Technology Working Group Final Report—April 2016

Tactical Strategies for the Technical Infrastructure of DPLA-OHIO

Executive Summary

The Tactical Strategy for Technical Infrastructure Working Group was formed to evaluate technical infrastructure options for a potential DPLA hub in Ohio. The Working Group consulted with various stakeholders, as well as current DPLA hubs, to gain a better understanding of what technical stack would be the most successful within different hosting scenarios. The Working Group has endeavored to provide a thorough discussion of the issues, potential technical environments, and ultimately has provided a set of recommendations that we believe will best suit the DPLA-OHIO during a 3-year DPLA pilot.

The Working Group recommends the following:

- *DPLA-OHIO should manage their own technical stack for the initial 3-year pilot, and Repox represents DPLA-OHIO's best option for a locally hosted technology stack*

- *Technical infrastructure decisions need to remain flexible and easily fungible*
  Remember, this is a 3-year pilot and the technology environment and available options will change.

- *The DPLA-OHIO program should include a standing technology working group, comprised of members representing participating communities.*

- *Metadata remediation will primarily be a local concern; remediating at the center will be minimal, and as needed during the 3-year pilot*

- *OCLC needs to be explored as a partner*
  Given the significant number of potential Ohio DPLA contributors that utilize CONTENTdm and Ohio's unique connection to OCLC, DPLA-OHIO should actively work with OCLC to identify areas of potential collaboration.

- *Evaluate additional ways to facilitate access and discovery to the primary resources about Ohio*

Regardless of technical infrastructure stack, DPLA-OHIO must provide 6 months for the technical infrastructure implementation.

Table of Contents

Background

Since July 2015, the DPLA-OHIO[1] steering committee has been working to evaluate the feasibility of hosting a DPLA statewide service hub for Ohio content. Working with a wide range of working groups, consultants, and a host of current DPLA hub organizations, the various working groups have created recommendations regarding the feasibility and potential structure of a potential hub.

As members of the Working Group, this report represents the concerted effort of the members to develop an environmental scan of potential solutions, identify potential issues, and ultimately provide a set of recommendations based on different hosting scenarios.

[1] http://www.dplaohio.org/main:about

DPLA Hub Model

When considering potential technology stacks, one of the most important parts of this discussion is around governance, or actually, around the host organization. Evaluating current DPLA hubs, a couple of things become very clear.

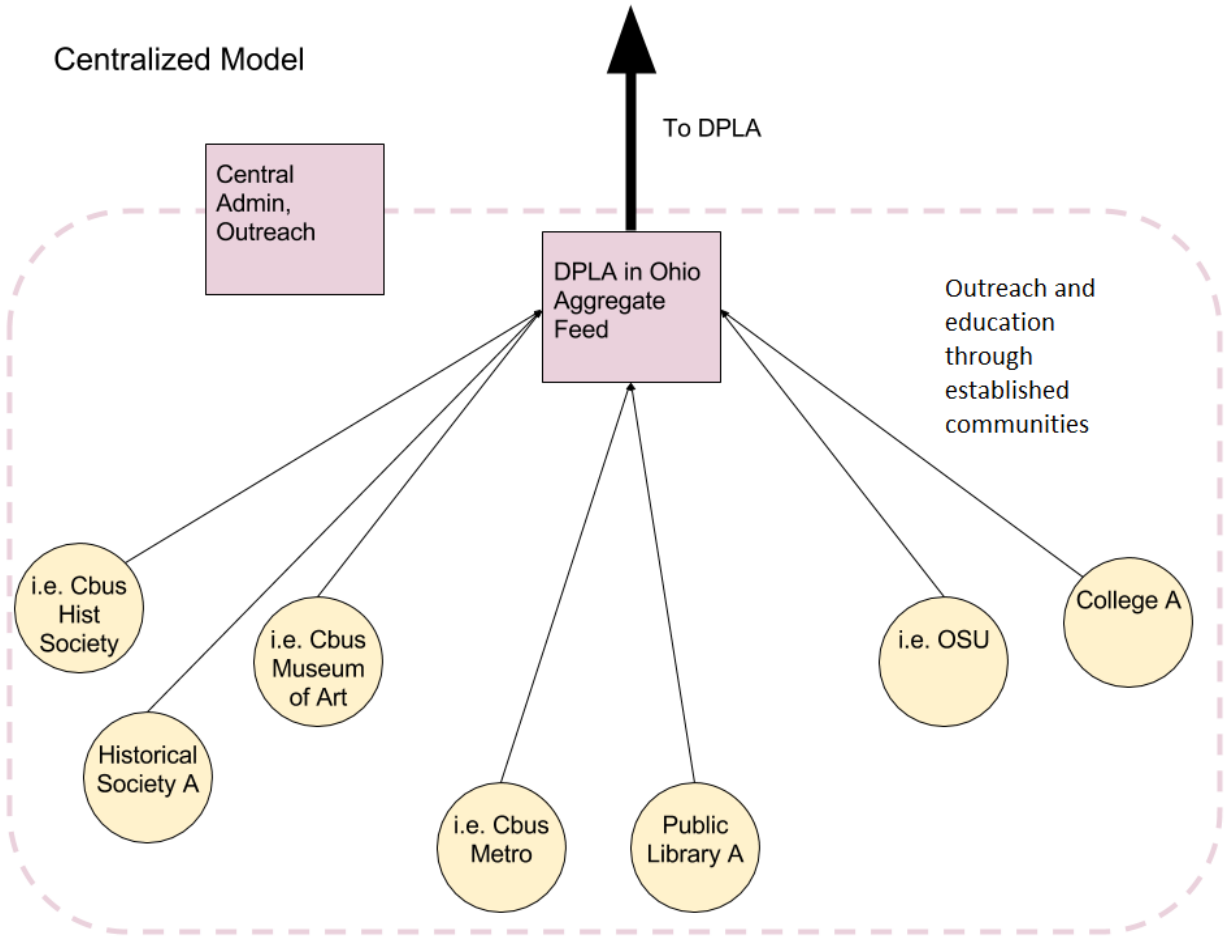
1) Early DPLA hubs organized around a strong central organization. This organization provided not only the technical infrastructure, but provided the legal entity that engaged not only with the DPLA, but the aggregation partners. This model has a number of distinct benefits, but assumes that a single organization has not only the technical but the political, and potentially financial, resources to make the hub successful.

