DPLA Content & Scope Workshop

Museum of Photographic Arts, San Diego
August 5 – 6, 2012

Introduction

The DPLA Content & Scope workstream hosted its second workshop at the Museum of Photographic Arts at Balboa Park in San Diego, CA on August 6, 2012. The goal for the workshop was to draft recommendations for the DPLA Steering Committee with regards to metadata solutions within the proposed technical framework, finalize the DPLA content provider agreements, and identify initial content providers and data hubs for the Digital Hubs Pilot Program.

August 5, 2012

The short evening meeting and presentation on August 5th began with quick introductions around the table and a brief welcome by Rachel Frick, Co-chair of the Content & Scope workstream. Frick then introduced the presenter for the evening, Patric Stillman, Innovations and Programming Officer of Media Center San Diego.

A founding officer of the Media Arts, Mr. Stillman has a long history with libraries and technology in San Francisco, Los Angeles, and San Diego. Since 2006, Stillman has engaged the greater California community as the lead trainer for digital storytelling, an active webinar instructor with Infopeople, a promoter of lifelong learning and is currently developing a new technology-based learning lab for youth at San Diego's new Central Library.

Mr. Stillman's presentation was on Digital Citizens and Virtual Communities. In his talk Stillman emphasized the importance of not only consuming content but creating or interfacing with it. He also discussed how children/youth are digital natives and we need to consider how they are exploring use and engaging with technology. Stillman concluded the lively presentation with the point, "the digital citizen is us."

August 6, 2012

The State of the DPLA

Maura Marx, Director of the DPLA Secretariat, opened the meeting with an overview of recent DPLA project developments and goals, with DPLA Director for Content Emily Gore and Technical Development Project Manager Jeffrey Licht providing details on the content infrastructure and technical platform:

- Governance: Currently working to create a standalone 501(c)(3) nonprofit organization, including the assembly of a Board of Directors (5-7) and the hiring

of an Executive Director. A five-person nominating committee has been formed to propose a slate of candidates for the inaugural Board of Directors, and the Governance Workstream will convene a workshop at the end of August to clarify a charter and bylaws.

- Content: The DPLA recently received a $1 million grant from the National Endowment for the Humanities to plan and institute a Digital Hubs Pilot Program, a project that will enable the DPLA to partner with 5-7 existing statewide or regional digital library projects (service hubs) and larger content repositories (content hubs) to define, test, and implement pilot projects and data provider agreements, thereby establishing foundational sites in the DPLA's content infrastructure. This pilot project has among its many goals to provide the content for the initial launch of DPLA and to build a model infrastructure for future content contribution. Service hubs are hubs that aggregate metadata and data from local public and research libraries, museums, archives, historical societies, and other cultural organizations; and content hubs are themselves existing large content repositories. The DPLA will harvest metadata and data from both types of these distributed hubs to create the full DPLA dataset. Service hubs will provide an on ramp for all cultural institutions in every state to participate in the DPLA. Emily Gore will be joining the DPLA team September 1 as Director for Content, to lead the Digital Hubs Pilot Program.

- Platform and Technical Infrastructure: The Technical Aspects Workstream and the Technical Development team based at the Berkman Center for Internet & Society have released an RFP for the design and development of a front-end prototype. Work also continues on the back-end content and metadata platform (*see:* Dev Portal Wiki for recent development). The plan is to build from the Audience & Participation Use Cases and from the initial provided content.

Framework and Goals

Content & Scope Co-chair Rachel Frick, set the framework and goals for the day by asking the group questions to consider throughout the workshop. She also emphasized that the initial DPLA content will be "Green Light" content – cultural heritage and other full rights data that resolves to a digital object. Questions to answer/consider:

- How do we leverage existing metadata aggregation data stores?
- What input do we need to feed iterative development in the coming years?
- How can we provide stretch goals for what we want to accomplish, setting a vision for participants not just for what we can do today, but the long-term goals of the DPLA?
- How can we use these long-term goals to help innovate what we can accomplish now?
- What are the benefits from achieving a data store for cultural heritage work?
- What are the implications for other opportunities: from use cases, for technical architecture, for partnerships?
- How can the DPLA be more than just another metadata aggregation service?

