The Digital Public Library of America
Concept Note
March, 2012

Overview

The Digital Public Library of America (DPLA) will make the cultural and scientific heritage of humanity available, free of charge, to all. The DPLA's primary focus is on making available materials from the United States. By adhering to the fundamental principle of free and universal access to knowledge, it will promote education in the broadest sense of the term. That is, it will function as an online library for students of all ages, from grades K-12 to postdoctoral researchers and anyone seeking self-instruction; it will be a deep resource for community colleges, vocational schools, colleges, universities, and adult education programs; it will supplement the services of public libraries in every corner of the country; and it will satisfy other needs as well—the need for data related to employment, for practical information of all kinds, and for enrichment in the use of leisure.

The process of planning for a DPLA takes as its jumping off point a statement drafted in the fall of 2010 at a workshop at the Radcliffe Institute in Cambridge, MA, at which leaders from a range of institutions agreed to work together toward development of "an open, distributed network of comprehensive online resources that would draw on the nation's living heritage from libraries, universities, archives, and museums in order to educate, inform and empower everyone in the current and future generations." The list of signatories to this statement is online at: <http://cyber.law.harvard.edu/dpla/Sign_On>.

Table of Contents

1. Community.

The community involved in the DPLA is broad, deep, and growing. The community includes a self-conscious balance between public library and academic library leaders from a wide range of organizations, as well as leaders from other cultural heritage institutions, technology and publishing companies, government agencies, and many other types of institutions. The bulk of the work of the DPLA is occurring in the six workstreams: 1) content and scope; 2) finance and business models; 3) governance; 4) legal; 5) technical aspects; and 6) audience and participation. All are welcome to participate, both online and in person at workstream meetings. These workstreams will tee up recommendations to the DPLA Steering Committee, which will seek to resolve all planning issues by consensus where possible. Members of the Steering Committee are posted online at: http://cyber.law.harvard.edu/research/dpla/steering.

2. Content and scope.

The Content and Scope workstream is determining the initial approach to materials to be included in the DPLA. No firm boundaries can be set to the collections of the DPLA; but if it takes the sky as its limit, it will never get off the ground.

Despite our ambitions to include all kinds of cultural products, we are concentrating at first on the written record—books, pamphlets, periodicals, manuscripts, and digital texts—but are designing the system such that we can move quickly to other types of materials, such as images, moving images, sound recordings, and the like. We aspire to include in the DPLA materials not just from libraries but also from other cultural heritage organizations, including museums and archives. We will avoid duplicating processes and services that are better provided through other means.

The DPLA will respect copyright, and insofar as it will include works that are commercially available, it must do so only with the consent of the rightsholders. With adequate funding, it might establish a pool of money to be distributed, according to the frequency of usage, to authors and publishers of works that are in print and covered by copyright. A similar arrangement might also extend to out-of-print (commercially unavailable) works that are covered by copyright. Some European countries have pursued that possibility with the help of collective rights management organizations (CMOs), and the DPLA could learn from their experience; but it must be extremely careful about risks, feasibility, and legal problems before adopting any such strategy. Many authors are now making their work available online according to open-access programs. The DPLA could coordinate and help implement such voluntary contributions to the general store of knowledge.

In order to lay a solid foundation for its collections and to demonstrate what the project may become in time for its April 2013 launch, the DPLA will begin with works in the public domain (including but not limited to books) that have already been digitized and are accessible through the Internet Archive, HathiTrust, a broad range of government material, and possibly private-sector initiatives such as the Google Books Project. These

can be supplemented by digital collections of research libraries and amalgamated holdings such as the digitized newspapers from the fifty states, and other related materials, that are now on deposit in the Library of Congress. A new program of scanning collections should then be undertaken with the goal of including all printed material up to 1923 from all of the major research libraries.

Further material will be added incrementally to this basic foundation of public domain works. DPLA participants have a range of views as to how to proceed from here; this issue is one of the biggest that we will need to resolve in the coming year. Participants have been discussing the best way to proceed after the inclusion of public domain materials. The Legal Issues workstream has taken up this issue, as has the Content and Scope workstream. Participants differ in their opinions as to whether to pursue law reform through this project, or instead to proceed solely on the basis of the existing legal framework.

A next layer of content to be added to the DPLA could be "orphan" works—those whose rightsholders have not been located. This layer could include works published between 1923 and 1964, a period when the extent of copyright is most problematic. Next, the DPLA should attempt to provide access to the largest possible number of books that are covered by copyright but are out of print, and potentially from there on to works both in copyright and in print. DPLA participants are considering various solutions to clear a way through the legal obstacles: e-lending models, such as those pursued by the Internet Archive with its partners; a fund to compensate the rightsholders; pay-for-view arrangements; a provision to protect the interests of authors and publishers who do not want to cooperate, voluntary agreements with those who do; or legislation to protect the DPLA from litigation on the grounds that, like all public libraries, it is a non-profit enterprise dedicated to the public welfare. A moving wall could be established to bring in new materials. However contemporary its holdings may become, the DPLA will remain steadfast in its respect for intellectual property rights.