2) More recent DPLA hubs have utilized a distributed model with mixed success. Within this model, multiple sub-hubs exist within a state, aggregating content together by geographical region or by type of organization; distributing the expertise and technology stack across a set of willing organizations. This model tends to work best for states where there already exist not only strong partnerships between organizations, but an organization within each of these subdomains that is willing and technically capable of providing both the technical and organizational support for the project.

Within the current DPLA hub structure, those hubs that utilized a more centralized technology support model tended to have had the earliest and most sustained success. While this model places more of an administrative, technical, and fiscal burden on a single host organization, it has been the fastest and most well tested avenue for success.

In considering Ohio and the current partnerships that exist within this state, a first glance may lead one to believe that Ohio might thrive within a distributed technical architecture. The state has many strong partnerships between public, academic, and the museum/archive community – with strong organizations representing these groups and working closely with their constituents. Distributing the initial harvesting aggregations and metadata work would greatly simplify the work at the center – the central aggregator that would ultimately send content to the DPLA. Within this more distributed model, the central aggregation would have nearly no responsibilities – save for providing a feed to the DPLA.

However, while a distributed approach to the architecture does have benefits, it also requires multiple institutions to step up to take on hosting responsibilities for the infrastructure. For this three-year pilot, there is not such a group of institutions in Ohio - institutions that are technically able to fill this role have competing local priorities at this time. Therefore, the technology working group believes that a centralized aggregation is much more likely to be successful within the current environment.



Within this model, the technology stack, legal agreements, and administration will be handled from a central administrative hub. As of February 2016, only the State Library has offered a budget and implementation plan related to the development and hosting of a centralized metadata aggregator.[2] This would mean that all organizations would aggregate content into a central hub, and that central hub would manage the aggregated feed of content to the DPLA.

While a single organization should host the central aggregator and technical infrastructure, organizations throughout the state can and should participate in related outreach, engagement and education efforts. The DPLA-OHIO community can leverage existing partners and networks for education and outreach, which will allow the central administrative hub to focus on the aggregation and ongoing relationship with DPLA.

Anatomy of a DPLA Hub

A quick scan of current and future hubs provides a pretty clear picture related to the varied set of technological infrastructure components that can make up a DPLA aggregation. Looking across hubs, the technical stack components includes:

[2] The proposed budget was drafted in January before all of the potential technical/infrastructure options were discussed by the Steering Committee. It will need to be revised upon adoption of a technical platform for DPLA-OHIO, with consultation of the Technology Infrastructure, Metadata, and Sustainability Working Groups.