Ms. Frick continued by noting that this is an iterative process, with this workstream helping to set the blueprint toward which the content pilot will strive. The workshop's main focus is to map out what is really important and what will encourage participation. There are a number of problems this community has not yet been able to solve in a comprehensive manner. Frick pointed out that this is a chance to set sights on services and solutions outside our current network and that were not possible on a local or regional scale.

In response to Frick's introductory questions, participants brought up a number of initial thoughts related to the DPLA's content strategy and infrastructure. One participant suggested that the DPLA have curation-specific services based around the notion of community, with activities defined both in terms of a specific collection and a corresponding audience. The term "curation" in the context of this meeting was defined in the traditional way, referring to selecting and organizing items into a defined order. Another pointed out the importance of understanding and reaching out to pre-existing content providers and, by extension, not demanding that they follow overly strict metadata guidelines, for instance, in the process of incorporating them into the DPLA infrastructure. Rather, the DPLA would take content on behalf of current groups and work with them on an ongoing basis, motivating them to change their downstream services. In essence, the DPLA would make "bridge services"—services that offer an on-ramp for cultural heritage institutions to make their data available on a national scale and also allow institutions to reabsorb that transformed, or enhanced, metadata—part of its core service offerings. If an organization were not able or ready to participate, and if they were unable to utilize DPLA bridge services, the DPLA would strive to "always leave the door open" for when the time was right for the organization to participate.

Workshop participants also pointed out the need to empower non-technical institutions and individuals by potentially providing a DPLA "App Store" that would enable them to contribute, reuse, and describe/annotate content of particular interest. The opportunity to participate in such activities and projects on a national scale is by its very nature compelling for many, participants noted, and provided a level of motivation that was not there before.

Discussion then turned to the idea of "emergent collections," or unplanned collections that form naturally around content input into the DPLA. Participants noted that the DPLA ought not to let its metadata records exist as flat resources, but that they exist as nodes that directly engage community members.

Framing the Basics

Rachel Frick continued the discussion by asking participants to map the project's first steps. Frick encouraged the participants to avoid thinking in the old ways. Using the term "metadata aggregation" to define the DPLA cloud of resource brings along with it many preconceived notions and assumed limits. Instead, she offered using the term "data store." "Data store" helps us think of our collections and their associated metadata as data, and by extraction, offers up a wide range of possibilities. What data

stores and other technical considerations does the DPLA need to address so that the possible functionalities and uses described in the DPLA front-end use cases might succeed? In response to this question, participants put forth the IMLS DCC as a model for how the DPLA could present results, collections, and various interactions or reuses thereof. While the DPLA should make use of previous efforts in the field, participants said that there is a significant opportunity to do something new and different, such as processing data to offer contextualizing information, or building upon existing metadata and providing functionality beyond simple field descriptions and raw technical notes.

Based on end-user scenarios found in the DPLA front-end use cases, participants discussed having the ability to bridge the gap between primary and secondary resources as well as digital and physical collections. The personal curation of individual objects into collections was brought up, and participants felt it important that such clusters of content do not skew general search results. Rather they could possibly be presented as a complementary set of results, akin to a recommendation service. It was widely agreed that the DPLA should also find ways to share value-added services back to upstream data providers, as well as strive for maximum transparency in terms of an item's provenance.

Ms. Frick then turned the group's attention specifically to the project's data stores. How universal should the data stores be? Would this first data store be exclusively for cultural heritage data? Would we need to create different data stores for different types of content? Participants agreed that while the DPLA should offer a vision that doesn't draw lines, it should enact an implementation strategy that does, using content that is already available for the initial release. The DPLA will have many types of items, collections, and data stores, participants noted. While there are local technical questions about implementation, a long-term technical vision will help map potential dependencies and other technical considerations down the road. Participants therefore found that it would be useful to expressly state what content the DPLA is presently focused on—and therefore implicitly state which content is of lower priority for now—while also recognizing that a large number of other cultural materials are ostensibly on the docket for future inclusion. It is important to acknowledge for infrastructure design, that although the initial DPLA data store would be comprised primarily of metadata that represent digital objects, it is within the scope of the DPLA collection strategy that eventually the data store would include digital objects as well.

