



Berkman

The Berkman Center for Internet & Society
at Harvard University

Technical Workshop Digital Public Library of America

June 14, 2011

On June 14, 2011, the Berkman Center for Internet & Society, in conjunction with the Open Knowledge Commons and with generous support from the National Endowment for the Humanities, convened a small working group meeting at the Associate Librarian for Library Services Office at the Library of Congress to begin to make recommendations for the overall technical architecture of a Digital Public Library of America (DPLA) and to converge on a set of key technical principles upon which the DPLA will be built.

The goal of the meeting was to surface and identify concrete insights, including areas for future research; development and policy proposals; and other tangible outcomes. The workshop explored existing technologies as possible components or platforms for DPLA; mapped out key decisions that will be foundational to DPLA's technical architecture; and explored how best to integrate the results of the ongoing Beta Sprint into future technical development. This document highlights a selection of central discussion points and questions; we hope that these takeaways will serve as input into future discussions about the DPLA.

DPLA Technologies: Foundations for Growth & Sustainability

Chris Freeland (Biodiversity Heritage Library) led an initial discussion on the overall goals and architecture of the DPLA.¹ He suggested that the DPLA serve as an outreach mechanism, a way to disseminate content to users. He questioned whether the DPLA would function as an aggregator or as a repository, what types of media it might include, who might be able to contribute content, whether certain types of content would be restricted, and what additional functions—gaming as a way to crowdsource the task of cleaning metadata, for example—might be built in.

During discussion, participants explored the various directions the DPLA prototype might take. One possibility raised was a content-focused prototype that serves primarily as a search engine; another suggestion was a prototype that offers social networking and other digital services on top of a content layer. Participants generally agreed that the final prototype should consist of a combination of content, metadata, and services built on top of this content and metadata.

Participants briefly discussed the content to be included in the DPLA, a subject explored in more depth at the March 2011 working group meeting.² “Easy content” includes content in the public domain; participants also raised the idea of supporting the digitizing of local historical content or

¹ Chris Freeland, “DPLA Technologies: Foundations for Growth & Sustainability,” <http://www.slideshare.net/chrisfreeland/dpla-technologies-foundations-for-growth-sustainability>.

² Digital Public Library of America wiki, “March 1 Workshop Notes,” March 2011, http://cyber.law.harvard.edu/dpla/March_1_Workshop_Notes.

the creation of new digital content in other areas. Participants noted the need to explore digital lending of in-copyright materials in order to appeal to the broadest possible set of users.

The question of how existing libraries might connect to the DPLA was also raised. Some participants suggested a “two-pronged” approach, in which users enter the DPLA in two ways: 1) through the DPLA’s own website/mobile applications; and 2) through seamless connections via their local libraries.

Throughout the discussion, participants emphasized the need to enable users to build services and extensions on top of the DPLA’s technical architecture. Ensuring that the DPLA is conceived of and built as a generative platform will allow for as-yet-unknown participants to build features that address as-yet-unknown uses. Participants noted that the Beta Sprint, which encourages the public to submit their ideas for the DPLA, is valuable in this regard.

The Shoulders of Giants: Leveraging Existing Technologies for the DPLA

In the day’s second session, **Martin Kalfatovic** (Smithsonian Institution) led a discussion of how best to use existing tools to build the DPLA.³ Initial conversation centered on digitization, with participants generally agreeing that the DPLA should encourage a spectrum of activities, from local scanning efforts to mass digitization projects. Participants suggested a variety of approaches to local scanning efforts, including partnering with community youth organizations, historical societies, and other local groups; training them in how to scan content and enter metadata; and providing certain resources to assist with scanning efforts in exchange for permission to include the resulting content in the DPLA.

Building on the May 2011 Global Interoperability and Linked Data Workshop,⁴ participants also discussed potential approaches to metadata. Participants discussed whether the DPLA should make use of the draft four-star classification scheme for metadata developed at the June 2011 International Linked Open Data in Libraries Archives and Museums Summit, which ranks metadata rights statements “by order of openness and usefulness: the more stars the more open and easier the metadata is to use in a linked data context.”⁵ In general, participants agreed that metadata contributed to the DPLA should be open and that the DPLA should create no new restrictions on metadata.

Participants disagreed over the importance of requiring attribution for metadata. Some argued that attributing metadata to the contributing institution, and requiring that attribution to be visible in all reuses of the metadata, is an important way of providing incentives for institutions to participate in the DPLA. Others argued that requiring attribution unnecessarily encumbers machine-based

³ Martin Kalfatovic, “The Shoulders of Giants: Leveraging Existing Technologies for the DPLA,” June 14, 2011, <http://www.slideshare.net/Kalfatovic/the-shoulders-of-giants-leveraging-existing-technologies-for-the-digital-public-library-of-america>.

⁴ Digital Public Library Wiki, “Global Interoperability and Linked Data Workshop,” May 2011, http://cyber.law.harvard.edu/dpla/Global_Interoperability_and_Linked_Data_Workshop.