Metadata Aggregator (required)

The Metadata Aggregator is the heart of a DPLA Hub. This is the tool that harvests partner metadata and prepares it to be pushed to the DPLA. DPLA has the ability to support a wide range of potential aggregator solutions, but the principal method for harvesting metadata to the DPLA is through OAI-PMH, utilizing a MODS metadata profile. Some aggregators include translation services that support metadata ingest of various schemas and via CSV.

Knowledge-base (required)

The Knowledge-base can be integrated into the metadata aggregation software, or exist as a stand-alone service (or database), but provides the rules necessary to harvest an institution's metadata. This includes the protocol necessary to harvest the data, as well as information about the metadata, needed to generate identifiers, links to thumbnails, etc.

Metadata Remediation (optional)

The DPLA requires hubs to provide metadata in a handful of supported formats. Part of the metadata work DPLA asks hubs to undertake is some degree of metadata remediation. Remediation may include enhancing metadata, normalizing metadata, other refinements in metadata to conform with DPLA's MAP and best practices[3]. The strength of DPLA's discovery will be in the data consistency found within the aggregations it harvests. However, while DPLA will work closely with hubs to identify areas for metadata enhancement, and recommend strongly for those enhancements, ultimately, this is at the discretion of the hub. While many current hubs provide a wide range of metadata remediation services – there are equal numbers that do very limited metadata remediation due to financial and other resource limitations, and simply accept the limitations that that will incur when users query for items within the DPLA.

Collection Hosting (optional)

In the early pilots, the DPLA provided funding for hubs to support digitization efforts to provide free and open content to cultural heritage collections. While this funding is now very rare, a number of state DPLA hubs provide collection hosting services. However, given the current environment within Ohio, hosting collections through a central hub appears to be out-of-scope. Ohio presently has a number of community supported options, like the Ohio Digitization Hub network supported by the State Library of Ohio and the Ohio Memory Project; which provide avenues for organizations looking for hosted digitization.

Technology Environmental Scan

Given the needs stated above, the working group evaluated current DPLA hubs, planned DPLA hubs, and possible future development to identify the following technology solutions.

Hydra/Fedora[4]

The Hydra/Fedora option represents the future for the DPLA. DPLA, through an IMLS grant, is currently working on the development of a project known as Hydra-in-a-Box[5]. The goal of this project is to create a simple, turn-key repository application that can function not only as a service for hosting digital collections, but also serve as an aggregation engine for DPLA statewide hubs. By far, this is currently the most technically challenging of all the potential solutions that the State of Ohio could embark on.

---

[3] http://dp.la/info/wp-content/uploads/2015/03/Intro_to_DPLA_metadata_model.pdf

[4] http://projecthydra.org/

[5] http://hydrainabox.projecthydra.org/

Presently, the Hydra-in-a-Box concept is still very much a work in progress, with development versions of the software still 12-18 months into the future.

However, given that the framework is open source, the state could embark on a project to build our own aggregation toolset using the Hydra framework or work with a set of dedicated consultants like Data Curation Experts to develop a custom Hydra stack. This is currently the approach that the State of Pennsylvania is taking. They are developing a custom Hydra application, dplah[6], to support the harvest and aggregation of their statewide content. Co-Chair Terry Reese installed this in-development application on a sandbox server. From this it was discovered that DPLA is built on old versions of Hydra/Fedora and did not work as expected out-of-the-box. Only one record was able to be harvested and there was a lack of sufficient incoming metadata profiles.

This approach makes sense if….

If DPLA-OHIO was looking at creating not just an aggregation, but also a statewide public portal for Ohio Collections, this solution would make the most sense. This is exactly why the Pennsylvania hub has taken this approach. The Hydra platform is serving not just as a platform for aggregating content, but will also serve as the statewide portal to the hubs digital content.

Dependencies

This solution will require a significant commitment of resources to be successful. This solution likely could not be managed by a traditional IT administrator. Rather, the solution would need to be handled by a DevOps team, which would include a programmer (1.0 FTE), 0.25 Project Manager, and an administrator (0.5 FTE). While this solution is likely the most flexible of all the options available, that flexibility is at the cost of convenience. Hydra is a framework, and in order to utilize this toolset, a "Hydra Head" will need to initially be built for DPLA-OHIO use. Time to build would likely take 8-12 months.

Repox[7]

Repox is a set of technologies developed by Europeana to support organizations needing to develop an OAI-PMH access point. Europeana, like DPLA, utilizes OAI-PMH to harvest metadata into the central aggregation, and this project enabled Europeana to offer technology that would allow partners to easily participate in the project. However, in addition to providing an OAI-PMH endpoint, the tool can also be used to aggregate metadata together.