The discussion then turned to what users and institutions might do with the content provided via the DPLA infrastructure. How will collaboration, augmentation, and selections be distributed? Could a condition of participation be some sort of exchange, such as some kind of p2p or information exchange? While participants noted that this conversation was largely about collaboration, both in terms of sharing and remixing collections, they also mentioned how this also brought up the question of preservation vs. access. Participants ultimately decided that the issue of preservation was not an immediate priority in light of other pressing topics, but was an area of potential collaboration with other emerging national digital preservation networks, such as the National Digital Stewardship Alliance, Academic Preservation Trust, and the Digital Preservation Network (DPN).

In preparation for the September 2012 Digital Hubs Pilot Project kick-off meeting, participants suggested a general vision for the meeting be drawn up so as to address larger questions such as whether people and institutions can collaborate on curation and collection development and other emergent projects within the DPLA content structure.

Iterating Use Cases/ Design Ideas

Rachel Frick asked the participants to review the Audience & Participation Use Cases and also for volunteers to create institutional use cases that are missing to add to the personal and technical cases already developed. George Oates volunteered to work on these. The goal is to have synthesized list of use cases (we need to have one list, not three, possibly merging with Audience & Participation's list), by mid-September, in order to have time to circulate and call for comment prior to the DPLA Midwest meeting in October.

Mid-morning Group Work

After the mid-morning break, the workshop split into four groups to discuss the following questions:

- What are the value services we should ensure we offer [to institutions, to people]?
- What are the biggest medium-term wins we are working towards (different from what's out there)?

Answers by group can be found below, as well as a summary of the comments from our remote participants.

As many of the "big wins" were agreed upon by each group, we have created a summary of the big wins and value services in **Appendix I**.

Group 1
A mix of media and issues

- Should we discuss being included in DPLA as an honor?
- How do we get people to feel they are active participants? Don't have a one-dump stand with metadata providers; active relationship is important. Encourage "ask the expert" relation with the contributor.
- Ask contributing institutions: whether they want that sort of link with a collection. This can also help build emergent collections.
- Iterative process between providers and DPLA work. If we start doing geocoding, they should get this back, and can keep improving what they put in.
- Suggest you can save time and money by participating - especially in areas you can't do locally. (Ask what they can and can't do now? What they can share and what they need? Induction survey?). Search engines, NLP, mobile apps, Int'l'z'n (both interface and metadata), linked data (as a service).
- Exploit current relationships at the network level.

Group 2
Collection level work

- Concordance as a service across collections.
- Collection matchmaking "find orgs with similar stuff" based on 10 things.
- Individual matchmaking "find people doing similar research" [include reference desks?].
- Connect primary and secondary sources; help direct people back to primary repositories.
- Be both a service and a destination; require that we offer both. Services at 3rd-party places may not be as deep/rich as DPLA's site, but make it possible for them to be (and make linking between them possible).
- Joanie Utter: while DPLA might not acquire the content, DPLA could serve as a matchmaker, and help capture related data.
- Some value in findability of offline materials too. Best when collated.
- When there are current events, you want to be able to draw together from across dozens of places. DPLA could be a hub for doing this right away (say, within less than a month of an event).
- Big scope document: one that encourages stretching.
- Provider agreement: one that encourages institutions to share, alleviate concerns, inspire don't scare.
- Data processing/normalization a la Google Refine.
- "Tiered discovery": let people find many different tiers, not just 'the most robust'. The highest tier could be locally indexed within DPLA, with fully parsed/contextualized data. A second might be using a partner's API. A third might be matches to an existing collection where metadata isn't fully available, exposing the longest tail.

Group 3

- Transform any kind of metadata and return good search results.
- Incentives for multiple audience - funders, US, bring together emergent collections.
- Enable dynamic curation and collection. What gets the highest use, how do collections form and update?
- Search and discovery for defined collections and all collections. Within your own works+.
- Agnostic framework that can handle any kind of data. Born digital --> web data
- DPLA toolbox and app store. Enable content reuse!
- Exemplars of content being used as a demo to funders – partnerships and non-traditional work.
- Lay out a warm roadmap to getting their content accepted/ reused.