⁵ MacKensie Smith, “Proposed: a 4-star classification-scheme for linked open cultural metadata,” LOD-LAM, June 6, 2011, <http://lod-lam.net/summit/2011/06/06/proposed-a-4-star-classification-scheme-for-linked-open-cultural-metadata/>.

reuse of metadata. Some participants suggested a “best practices” approach, in which attribution may be the best practice except in cases of machine-based reuse. Others suggested that the DPLA store and publish attribution but not require those reusing the data to include this information.

Specific licenses for metadata were also discussed. Some participants were in favor of waiving all copyrights via a CC0 license⁶ or by simply stating that all metadata is in the public domain. Others preferred a CC-BY⁷ or OCD-BY license⁸, both of which would require attribution.

In terms of content, participants also favored a “no new restrictions” principle on both contributed content and content digitized as part of the DPLA.

The Beta Sprint: Incorporating Outcomes in the Technical Workstream

John Palfrey (Berkman Center for Internet & Society; Harvard Law School; DPLA Steering Committee Chair) introduced a discussion on the Beta Sprint, which received over 60 statements of interest before the June 15, 2011 deadline. Participants suggested a number of ways to support the sprinters as they develop their betas, including:

1. Asking sprinters to identify specific technical, design or other needs they have so that community members who possess related skills or expertise can contribute. This might also allow those with specific tools, services, or data who did not submit a statement of interest but are interested in becoming involved to participate in beta projects. This is currently being coordinated through the DPLA wiki: [http://cyber.law.harvard.edu/dpla/Beta_Sprint: Specific Needs/Ways to Contribute](http://cyber.law.harvard.edu/dpla/Beta_Sprint:_Specific_Needs/Ways_to_Contribute).
2. Sending out a call for volunteers who can help test betas, perhaps in late August.
3. Based on lessons learned through the Beta Sprint, creating documents in support of standards or best practices for digitization, metadata, and other aspects of the DPLA’s work that might be helpful as a series of guides.
4. Building out the use cases crowdsourced through <http://allourideas.org/dpla> into short descriptions; making these available to the sprinters and to the Technical Aspects workstream as a whole.

After the Beta Sprint concludes, a group of sprinters will be invited to present their betas at the October 2011 public meeting in Washington, DC. The Technical Aspects workstream will work with sprinters and a core technical development team to develop the DPLA prototype. Participants noted that much remains to be done in terms of defining how this process will work; much of this work will come after the formal launch of the Technical Workstream.

Technical Dissemination—Where and How?

Nate Hill (San Jose Public Library) emphasized the need to make the DPLA useful for both users and libraries, particularly public librarians who may not have high levels of technical capacity.⁹ He

⁶ Creative Commons, “CC0,” <http://creativecommons.org/choose/zero/>.

⁷ Creative Commons, “CC-BY 3.0,” <http://creativecommons.org/licenses/by/3.0/>.

⁸ Open Data Commons, “Open Data Commons Attribution License,” <http://www.opendatacommons.org/licenses/by/>.

stressed that the DPLA should strive to strengthen public libraries, something that designing a DPLA only for the end user would not achieve. Participants discussed how best to reach out to public libraries and their users for input, suggesting that perhaps after an initial DPLA prototype is developed various “town hall” meetings might be held to solicit feedback.

Participants discussed possible use cases and related services that would address the needs of public librarians, for example providing embed code that would enable local libraries to include DPLA widgets on their local websites or offering training to librarians who lack technical expertise. Another example was to offer local versions of the DPLA (<http://dpla.org/city>) that would allow local public librarians to curate collections based on local interests and consisting of materials drawn from the entire corpus of the DPLA. These collections could be shared among libraries as “playlists,” remixed and supplemented with local materials as desired.

Participants suggested a number of service-oriented goals for the DPLA:

1. To make it easier for small libraries to digitize and store digital content and enable users to access this content.
2. To take a service-based approach that spans collections, collection guides created by librarians, and other materials.
3. To allow users to subscribe to collections from various libraries.

Principles for the Technical Issues

David Weinberger (Harvard Library Innovation Lab) led a discussion of the technical principles the DPLA should adapt. Participants collaboratively developed a draft set of four principles:

1. Metadata

- a. All metadata contributed to or funded by the DPLA will be placed in the public domain. The consensus is that the metadata deposit should be as open as possible, though the optimal marking has not yet been decided. It’s important to keep the perspective that most metadata (as distinct from the content-data) are uncopyrightable facts, and the DPLA does not want to encourage their licensing.
- b. The DPLA will make the metadata it aggregates and develops freely available in reusable form, except where doing so would violate personal privacy.

2. Code

- a. All code funded by the DPLA will be free and open source.
- b. All code will be posted to public repositories with version control; regular releases will be issued; the DPLA will accept patches.
- c. The DPLA will try to use existing services and code where possible. Free and open source software is preferred, and the DPLA will move toward a completely free tool chain.
- d. Code and services will not accept any intellectual property (such as patents) that is not licensed royalty-free on a non-discriminatory basis to all users.
- e. As far as possible, the DPLA platform will be open and accessible for others to fork/host/replicate with no discrimination based on use or field of endeavor.