In speaking to a number of DPLA Hubs, Repox has been the primary technology of choice for many organizations. The tool is relatively easy to setup and configure as a stand-alone java application. The tool comes in two flavors. The older version of repox (2.x series) is a java application that uses a java client to interact with a MySQL database. Depending on the hosting scenario and information security requirements, multiple users could be created. The new version of repox (3.x series) is a web application developed using Jersey. Instead of downloading a java client multiple users could interact with the application using a web browser. Europeana encourages use of the 3.x series.

[6] https://github.com/tulibraries/dplah

[7] http://repox.sysresearch.org/, http://labs.europeana.eu/apps/repox

Repox handles metadata transformations through XSLT stylesheets. For each collection being harvested, a separate stylesheet will need to be created to capture and map the content. In speaking with current DPLA hubs using the software, this part of the process represented the lion's share of the technology support for the software. It also represented the biggest complaint around the toolset. For hubs that only provide aggregation services without any metadata remediation, Repox provided those hubs with a near problem free environment. For hubs that wanted to provide a significant level of metadata remediation at the point of harvest, Repox's use of XSLT to support metadata refinements made this process both difficult and frustrating.

This approach makes sense if…

A significant number of DPLA hubs exist because they utilize Repox and it provides these hubs with a solid, easy to manage environment. This solution makes the most sense for statewide hubs that only intend to provide a metadata aggregation to the DPLA.

Dependencies

Support for the system can easily be managed by a traditional IT Administrator (0.25 FTE), though a developer or metadata specialist will be necessary to create the various XSLTs needed to harvest metadata from partner's collections (1 FTE initially, 0.25 FTE after setup)

OCLC Gateway/Collection Worldshare[8]

This is an interesting option, because it currently does not exist in a form that could be utilized by the DPLA and OCLC has indicated through a proposal a strong desire to explore potential partnerships. (See Appendix D.) In this case, OCLC would host all the technical infrastructure for the project. Organizations would work with OCLC to harvest materials utilizing the OCLC Gateway software. Metadata would be harvested and managed by OCLC, and ingested into WorldCat. However, at this point, getting the data from OCLC to the DPLA will still need to be a mediated process. OCLC's current infrastructure does not support the ability to automatically harvest data to an end point. The present workflow then, would be for the central hub to download "sets" of metadata and then provide them to DPLA. As part of the proposed pilot, OCLC has proposed development of new functionality that would establish a streaming endpoint that could potentially be harvested directly by DPLA or other service. In addition to the development work proposed by OCLC, DPLA-OHIO would need to work closely with the DPLA to ensure that this method of aggregation could be supported by the current DPLA legal framework. Utilizing OCLC would represent an entirely new aggregation model and close communication with both OCLC and the DPLA would be necessary to ensure a successful outcome.

This option lowers the technology barriers significantly. In each scenario but this one, the hub would need to manage, develop, and/or host some type of software and work closely with partners to ensure metadata was harvested successfully into the hubs central aggregation. OCLC's Gateway solution flips that model. Organizations would work the hub to identify best practices and notify when new collections have been profiled by the OCLC Gateway – but central technology would not be necessary to manage the resource.

This approach makes sense if…

This approach seems best suited for states without a willing host or lacking organizations that can make a commitment to function as a technology hub.

---

[8] https://www.oclc.org/digital-gateway.en.html

Dependencies

So, while this solution appears to have the lowest technical bar, it has some very significant dependencies that may or may not, make this solution a good fit. First and foremost, while OCLC's proposal indicates a willingness to provide a metadata feed under a CC0 license, the feed would likely represent a subset version of the data. While the proposal demonstrates a strong desire to meet the DPLA licensing requirements, additional discussion will be needed with both OCLC and with DPLA to ensure that any potential metadata licensing issues are addressed.

Second, OCLC has a handful of other requirements for this service.

