure what would have happened with algebra).

Today I wonder what it would be like if we all had this kind of access? What if any of us could enter the Starship Enterprise's holodeck and in a matter of moments be walking through computer-generated versions of nineteenth-century New York, ancient Rome, or any place and time that we had sufficient data to recreate? What would our education be like, our entertainment and jobs, our political and social systems? Information is freedom and power. Access to information is transformative.

Information is also data. And data is where the rubber hits the road.

The data of our lives was once kept in our memories, or communicated in real time through sound waves from one of us to the other. Through time it has been inscribed on cave walls, clay tablets, and plants. It has been written and printed in books, on magnetic tape, and via other media. And now it is in digits. For the vast majority of human history, data from the world around us has been something that could be processed by human senses and the human brain. We were the primary processors, translating data into meaningful information. In digital form, for most of us (realizing some of us may read in hexadecimal), the data must be preprocessed and translated from binary representations in order to be made manifest in a form we can readily use.

There are great benefits to this arrangement, including the small size of digital data in comparison with analog, the ease of transmission, and the scope and sophistication of tools (from web browsers to Adobe products to 3-D design

labs) that can be used to render the information to a broad variety of audiences. These strengths can also be weaknesses, however, as we know—particularly with regard to issues of authenticity, reliability, and preservation. The media we use for digital information are often fragile, easily subject to damage and decay, as well as malleable, subject to accidental change or intentional tampering. Securing the data, then, is a first priority and prerequisite to any kind of reliable or long-term use.

Significant progress has been made, beginning in the 1990s, on questions related to the preservation of digital materials, resulting in robust guidelines for creating and maintaining trustworthy repositories for digital data. As of this writing, four repositories in North American have been certified as "Trustworthy Digital Repositories" by the Center for Research Libraries (CRL), a body that has taken on the work of certification on behalf of its extensive community. These repositories are Portico, HathiTrust, Chronopolis, and Scholars Portal. Each of these repositories has a different focus and model of services: Portico preserves licensed e-book and e-journal content that it may make available separately from licensors in cases of designated "trigger events"; HathiTrust preserves and provides access to book and journal content digitized from partnering libraries; Chronopolis provides preservation services for a broad variety of data; Scholars Portal offers a wide range of content but has been certified in particular for its ability to manage electronic journal content. Even with their different foci and service models, these repositories share underlying missions to preserve digital content for their communities, and their formal certification contributes to their legitimacy and importance.

Of course, there are many more preservation efforts in existence than the ones that have been certified, and there are more frameworks in use to evaluate repositories than the framework used by the Center for Research Libraries. A significant number of academic, research, or governmental institutions in North America, Europe, Asia, and Australia have their own digital preservation programs, and a number of shared preservation initiatives—such as LOCKSS, CLOCKSS, MetaArchive, and the emergent Digital Preservation Network (DPN)—have gained increasing support and importance in recent years. Meanwhile, efforts such as the International Internet Preservation Consortium (comprising primarily national libraries from around the world, but also including nonprofit organizations such as the Internet Archive, Internet Memory Foundation, and National Film Board of Canada), are tackling issues around the preservation of the World Wide Web.

In the next several years, I believe we will see expanded collaboration among publishers, nonprofit organizations, and cultural heritage institutions to increase the amount of "secured" data that will flow, or at least have the possibility to flow,

into the Computer of the future. There are a number of areas of concern, however, that are not necessarily covered by current initiatives, where I believe there will be progress in the near term, but where progress will be more slow. These concerns relate to the range of the participants involved in digital preservation initiatives and the scope of the data that is preserved.

In particular, while an increasing number of publishers and major academic institutions are securing more and more of their collections, there remain a large number of colleges and universities, archives, historical societies, public libraries, museums, and individuals with collections of importance that either are not on the map to digitize, or are digitized but with no plan for preservation. Furthermore, many institutions are only beginning to address the challenges of preserving legacy audio and video collections that are in grave danger of being lost due to obsolescence or decay. Initiatives to preserve new forms of digital production—including art, multimedia publications, games, and others—are also still building capacity. Some of the major challenges preservation initiatives will face in the next couple of years are to bring more organizations, institutions, corporations, and individuals into the scope of collaborative projects, and to develop strategies to identify and retain the materials, among all of those created, that are valuable to preserve over the long term.