Group 4

- Enhancing metadata (for users)
    - Bootstrap existing data, linked data – URIs.
    - Augmenting, editing data; adding expertise. Open collections to many levels of discovery.

    - Articulate value: communicating to contributors; in the context of use-cases.
- Enhancing metadata (for institutions and hubs)
    - Tools for enrichment; crowdsourcing platforms, quality control at scale
    - Tools for collection-making.
        - Both at many levels: users, contributors, and others.
        - Both need definition - what tools are needed, exist; who is responsible?
- Baseline requirements and recommendations for data. Example: geospatial data!
- Incentives for high-value / great collection.
- Addressing concerns
    - Preserving institutional identity and provenance.
    - Implications of data sharing. Education & awareness.
    - Reporting and analytics: understanding the effects of exposure.

Remote Participants

- Coordinated, collective ingest on a statewide scale.
- Service Hubs should make metadata better so that it is available for transformation. And/or make metadata available to Google, Bing, Baidu, etc. in a schema.org or other linked data formats yet to be developed. This is out of scope for most institutions using commercial or in-house systems.
- Capture EAD and other finding aids (from the Archivists Toolkit or other systems)
- State services at a transformational level: give them tasks to do.
- Find the immediate return for reaching each level of service and work.
- Develop services that allows for integration of DPLA content into other systems (courseware, such as Blackboard, Epsilen, CourseEra, iTunesU; online services such as Google Maps, HistoryPin, Ancestry.com). Again, this is out of scope for most institutions.
- Stuff available at point of need (whether that is in Google, a courseware system, a research project via bulk import, etc).
- Provide a utility that individuals/institutions can create and curate their own collections of interests. Something similar to the exhibit function that is in Europeana but perhaps at an easier scale. This could have great facility in the K-12 arena.
- DPLA gets a seat at the table with people (search engines, W3C) developing new protocols. I think OCLC is almost there; DPLA could be a second voice.

Review of Data Provider Agreement

After lunch, the group reviewed the draft Data Exchange Agreement Summary created by Content & Scope workstream convener Robin Dale. Workshop and remote participants commented directly on the document, which can be found at the link above. One of the main comments repeated in the meeting was that it would help to have a glossary at the top of the document as guidance for these new terms.

The Data Exchange Agreement (previously referred to as the Data or Content Provider Agreement) is something the Content & Scope workstream had discussed at previous meetings. Participants agreed that the DPLA needs to have a strong agreement in order to build trust with its partners. Only Content and Service Hubs will be signing the agreement. The first phase of the DPLA will involve harvesting metadata, previews, and actual links pointing to digital objects. It was clarified by Robin Dale that while the metadata and links to digital objects will be required, previews will be optional (though desired). The DPLA is not focusing on aggregating content itself, which is part of the scoping statement.

Sarah Shreeves and Emily Gore drafted the DPLA Metadata Requirements after the first Content & Scope workstream workshop in Philadelphia in February 2012. Although not yet an official decision, it was concluded after the last Content & Scope workshop that all metadata contributed to the DPLA must be made available under a CC0 license. Within a CC0 metadata record, there can references to provenance and links back to the home institution, thereby establishing "credit" or acknowledgement. The participants at the San Diego workshop agreed the DPLA should set an example by promoting awareness and best practices for legal barriers in terms of intellectual copyright. The DPLA Metadata Requirements will become the official requirements document for those who want to be Service Hubs. The technical requirements for providing data will evolve from the metadata requirements.

The group also commented directly on the draft Data Exchange Agreement Legal document, which is a more formal version of the Data Exchange Agreement that will serve as a jumping off place for the contracts with Content and Service Hubs. It still needs to have technical elements incorporated. The DPLA Legal Issues workstream will be taking these drafts and moving them forward.

Hub Responsibilities

The draft Hub Responsibilities document is a high-level document for beginning a conversation and relationship between DPLA and the Service Hubs and will be the basis of negotiation for Content Hubs. It is not an actual legal agreement itself. It contains points and goals to talk about with those providing data and provides interested partners an idea about the responsibilities they would take on as a DPLA Service or Content Hub. The DPLA will not have a huge overhead, so part of the Service Hub responsibilities involve managing content providers (dealing with take down issues, ensuring valid links to metadata and content are correct, etc.). Workshop and remote participants commented directly on the linked Google document above.