1) The OCLC proposal has noted that for the purposes of the pilot, members utilizing the service would not have to be OCLC members. However, as the end of the pilot and the service moved into production, participants would take on a service level agreement and be made members of the cooperative.
2) This is a subscription service, which means that the cost for this service will need to be bore either by the DPLA-OHIO service hub or by the individual partner organizations. Presently, OCLC's proposal indicates that service level costs will be bore by each organization utilizing the service, though this may be open to change through negotiation. The cost of this subscription service is not yet determined.
3) The current service was designed around the need to generate MARC data. DPLA requires metadata to be provided in a different format, and prefers data to be made available via OAI-PMH. Currently, this is beyond the scope of OCLC's collection manager tool, and would represent significant new development work for OCLC to take on to enable this tool to work within the DPLA model. The OCLC proposal, as currently written, indicates a timeline starting April 2016, ending December 2016. While these timelines are likely open to discussion, OCLC's proposal does infer a strong desire to follow an aggresive timeline for development and execution of the pilot.
4) All data harvested by OCLC will be loaded into WorldCat. On its face, this seems like a good thing, but many organizations make a conscious decision to not load their digital collections into WorldCat for a wide range of reasons. Forcing ingest into WorldCat as a prerequisite to participate in DPLA-OHIO has a number of potential problematic components:
    a. Tying ingest into WorldCat as a prerequisite to participate in DPLA-OHIO requires members into a relationship with OCLC around their digital collections metadata -- a relationship that many members may have already evaluated, and for local decisions, passed on.
    b. DPLA was developed to promote metadata freedom, collaboration, and choice. Tying ingest to WorldCat as prerequisite to participate in any DPLA-OHIO project would betray those initial goals from an ideological perspective. This type of relationship would make Ohio an outlier and could impact which collections contributing partners may make available via the hub.

Primo/Encore/Summon/EDS

A wide range of vended solutions exist that could potentially provide aggregation functionality[9]. These services, traditionally used to support discovery across a wide range of library resources, could be

[9] The Mountain West Digital Library is an example of an organization that utilizes Primo to support their DPLA hub.

purchased and scoped to support a simple aggregation (no metadata remediation) and public interface for that aggregated content.

This service makes sense if….

This approach seems best suited for states without a willing host or lacking organizations that can make a commitment to function as a technology hub.

Dependencies:

The most significant dependency is fiscal. These solutions tend to be expensive, in part, because they provide a wide range of functionality and are usually negotiated as part of a package with other products. While any of these solutions could potentially provide the type of aggregation service for the DPLA-OHIO, it might be difficult to justify the financial costs given the limited needs of the project. Though, if this solution is deeded to be viable, a more thorough investigation of the costs will need to be undertaken.

It should also be noted that choosing a vended discovery solution for the technology stack will likely require an RFP process for DPLA-OHIO that affects the implementation timeline.

Local Development

While this probably is the least attractive option, the creation of a local aggregator and OAI-PMH server is a possibility. At the same time, this solution would require the most technical support upfront, as all the work would be done by a developer.

This service makes sense if…

We have no other options.

Dependencies

Developers – this solution would succeed or fail based on the developer(s) (1.0 permanent FTE) and the project manager (0.5 FTE) handling the requirements gathering and sprint planning. Additionally, the organization would require a system administrator (0.5 FTE), that would also likely need to have some development skills to function in more of a DevOps role.

Potential Tactical Strategies for a Technical Infrastructure

At this point, providing a set of recommendations is murky, at best. While the working group has provided a set of recommendations based on our best understanding of the current environment and available technical landscape, the working group sees a number of potential scenarios where different technology infrastructures would thrive. To that end, the working group will propose a set of potential outcomes based on the following scenarios:

Scenario 1 (Technical Hub: State Library of Ohio [or like organization such as OhioLINK])

The first scenario assumes that in committing to supporting the initial 3 years of the project, the State Library of Ohio (or like organization) intends to commit to supporting the technical infrastructure. In this scenario, the State Library of Ohio (or like organization) traditionally would contract IT support services for a project like this, meaning that the host organization would likely be able to procure primary support

for the development and maintenance of the service from a general system administrator, without the need for developer support.

With those restraints in place, the working group would strongly recommend the use of Repox. In all the conversations with hubs currently utilizing Repox, the initial setup and long-term maintenance of the system was the most straightforward and least time consuming. Additionally, by choosing Repox, the technical working group would recommend that the DPLA-OHIO metadata group focus primarily on providing best practices, but that the initial 3 years of the project provide zero metadata remediation services. Additionally, each collection managed through Repox will need to have an XSLT created to support harvesting. This work will likely need to be done outside of the hosting organization – the working group would recommend that this be work that the hub's governance body contract or find support from within the partner institutions.

Scenario 2 (DPLA-OHIO as an Administrative Hub)

This scenario assumes that no organization can be found to support the technical infrastructure for the project. In this case, the working group believes that the project has the following two options:

1) Contract these services out to a vendor utilizing a tool like Primo. Companies like Ex Libris are interested in getting a toehold in Ohio – so there is a high degree of likelihood that the state could negotiate a very favorable contract. Organizations in the state already have a long standing partnership with III.
2) Work closely with OCLC to explore a potential option related to the OCLC Gateway

When considering Scenario 2, one additional consideration is a fiscal one. Subscription or hosted services like Primo, Encore, or OCLC's Gateway service will require an annual licensing and maintenance subscription. Costs related to the subscription service will vary, but this raises additional barriers to the project as these costs will need to be borne initially by the host organization, but ultimately by the project participants.