Though the purpose of the DPLA is primarily to provide access to digital materials, it may eventually provide for the future by collecting and preserving a wide variety of information in many formats. But the DPLA cannot be everything to everyone. For it to fulfill its mission, its scope must be carefully defined, and it must be erected incrementally, according to a realistic plan.

3. Technological architecture.

The Technical Aspects workstream is hard at work designing a technical architecture for the DPLA. A small, interim core group of developers is putting together a specification (which describes the development plan in prose and graphical forms) and an initial platform for the DPLA (using an agile development approach). This *platform* is, a set of services designed to be used by developers who want to create innovative applications using the assets (metadata and some content) gathered by the DPLA. The plan is for the platform to provide access to a central store of metadata about collections held by individual institutions and by nodes that aggregate collections from many institutions,

distributed all across the Web. It will, in addition, contain metadata about local library collections, both for the rich contribution that data can make to our culture, and to provide useful services back to local libraries. The only content the platform plans to host will come from the DPLA's own sample digitization projects. All other content will remain in its existing repositories; if a developer wants to give access to end users to content in a contributing institution's collection, that access will be through a URL that points into that collection.

The metadata about these collections will be harvested and maintained through an API (application programming interface). The metadata is likely to include catalog information, use information, and other information that the community of developers might find interesting. For example, the prototype platform includes a mashup of information about public libraries from IMLS and U.S. Census Data about their localities. The platform development team currently intends to "massage" this large collection of metadata to make it more useful to developers by identifying relationships among its elements, a task that is notoriously difficult and will be approached incrementally, engaging the help of the broad community, and with realistic expectations.

The Technical workstream is committed to using existing standards, and extending them when necessary, rather than creating new standards. This will lower the barrier to institutions that want to participate, and to developers who want to create applications using this metadata. It is the best way to ensure interoperability quickly and deeply.

While initial development efforts focus on the platform, we recognize that we will need to develop (or to support development of) beautiful, intuitive user interfaces before a major public launch of the system. The design must promote interoperability and also be user-friendly—that is, it should be simple enough at its front end to satisfy the needs of ordinary citizens, while the engineering at its back end links all the platforms together in such a way that the search and discovery tools operate smoothly everywhere.

An initial technical development team has developed a scope document, made public on the dp.la web site, and will continue to publish updates of both the scope document and the initial code-base incrementally and openly, up until the April 2013 DPLA launch.

4. Metadata.

The DPLA is committed to a policy of open access to metadata. Aside from its engineering requirements, the system cannot function without adequate metadata. Legacy digitized collections require enriched semantic metadata, and current scanning operations need updated tools to create such metadata. The Technical Aspects workstream has taken up the topic of metadata standards. The experience of HathiTrust suggests that this kind of collaboration is feasible; HathiTrust might well be taken as a model. But it could be necessary to design new tools in order to promote maximal compatibility. In any case, open linked data will be necessary to provide both discoverability and context.

As a practical matter, the DPLA cannot certify that the metadata about all the items in its collection are reliable and accurate. That kind of certification must be devolved upon the institutions that originally collected them. But the DPLA could certify attribution and authenticity—that is, it could devise mechanisms to inform users about the identity of the creators of the documents and to verify that its copies are unmodified replicas of the originals.

5. Access.

Whatever its administrative structure may be, the DPLA must be open to all Americans, free of charge. Its openness should extend to everyone on the globe where possible, subject to legal constraints that may arise. Thanks to the world-wide reach of modern technology, the DPLA will be a vital part of the world of knowledge, and its activities should be coordinated with those of digital libraries in other countries. Its holdings will correspond to its global dimension, because they will include many languages and many means of communication. The use of them should be unrestricted, unless exemptions from copyright requirements may exclude commercial applications. Through the DPLA's partnership with Europeana, our plan is to connect our materials and metadata through open linked data to the materials in digital collections elsewhere in the world.

6. Digitization.

DPLA participants aspire to kick off a massive digitization project in America. This digitization effort should link the efforts of libraries, government agencies, technologists, cultural institutions of all types and sizes, and local historical societies. Digitization also calls for the preparation of standardized metadata which must be coordinated in ways to ensure adherence to common standards and interoperability. The technical requirements of the scanning must be determined after careful study—no easy task, because high-quality scanning can be so expensive as to put unacceptable pressure on the finances of the whole operation, yet the quality must be adequate for the use of scholars as well as ordinary viewers and for storage and migration through various formats.