COLLABORATION

For all of its processing power, the Computer in *Star Trek* would certainly not have been what it was without the vast database of information behind it, and the database was surely only as good as the information compiled from all its contributors. There are two main factors driving contributions to, and collaborations in, preservation and access initiatives for cultural heritage institutions today. The first is a deep commitment to the traditional roles our institutions have had as stewards and disseminators of our collected knowledge; if we do not this, there is no one with the knowledge, skill, and enduring interest(at least, in the long run) who will. While corporations and for-profit entities are critical partners, their interests do not necessarily align with interests of the scholarly and educational communities. At the same time, academic and cultural heritage institutions are not without their own bottom lines. If libraries at universities, for instance, do not meet the needs of their students, faculty, and staff, they run the risk of losing administrative support and being marginalized as information sources for their constituencies. The same

could be said of any public library, historical society, museum, or other similar institution. The second major factor, then, is the knowledge that we will never be as comprehensive as we would like and will never be as comprehensive and financially feasible as we must be if we do not combine our efforts and collaborate effectively in areas where collaboration is a possibility. Realizing the value of our collections and services in aggregate is the Web 2.0 of digital initiatives.

HathiTrust is an exemplar of a new kind of collaboration that libraries are seeking in order to prepare themselves not just to be relevant, but to be leaders and innovators in the next century. HathiTrust is a broad collaboration of academic and research institutions that are pooling efforts to preserve and provide access to "the record of human knowledge."[1] The partnership launched a large-scale digital repository in 2008, which currently contains close to 11 million volumes digitized from their library collections. Nearly 3.5 million of these are in the public domain and available on the Web. There are a number of aspects that make collaboration in HathiTrust different from collaborations in the past. The first is the depth of collaboration. Rather than being a common software package that institutions have implemented separately to preserve and provide access to their collections, or a place where institutions are *also* putting their digitized materials to provide access to them, HathiTrust is a deep sharing of content where institutions' preservation copies are being placed in a repository that is owned and managed collaboratively by the partner institutions. It is a shared collection on a scale that we have never before seen. The partners believe that this kind of sharing is the best way to maximize the cost efficiency and impact of services offered for the deposited materials. They are not building and operating multiple systems individually; they are building one system, with adequately distributed components, together.

The second difference in HathiTrust is in the scope of the collaboration. The scope of HathiTrust's mission and goals extends to helping institutions address the long-term costs associated with the storage and management of print materials, as well as to developing new models of scholarly publishing. By sharing in the management of a collective collection of digital materials, libraries have the possibility of gaining a greater understanding of correspondences between all their collections, whether in analog form or digital. Through shared goals and shared governance, libraries have the capacity to coordinate decisions about the materials that are accessioned, retained, or deaccessioned from the shared collection and from their local collections. Libraries are able to make decisions—informed by collective holdings—about how best to allocate time and resources to most effectively meet the needs of their communities.

With regard to publishing, the base of financial and organizational stability that is achieved through broad participation in HathiTrust allows for exploration of new possibilities and new models. In particular, HathiTrust is pursuing an initiative to allow publishing of open access materials directly into the repository, thus combining processes involved in publication and archiving, and opening a new channel for libraries and publishers with common interests to work together toward common ends.

HathiTrust is a prime example of twenty-first-century library collaboration, but it is by no means the only example. Europeana, the Digital Public Library of America (DPLA), Digital Preservation Network (DPN), and other repository efforts mentioned above are premised on the value that can be gained from leveraging resources in aggregate. The DPLA seeks to aggregate and provide access to publicly available materials of all kinds through a platform that allows sharing and integration in the widest possible number of contexts and applications. The DPLA is an access layer that leverages both individual preservation infrastructures at institutions it aggregates from, and collective preservation infrastructures, such as HathiTrust, in order to provide reliable access. DPN, for its part, is a collaboration seeking to build a preservation safety net that underlies collaborative repositories in order to ensure durable access to the scholarly record. We are seeing today an increase in collaboration among libraries and other cultural heritage institutions across shared areas of interest (e.g., preservation, access, or publishing). We are also seeing, in aggregate, the emergence of a new library ecosystem. In this ecosystem, different collaborations are taking complementary roles and functions, becoming systems of collaborations, united by common desires to better serve local and collective user communities more effectively and at lower cost.

I believe that broadening collaboration, leveraging resources in aggregate, and developing collaborations with complementary roles and responsibilities are trends that will continue in the next couple of years. Some of the issues I believe are important to continued positive development in these areas are:

1. Models of governance that are adequately inclusive and representative of members, and at the same time nimble and efficient. There are particular issues in this arena regarding collaborations across governmental boundaries and legal regimes.
2. Continued alignment of the goals of collaborative enterprises with the goals of their participating members.