Scenario 3 (Central Hosting With Developers)

Scenario 3 assumes that DPLA-OHIO has found an organization that not only will provide a technical home, but will dedicate significant hosting, administration, and developer resources to the effort. In this scenario, any potential technical stack would be possible. Though, this scenario represents the only instance in which an organization should consider the Fedora/Hydra stack or custom development to support the DPLA-OHIO hub infrastructure. At this time, both of these options require significant development resources to not only implement a solution, but to provide the long-term care and feeding of the project. Of all the proposed scenarios, scenario 3 offers DPLA-OHIO the greatest level of flexibility (in terms of developing a metadata remediation plan, supporting multiple harvesting protocols, etc.), but does require the highest level of commitment from a hosting organization. Within the state of Ohio, there are a handful of institutions that could potentially fill this role.

Scenario 4 (Technical Hub in the Cloud; Shared Technical Administration)

One of the unique aspects of DPLA-OHIO is that much of the work and interest around joining the DPLA has come from the bottom; curators, collection managers, individuals that work every day with users and want to see Ohio Collections better represented. This has had a lot of advantages, but it has also contributed to the uncertainty around the technological home for the project.

In this scenario, DPLA-OHIO would continue to build on that collaborative approach, and develop their infrastructure outside of any one institution -- hosting the content in the cloud. There are a number of interesting benefits that come from this approach.

- Infrastructure costs are fixed -- services like Amazon Web Services provide a well defined set of costs for hosting and maintaining services. Strategic use of a cloud infrastructure can reduce hardware costs while allowing for more fine grain resource management.
- Security challenges are minimized -- one of the most significant issues for many hosting institutions is navigating their organizational security policies. By pushing this outside of any one organization, those issue become mute
- Instant scalability -- utilizing a cloud service would allow DPLA-OHIO to scale a service to meet the needs of today, and then easily add resources as needs and services expand. With cloud hosted servers, it's much easier to add additional resources when needed than to upgrade physical servers located in a data center.

This approach also offers another opportunity that could be uniquely Ohio. Like many open source efforts (Fedora, ArchivesSpace, DSpace), partners make commitments to support the project with their time. Rather than having the management and administration of a service tied to a single organization, the DPLA-OHIO would create a standing technical group, with responsibility for shared administration of the service. By utilizing a shared group, knowledge would be passed throughout the partners, no one institution would be responsible for the service, and we could develop a new collaborative model for DPLA members.

There are obviously risks to this approach. Shared administration would require active participation by partners to maintain a set of services, and would require the DPLA-OHIO to consider potential services based on the expertise available to them. Additionally, partner organizations would need to see the value of providing their staff resources to the project; though, we have many examples (like OhioLINK) where this type of value proposition has been considered and found to be beneficial.

What type of technology stack would work well in this type of scenario? The Local development options, the Hydra Fedora stacks, or Repox. All three of these solutions would thrive within this environment, though the level of expertise to implement and manage each of these solutions would vary greatly -- as noted in the environmental scan above.

Scenario 5 (Technical Hub in the Cloud; Single Administration)

In this scenario, a single organization would be responsible for staffing the administration and management of the technical infrastructure -- but rather than the infrastructure living at anyone one institution, it would be hosted in the cloud. While this solution wouldn't mitigate the people costs related to managing the technical infrastructure, it would have a handful of practical benefits:

- By hosting in the cloud, it would make it possible for the organization managing the technical infrastructure to change without requiring infrastructure to move.
- Hosted infrastructure would live outside of local organization enterprise security environment and policies.
- Ability to scale infrastructure to meet technical needs allowing finer control of hardware costs.

In considering the initial 3-year pilot, flexibility related to infrastructure and hosting may provide significant advantages, as it would allow the project to more easily pivot technical infrastructure management and hosting long-term.

Recommendations

In considering the various scenarios, available technology solutions, and potential technologies that could be implemented by the eventual DPLA-OHIO hub - recommending a particular technical stack without first considering the expertise and capacities of the hub organization would be problematic. As a result,

this working group won't forward such a recommendation. With that said, there are a handful of recommendations that the Working Group can make, and we believe should be considered by the Steering committee and any potential DPLA-OHIO host.