Potential Content Hubs

The workshop participants brainstormed potential institutions and organizations to contact after the pilot phase to potentially recruit as Content Hubs. There was some conversation about large research library digital collections, like Harvard and New York Public Library. Size cannot be the only determining factor for defining a Content Hub. While the focus of the conversation was on Content Hubs, the group did discuss Service

Hubs as well. The suggestion was made that there may be a role for a Service Hub that worked with large research libraries, as modeled in the past by the DLF Aquifer project. A list of potential DPLA Service and Content hubs is open for comment and contribution on the Content and Scope Wiki.

Next Steps

Workstream members and workshop participants are encouraged to review, flesh out, and prioritize the above Service and Content Hub lists on the Content & Scope Wiki. People can also volunteer on the wiki to share their collections.

George Oates and interested volunteers will work on institutional use case scenarios to publish on the Content & Scope wiki for comment. Jeffrey Licht mentioned that the Technical Development team has been working on these types of institutional use cases. They are available here: http://dp.la/wiki/Scenarios.

Comments from the workshop participants will be incorporated into the draft Data Exchange Agreement and Hub Responsibilities documents. These will be passed on to the Legal Issues workstream for review and comment, noting that their feedback was not deemed official legal advice. The documents will have to be vetted by hired legal counsel before office use by the DPLA organization.

The Content & Scope workstream will be working closely with the Digital Hub Pilot Program moving forward.

Appendix I: Priority Points: Big Wins and Value Services

The DPLA Content & Scope Workshop held on August 6th at the Museum of Photographic Art yielded a fruitful conversation. One was on the topic of "big wins," or how we as a community could push digital library development forward, as well as solve some of the more problematic challenges that we could not resolve when we tried to aggregate content merely on a local scale. The question: "What are the big wins that we could achieve now when we look at aggregations at a national scale?"

The biggest wins:

- Discovery and creation of emergent collections and enabling dynamic collection building. Collections that can only be created virtually by combining information from a wide variety of collecting institutions, and ultimately, individuals.
    - Creating tools/apps to facilitate local and individual collection building, as envisioned by the DPLA use case of Joanie Utter.
    - This big win works both from a top down, as in DPLA creating time sensitive collections in response to current events, historical event anniversaries, or topical events. A good example is the European immigration/emigration exhibit.
    - This also works with the concept of the DPLA collection enriching the local collection. Emergent collections could be used to engage local communities in conversation about themselves and their local history and how they and their communities connect to others throughout America. This opportunity can facilitate teaching, learning, and understanding on many levels.
- Agnostic framework that can handle any type of metadata
    - There is the priority to work with what we know, and how we know how to do it (OAI-PMH/DC). This is an opportunity to go beyond that to something new. Beyond the NISO sponsored Resource Synch there is the entrepreneurial effort of UVA and a similar one at Princeton that is "atomizing" metadata records into their basic relationship counterparts, in order to "remix" the data and represent it on the fly depending on the service that "calls" on the data. Is it possible to "break" the metadata record and atomize the record's elements and store only the "relationships" represented in the record?
- Being able to provide geographical, thematic, and time (dates) points of reference/navigation through enriched metadata
    - Location, date, subject/themes, and/or event information can be used to help users navigate through a large pool of data, to provide context to individual items, and to build other collection based services. See the UIUC/ DLF Beta Sprint entry for examples of this type of navigation. This requires metadata enrichment at the point of ingestion, and a front end service that can interpret the data into a navigational interface/service.
- The potential to transform data to linked data, and in return being a datastore, that in turn can enrich other digital resources and collections (like Europeana does for Wikipedia).