3. "Sourcing and scaling" of collaborative work. Sourcing and scaling are concepts that Lorcan Dempsey has written about; they involve identifying appropriate levels at which support for initiatives should be "sourced" and appropriate scales for collaboration to occur.[2] We know that collaboration is not appropriate for all activities; we know that the same level of collaboration is not effective for all undertakings. One of our greatest challenges going forward will be finding ways to source and scale initiatives in order to achieve optimum balances in efficiency, impact, and value.

SERVICES TO USERS

It is hard to imagine a service that could be better than one where it is possible to ask aloud any question you would like and immediately receive an answer. Right? "Computer, calculate the average rates of climate change over the last one hundred thousand, ten thousand, thousand, and hundred years, and model the effect of those changes over the next hundred years based on expected human social and political responses." I would say: not necessarily. For sure, we want to support full-text search across our collections where possible. For sure, we want users to be able to find our collections where relevant, no matter where they are searching. For sure, we want the most usable systems possible, overcoming all the attendant contradictions and challenges our user experience teams encounter every day. It is worth considering for a moment, however, where our current systems are taking us, and where we would like to be.

Although we have common desires in seeking information, we can have vastly differing skill sets, needs, and tolerances when it comes to the search process and using the information we find. Some of us seek information by querying resources directly through keyword searches or browsing tools. Some prefer to search via social networks or by examining citations in relevant books and articles. Sometimes we appreciate people or websites that remember our preferences or what we searched for the last time; sometimes we do not.

Below are some of the trends we are seeing today in user services (with *user* defined broadly as those who benefit from accessing the resource). These point the way to what we can expect to see more of in the future.

Making data available. One of the major trends taking place in digital repositories and digital resources in general today is a strong drive to provide access to underlying data: through linked data, APIs, data feeds—any ways that data can be made available for processing, recombination, and reuse. The availability of data facilitates its gathering and reuse by researchers. It is also a prerequisite for interoperability among administrative systems, and allows administrators to customize interfaces for their user communities, to test new functionality, and to build customized user services. APIs are key service features highlighted by HathiTrust, the DPLA, and Europeana.

Use and understand. A second major trend is a call from multiple sources for a shift from what Eric Lease Morgan has called "find and get" services, where the user puts information in a search box and selects from results, to "use and understand" services, where users are able to go beyond retrieving results to "doing" things with those results.[3] Some examples Morgan gives are the abilities to analyze, annotate, cite, compare and contrast, count and tabulate, discuss, evaluate, graph and visualize, and summarize discovered information. Morgan discusses how using text mining and natural-language processing in back-end processing can surface new information such as how long a book is (by word count rather than page), its reading difficulty, what concepts are present, the proximity of a particular word to other words, the location of a particular word within a text, and much more. Many types of analyses that have become more popular in recent years (e.g., word clouds, topic modeling, and network analyses) are based on these kinds of semantic processing. Ultimately, such processing and added functionality rely on the availability of underlying data, which enables researchers, application builders, and others to answer questions and meet their own needs, or the needs of their constituencies, through access to raw data.

Personalization and recommendation. Two other trends that are active today are an increase in personalization and in recommendation services. More and more in library and other information resources, we have the ability to customize dashboards, save personal collections, and set resource preferences. Interfaces and services respond to the devices we use, the locations we are accessing from, the information we last accessed, and our account privileges. Systems may also suggest alternate resources based on our search terms, browsing history, or resources

> others have viewed. Services are being designed not only to allow us to find information and do interesting things with it, but also to respond to our behaviors and preferences—and sometimes even our physical movements as well.

What do these trends mean for collaborations between large academic and cultural heritage institutions, and for the digital library of the future?

- There will certainly be a greater availability of data, including new data created or derived from the primary data.
- We will certainly provide more tools and opportunities for users to do interesting things with data.
- We will almost certainly be gathering and using more information about our users via the provision of services.

All of these have important implications, but the last in particular is worth considering with some care as we go forward. Part of what is driving our collaborations today, and innovations in services, is the knowledge that we, as stewards of our cultural heritage, must continue to change and adapt to the new information environment in order to be true to our missions, convictions, and values, and, at a minimum, to continue to be relevant in a culture that seeks information free, here, and now. As corporations innovate their consumer services, users approach libraries and other institutions with expectations that impact the services we offer. A single search box is a great example; streamlined interface designs and the use of analytics to determine how our websites are used and how they can be configured to best meet the needs of our users are others.