⁹ Nate Hill, “Technical (and not-so-technical) Dissemination of the DPLA,” PLA Blog, June 13, 2011, <http://plablog.org/2011/06/technical-and-not-so-technical-dissemination-of-the-dpla.html>.

- f. In order to facilitate and maximize interoperability, the DPLA platform will support open standards, including Linked Open Data.

3. Content

- a. The content that is contributed to or funded by the DPLA will be made available, including through bulk download, with no new restrictions, via a service available to libraries, museums, and archives in the United States, where use and reuse is governed only by public law.
- b. The DPLA claims no rights based on digitization.

4. Participation

- a. The DPLA will be designed as a participatory platform that facilitates the involvement of the public in all aspects of its design, development, deployment, maintenance, and support.
- b. As much of its content, data, and metadata as possible will be made available in forms that enable their re-use and extension. This includes the provision of tools, services, and bulk download capabilities.
- c. The DPLA will actively support the community of developers that want to re-use and extend its content, data, and metadata.

These principles are still a work in progress and will be further developed by the Technical Aspects workstream; community members are encouraged to contribute their thoughts via the DPLA wiki and the public listserv.

Summary and Concluding Remarks

Chris Freeland (Biodiversity Heritage Library) worked with participants to summarize the key takeaways from the workshop.¹⁰

1. The DPLA should enable, provide, and facilitate.
2. The DPLA should aggregate existing data and create new data.
3. The DPLA should focus less on the front door and more on creating multiple points of entry.
4. Code and metadata should be open; content should be as open as possible. The DPLA should create no new gatekeepers.
5. The DPLA should work to identify gaps and needs in current systems, including change management service, trackbacks, and metadata versioning.
6. The next steps in the DPLA Beta Sprint should include a call for data providers and usability testers.
7. Participants were in favor of a “Geek Squad for Public Libraries,” which would provide digitization services, shared experiences, and support.
8. The DPLA should focus on contextualization, multimedia playlists, and facilitating social interactions around objects. The DPLA needs a unified framework and disambiguation services.

¹⁰ Chris Freeland, “DPLA summary from June mtg,” June 14, 2011, <http://blog.chrisfreeland.com/2011/06/dpla-summary-from-jun-mtg.html>.

Next Steps

October 2011 Plenary

The first public meeting of the DPLA community will be held on October 21, 2011 at the National Archives in Washington, DC. The Steering Committee encourages those who are unable to attend to hold satellite viewing events at additional locations throughout the country. The goals of this meeting include:

- To clearly communicate the vision and goals of the DPLA to the public.
- To discuss the Beta Sprint and its results.
- To hold a large, broadly open meeting in DC.
- To launch a national effort to build the DPLA via public involvement in the workstreams.

Beta Sprint

The DPLA Beta Sprint received over 60 statements of interest. Sprinters have until September 1, 2011 to submit their final betas; the Steering Committee encourages them to work with one another and with the broader community throughout the development process. The DPLA Secretariat is working to build resources to facilitate collaboration; additional information is on the wiki: http://cyber.law.harvard.edu/dpla/Beta_Sprint.

A review panel appointed by the Steering Committee and composed of experts in the fields of library science, information management, and computer science will review Beta Sprint submissions in early September. Creators of the most promising betas will be invited to present their ideas to interested stakeholders and community members during the October 2011 plenary. More information is available at <http://blogs.law.harvard.edu/dpla/>.

Workstreams

The work of the DPLA over the next two years will take place largely within six, or possibly seven, workstreams:

Audience and Participation

This workstream will collaborate with each of the other workstreams to ensure that decisions are being made that will best support the current and future needs of the broadest possible user group. Specifically, it will examine models to establish and serve communities of users and stakeholders and to define the privileges and benefits the DPLA will offer to them.

Content and Scope

This workstream will make recommendations for a collection development policy for the DPLA. One primary goal is to begin to identify and articulate the criteria for including materials in a proposed DPLA. This workstream will also confront questions regarding management of and access to distributed materials.

Financial/Business Models

This workstream will make recommendations for a sustainable business plan for the DPLA.

Governance

This workstream will make recommendations for a system of decision making and management for the DPLA. The DPLA must be as broad, open, and non-partisan as possible.

Legal Issues

This workstream will make recommendations regarding how to approach and utilize the legal and copyright environment in order to support equitable knowledge distribution in a digital world.

Technical Aspects

This workstream will explore the desired architecture for the DPLA and will make recommendations regarding technology to be used for its development and to build or facilitate building the discovery environment.

(Research Uses)

This potential workstream, closely related to the Audience and Participation workstream, will make recommendations regarding how the DPLA might support various forms computational research, teaching functions, both large- and small-scale humanities projects, etc.