- *DPLA-OHIO should manage their own technical stack for the initial pilot*
  It would be tempted to outsource the technical infrastructure for this project, given the wide range of activities that will need to be accomplished to make the DPLA-OHIO a success. A new governance model, partnerships, metadata best practices -- these are all things that take significant time and resources and adding the management of a new technical infrastructure to these activities does complicate the project. But in this case, it is a necessary complication. The goal of the 3-year DPLA-OHIO pilot is to evaluate the sustainability of creating a DPLA hub within the State of Ohio, and ensure that decisions made today, don't preclude or overly complicate the long-term viability of the project. Given these goals, it is imperative that DPLA-OHIO host their own technical infrastructure in order to better understand and assess the overall technical needs for the project. While these needs appear daunting today -- one must remember that the technical challenges posed during the initial year of startup will largely diminish as the project moves from startup to long-term maintenance and assessment. By outsourcing the management of the technology stack, DPLA-OHIO would compromise these goals as an assessment of the technology and technology hosting requirements would be as unknown in 3-years as they are today. As DPLA-OHIO considers the technology stack and management options, the following issues need to be considered:
    a. Hosting our own technical infrastructure will provide the DPLA-OHIO the information that it will need to better understand the long-term sustainability needs and requirements related to running an aggregation hub, and allow the organization to make the best long-term decisions for the cooperative.
    b. The technical solution needs to hold members harmless...i.e., many outsourced solutions may require members to harvest content into vendor-based metadata cooperatives to participate in DPLA. While DPLA-OHIO may see this relationship as a 3-year pilot, this decision potentially binds DPLA-OHIO members to these aggregations beyond that period.
    c. The vendor market is still maturing, with new options to likely develop over the next three years. A lean, self-hosted, 3-year pilot with a known timeline will allow DPLA-OHIO to launch while giving the vendor market time to mature and introduce competitive solutions.
    d. Given that this is a pilot, the ability to change technology stacks and the long-term ramifications related to the uncoupling if the technology stack is outsourced, needs to be given full consideration.

- *Repox represents DPLA-OHIO's best option for a locally hosted technology stack*
  While the eventual host organization should have the ultimate decision around the technology stack that they are willing to support, the above scenarios do provide some helpful guidance. For the 3-year pilot, Repox likely represents the lowest barrier to developing, and implementing an aggregator for the state of Ohio. The tool has served as the initial aggregation system for many of the DPLA hubs, in part, because its minimal technical resources minimizes the potential for failure. This simplicity comes at a cost, it assumes that metadata remediation will be minimal and that the hub will only function as an aggregator -- but for the 3-year pilot, this is likely exactly what DPLA-OHIO needs to be successful. What's more, Repox could be installed and managed by a significant number of partner institutions (assuming they were willing) and the expertise needed to develop the XSLT crosswalks lives throughout the state of Ohio. Certainly, dedicated staff will be necessary to ensure the greatest level of success -- but unlike many other technical

solutions, this option will allow the host organization to leverage expertise from around the state in a way that wouldn't be possible using many of the other hub technologies.

- *Technical infrastructure decisions need to remain flexible and easily fungible*
  Remember, this is a 3-year pilot, and the technology environment is going to change. The State Library of Ohio has pledged initial support for a 3-year pilot investigating the feasibility and long-term sustainability of hosting a DPLA hub for Ohio. This means that this initial 3-year period should be focused on:
    a. Working with DPLA to understand where our and their interests intersect
    b. Cultivating the initial partners interested in making their content available to the DPLA
    c. Developing a shared set of metadata best practices which, while not prescriptive, will provide guidance to partner institutions and provide the information necessary to understand how metadata decisions impact indexing and discovery within the DPLA
    d. Develop an aggregator and begin sending DPLA content. Most DPLA hubs bring up 8-10 partners a year. DPLA-OHIO is potentially looking at bringing content from ~50 CONTENTdm instances and numerous other systems to the DPLA. This will take time, and a planned implementation will be needed.
    e. Develop an assessment protocol for evaluating the impact DPLA ingest is having on partners.
    f. Educating partner institutions around the harvest technologies, and the benefits of potentially making content available via other avenues like OCLC's WorldCat, or making content search engine harvestable.
    g. Charge a standing technology working group to reassess the information landscape in Year 2 of the pilot. A number of new technologies are being developed specifically to support the DPLA and DPLA hubs. Understanding not just what these new technologies are, but their benefits to the current aggregation stack, and the time to adopt a new technology stack. These technologies should all be developed prior to the end of the 3-year pilot.