There are some issues I believe it is crucial to keep in mind as we move forward in the services we offer. The first is the particular constituencies and particular needs we are trying to meet. Education today is ever more commoditized, but it is important for us to pay attention to the actual trends and innovations happening in educational services. Some of these are:

- trends in approaches to learning that are more entrepreneurial, playful, interdisciplinary, and collaborative
- trends toward "active classrooms" and greater involvement of students in designing the learning experience
- trends toward increased peer critique

We should not only concern ourselves with offering improved personalization services, but with how our services can support greater collaboration, greater

entrepreneurship, greater empowerment of students in creating, both individually and collectively, their own learning experiences. Because of our deep experience and ties to the educational and scholarly communities, these areas may be ones in which we as libraries and cultural heritage institutions are well placed to innovate.

Another crucial issue is the traditional value that cultural heritage institutions place on maintaining and protecting the privacy of our patrons. The relationship of trust that we have as a community with our users has been built over literally centuries of time. There are numerous threats to this relationship today, including from governmental forces completely removed from issues related to the enhancement of user services. As we go forward, I believe our institutions must ensure the strength of this relationship, through transparency about our data collection practices, through opt-in and opt-out strategies, through any means at our disposal. The trust that we have with our users is one of the greatest values we bring to the marketplace of information, and is another area in which, in the twenty-first century, we are well placed to be the innovators.

PREPARING FOR SUCCESS

I'm not sure if anyone has ever said that reputation is the enemy of collaboration, but the relationship between these elements is something that cultural heritage institutions should consider carefully. For centuries our institutions have built their reputations on how much data they hold. They have amassed information, and have grown in prestige from the fact that that information could be made available only to a select few to digest and impart to others. The winds of sharing are changing, however. From open-source software to open-access licenses, from open educational resources to massive open online courses, more and more prestige is being ascribed to institutions that share their resources and expertise, and collaborate with others. Over the next several years I think the boundaries of sharing and reputation will be tested. If we get it right—if our institutions and collaborations are able to source and scale appropriately so that we do in fact maximize our value and impact—it is a very real possibility that a significant portion of our data sharing services will become part of the invisible infrastructure that underlies our collective web of knowledge. If all our data is made available (as much as it can be, bounded by respect for copyright law and concerns about user privacy), the exact source of the information may become less important to researchers and scholars than the fact that the information is available at all, and

that it is trustworthy. If a researcher knows that materials are made available via HathiTrust, for instance, or provided through the DPLA, the particular institution that contributed the material may fade into the background. To the external observer, the collaborations become great, and individual institution may be accorded status because of their participation in the collaborations, rather than because they have achieved greatness on their own.

There is a fear in this for libraries and other cultural heritage institutions; a fear that a lack of acknowledgement and recognition, if it were to occur, would lead to less funding and less capability to make information available and offer important services. I believe that the more we pool our resources and learn to work together, the stronger we will be. But we must get the balances of contributed effort and recognition correct. The progress we make in the next few years toward better services and more efficient operations will depend less on what we are technically capable of and more on the perceived value of our collective activities to ourselves and to those who fund our institutions and programs. As measures of external prestige shift from holding data to sharing data, and from sharing data to packaging data and offering personalized services in the future, we must be sure within our collaborations that appropriate means of attribution and recognition continue to exist for all activities. This is the job of the governance we choose to take us forward: to help us see the value we can achieve by acting collectively, and to ensure that when others see the power of our results, they know who was responsible for getting us there.

SOME FINAL WORDS ABOUT TRUST AND THE *STAR TREK* COMPUTER OF THE FUTURE

We should always be suspicious of information. Just as we cannot always trust our senses and perceptions, we cannot always trust the data we encounter to be accurate or the algorithms and processes running across the data to be correct or functioning properly. We are living our lives somewhat in translation in the information age, often having multiple layers of technology between us and other people; between us and the answers we seek. Our Computer of the future, if it is to serve the needs of humankind—if it is to help build a more just, free, and equal society—must be built on trust. Those of us with responsibility for collecting, preserving, and sharing our collected knowledge and creativity must continue in the best traditions of our field. We must secure the data. We must work to ensure

that the data is lawfully accessible for purposes of education and governance, and as a matter of public policy. We must be flexible, entrepreneurial, and innovative. We must keep our purpose and the needs of our communities at the forefront of our minds. And we must work together.

NOTES

1. "Mission and Goals," accessed August 13, 2013, HathiTrust, www.hathitrust.org/mission_goals.
2. Lorcan Dempsey, "Sourcing and Scaling," *Lorcan Dempsey's Weblog*, February 21, 2010, http://orweblog.oclc.org/archives/002058.html.
3. Eric Lease Morgan, "Use & Understand: A DPLA beta-sprint proposal," Infomotions, September 1, 2011, http://infomotions.com/blog/2011/09/dpla/.