7. Storage and preservation.

The DPLA is about access at its core. However, provisions for storage and preservation must be built into the budgets for digitizing. No one has solved the problem of permanently preserving digital works, but the DPLA should work with the leading preservation efforts—HathiTrust, DuraSpace, and LOCKSS, Portico, and/or potentially others—to build out the nation's existing preservation architecture. Many research libraries understand the need to rework digital texts and to migrate them through different formats in order to preserve them from obsolescence and decay. But all of the contributors to the DPLA should adopt compatible measures. In fact, the need for migrating digital files could reinforce the argument for creating an additional, catch-all data base or perhaps a "dark" archive, in order to mitigate the risk of loss.

8. Administration and governance.

The system of decision-making and management of the DPLA, like its architecture, is being determined through a consensus-based process. The primary discussions of this sort are being carried out through the Governance workstream. Our inclination is to establish a broad-based, federated structure, rather than a traditional top-down organization. We intend to create coherence out of diversity by erecting one virtual library out of a multiplicity of collections. But it cannot hold together without adherence to common practices and common standards, and those coordinated modes of behavior cannot be sustained unless there is an adequate administration to govern the whole system. We are actively considering a range of possible organizational options, which can be found on the Governance workstream's wiki page. The most likely outcome, based on preliminary meetings of the Governance workstream, is to create a new 501(c)(3) organization to manage the administration of the DPLA. Other options include grafting the system onto a structure that already exists, like the Library of Congress, CLIR, or ALA. The organization should be set up so as to be free from political pressures. The administration and governance of the DPLA should be reviewed after the first few years of operations to ensure that the model continues to support the initiative's goals.

9. Leadership.

The DPLA will require strong leadership to thrive. The leadership model should be inclusive and should draw upon the strengths of a diverse group of leaders from a range of related fields, including the world of libraries (public as well as private), education, information technology, publishing, and the general public.. The DPLA will most likely be governed by a strong board of trustees or directors. It may include deputies from the research libraries whose holdings will be integrated into the system. As soon as practicable, the DPLA will launch a search process for an executive director to lead the DPLA's day-to-day operations. Such an executive director could join the team at any time, though she or he should have been identified in advance of the April, 2013 plenary meeting in any event.

10. Incubation.

During the planning phase, which is well underway early in 2012, the Berkman Center for Internet & Society at Harvard University is managing the secretariat that is supporting the DPLA planning efforts. The research and planning work is being recorded and developed on a public wiki, online at: http://cyber.law.harvard.edu/dpla/Main_Page. A broad community of volunteers has been formed through participation in these workstreams, which in turn are meant to prepare the way for a "big tent" project involving the general public. The work of this planning phase is guided by members of the Steering Committee. The Berkman Center is committed to serving as the "incubator" for the DPLA but will not be its eventual home. The Berkman Center plans to manage the spin-out of the DPLA as a standalone entity or project by the time of the April, 2013 plenary meeting or shortly thereafter, through an orderly hand-off process.

11. Funding.

The initial funding for the DPLA has come from two generous foundation grant awards: $2.5 million from the Sloan Foundation and $2.5 million from the Arcadia Fund. The Sloan Foundation has also made related grant awards of nearly $1 million to Berkeley, in support of the legal workstream and the (already completed) pre-planning stage efforts. Additional grants have been made to DPLA-related activities by the Mellon Foundation (to CLIR/DLF for its beta sprint work); the National Endowment for the Humanities (to OKC for a technical meeting in June, 2011); and the Soros Foundation (to OKC for a technical meeting in Amsterdam to coordinate with Europeana, in May, 2011). The IMLS has also made grants to projects related to the DPLA, such as support for a meeting of public library leaders that took place in Los Angeles in fall, 2011, which Sloan Foundation also supported.

We expect that most of the financial support for the DPLA will come from private funding sources in the United States, such as foundations. It may, at some point, become desirable for Congress to appropriate funds to support this public good. But because the DPLA will be entirely independent of the U.S. government, its funding, at least initially, should come from a coalition of foundations who agree to establish the DPLA. One of the initial workstreams, entitled Business Models, is exploring the options for financial and business models that might support the DPLA on a sustaining basis.

12. Expectations for the Formal Launch.

In April, 2013, the DPLA planning phase will give way to the DPLA's operational phase. The expectations for this launch include a detailed operational plan for the DPLA; a technical prototype for the DPLA, including sample materials and metadata; initial digitized materials and associated metadata, including a demonstration project with Europeana related to the topic of immigration; either a new organizational structure or detailed plans to establish it immediately after the launch; and identification of an initial executive director. We also aspire to have lined up financial support for a proper system launch from a range of foundations and others who agree to support organizational costs, technical development, and digitization that will be required for the post-launch phase of the DPLA.

(Document history: We posted the first draft of the DPLA concept note in March, 2011; this second revision is being posted in March 2012.)