- CCO for metadata is the only way we can provide remix, reuse, and/or semantic services, like a data store of linked open data.
- It is our special collections—our unique materials—that will provide the greatest impact in a national digital library.
- Tiered discovery – Not everything has to go in the big DPLA bucket, but should be presented in some sort of unifying way through the main DPLA search function, as well as data harvesting/remixing services. Again, the UIUC/DLF beta sprint offers an illustrative example of this type of discovery approach.
    - The highest tier could be locally indexed within DPLA, with fully parsed/contextualized data. Initially this is seen harvested or data pushed via ATOM to a central cultural heritage DPLA datastore, this the idea that there is the potential to have several DPLA data stores, like a scholarly communications data store (data from University/college institutional repositories) etc.
    - A second might be using a partner's API. The example used was HathiTrust.
    - A third might be matches to an existing collection where full item metadata isn't full available, only collection level, exposing the longest tail.

<u>Other critical advantages/ concepts</u>:

- Potential time and money savings by providing services most would not be able to do locally (or that would come at a major cost).
- Provide analytics on use of materials to data providers and content home institutions.
- Connecting Content Experts to Users
    - "Ask the expert" type service.
    - Connect users to others doing similar research.
    - Local collecting institutions can focus on what they do best understanding, interpreting and sharing their local collections.
    - Amateur enthusiasts can provide support, feedback and even metadata enrichment if the right services are built into the user interface.
- Metadata enrichment: add value and make improvements to metadata being contributed (crowdsourcing, quality control, geocoding, linked open data, contextualization, etc.).
- Iterative relationship: improved metadata will be sent back to data providers.
- Tools to enable content reuse, at hub level and DPLA level (DPLA toolbox/app store).

Appendix II: Reference Links

Unedited collaborative real time notes:
http://piratepad.net/dpla-aug6

Draft Data Provider Agreement Summary document:
https://docs.google.com/document/d/1d4W3HZmivDunrhSesaAebmZ-3SMxJ-eNN4rOuhNBGX0/edit

Draft Data Exchange Agreement Legal document:
https://docs.google.com/document/d/1Gtt3E1eI-fc4xEnR2JWS9MGWZjWlL-d3hZI-yaVG-Sg/edit

Draft DPLA Metadata Requirements:
https://docs.google.com/document/d/1uZNX6S4FtxW7LTRyxAvgn5yGAjJqk7T3kWRysZ6cj3E/edit

Draft Hub Responsibilities document:
https://docs.google.com/document/d/1-G941L2ZPg0sCCIUwL0B9bD95ctnIs1OIhjLSmXFv0/edit

Audience & Participation DPLA Front-end Use Cases:
http://dp.la/use-cases

Appendix III: Workshop Participants

JOHN BUTLER – University of Minnesota
PERRY COLLINS – National Endowment for the Humanities
AARON COPE – Smithsonian Institution
BRADLEY DAIGLE – University of Virginia
ROBIN DALE – Lyrasis and Content & Scope Workstream
RACHEL FRICK – DLF Program at CLIR and Content & Scope Workstream Co-chair
EMILY GORE – Florida State University and Technical Aspects Workstream
ROBERT HORTON – Institute of Museum and Library Services and Content & Scope Workstream
JACOB JETT – University of Illinois at Urbana-Champaign and IMLS/DCC project
MARTIN KALFATOVIC –Smithsonian Libraries, Biodiversity Heritage Library, and Technical Aspects Workstream Co-chair
SAM KLEIN – OLPC, Wikimedia, and Technical Aspects Workstream Co-chair
KATHERINE KOTT – Katherine Kott Consulting
BETSY KRUGER – University of Illinois at Urbana-Champaign and Content & Scope Workstream
JEFFREY LICHT – Pod Consulting and Technical Development team
MAURA MARX – Harvard's Berkman Center and DPLA Secretariat
SHEILA MCALISTER – University of Georgia, Digital Library of Georgia
SANDRA MCINTYRE – University of Utah and Mountain West
MARY MOLINARO – University of Kentucky
GEORGE OATES – Smithsonian Institution
UCHE OGBUJI – Zepheira
TERRY REESE – Oregon State University
JENN RILEY – University of North Carolina at Chapel Hill
SARAH SHREEVES – University of Illinois at Urbana-Champaign
JOHN WEISE – University of Michigan and HathiTrust
PAM WRIGHT – National Archives and Records Administration