- *The pilot project should include a standing technology working group*
  The DPLA-OHIO effort will require technology input and support - regardless of the final stack composition - from multiple partners throughout the lifespan of the 3 year pilot project. The final governance model for the DPLA-OHIO should include a standing technology group comprised of members representing participating communities. This will help ensure that the expertise and responsibilities related to the implementation of a DPLA hub within the state are not limited to just one organization. Specific responsibilities will depend on the final program structure; this group might provide hands on support for XSLT development or may simply function as a sounding board for the host organization or potential participants.

  A standing technology working group will need to develop onboarding paths for institutions without OAI-PMH. This group would also be responsible for conducting an environmental scan and assessment during the 2nd year of the pilot to provide feedback to the advisory group and host organization regarding the long-term sustainability and technical landscape.

  A standing technology group would also be able to help promote innovation and technical development within the program. Activities like hackathons events to encourage engagement with the DPLA and it's data, or the ability to foster creative projects related to Ohio using the DPLA-API.

- *Metadata remediation will primarily be a local concern; remediating at the center will be minimal, and as needed during the 3-year pilot*

As the DPLA-OHIO hub is initiated, the hub will quickly find that DPLA has a list of metadata best practices that they'd like to see applied to all aggregated data sets. For the three year pilot, we'd recommend that the hub only perform the minimal level of metadata remediation necessary to provide aggregate data to DPLA. There are a number of reasons for this:

a. Given the number of potential partners, and the need to ramp up the organization and aggregation, it adding new collections and partners will take time. Metadata can always be enhanced and reharvested. The most important thing early on is having early success; it is imperative to make the barriers to participation as transparent and minimal as possible.
b. In discussing challenges new hubs faced during their first couple of years, the most common challenge cited was metadata work. The desire to provide clean metadata can bankrupt a project. For the pilot to succeed, metadata remediation will need to be done as an incremental process.
c. Over the pilot, particular issues should be considered:
    - How do we define remediation in the post-pilot? Normalization, enhancement, refinement?
    - How do limitations in member systems impact local remediation recommendations?
    - If some or all of the metadata remediation were moved to the center, what skill level would be necessary to handle that work? And how would it impact the stack (i.e., Repox's stack doesn't fit the -- remediation at the center -- model, while a solution like Hydra or OCLC hosting might).
    - If some or all of the metadata remediation were moved to the center, how would that impact technical resources? Metadata manipulation and crosswalking is a resource intensive process. How does this impact the hardware resources at scale?
d. At a practical level, minimal metadata remediation will give the organization the ability to build faster, and with more varied staff expertise.

- *OCLC needs to be explored as a partner*
  Regardless of OCLC's ability to function as a central aggregator, OCLC's production of CONTENTdm and their Gateway tools provide an interesting avenue to potentially supporting those institutions that lack the ability to provide an OAI-PMH data feed. While some details of from the OCLC proposal would need to be worked out, working with OCLC throughout the 3-year pilot could provide significant benefit not only for DPLA-OHIO partners and OCLC, but also other potential DPLA hubs.

- *Evaluate ways to facilitate access and discovery to the primary resources about Ohio*
  One benefit of placing metadata into the DPLA is the ability to leverage the DPLA API. As Emily Gore quoted at the DPLA-OHIO symposium, "the most interesting thing to do with your data will be thought of by someone else." While the initial focus of DPLA-OHIO should be getting the service hub off the ground, it should not lose sight of the end goal of facilitating access and discovery. DPLA-OHIO sponsored hackathons or workshops about using the DPLA API could result in amazing tools such as a statewide portal of resources about Ohio that includes content from all DPLA contributors. Co-Chair Terry Reese demonstrated a proof-of-concept for such a portal at the DPLA-OHIO symposium. The interest and expertise to create tools using the DPLA API already exist in the state. DPLA-OHIO should find ways to leverage those interested parties to further this goal.

Significant Challenges

There will be a number of challenges related to the development of the DPLA-OHIO hub, in part because this represents a new program structure, developing new partnerships and a shared history. And while these challenges will test the group, and likely push members to consider local practices, a number of specific issues that relate directly to the implementation of a technology stack. These challenges are:

- *Metadata remediation will be a significant barrier for many partners*
  While surveying stakeholders, one thing that became readily apparent is that metadata remediation will be an ongoing challenge. Issues related to compatibility with local best practices, the availability of necessary staff, and available metadata expertise all resonated with stakeholders. These issues were to be expected. What was more surprising was the significant limitations many potentials partners had to perform local metadata remediation due to limitations in their local content management systems. This concern was called out many times during the symposia, and by many members of the working group. As such, one goal of the initial 3-year pilot should be to determine not only the success the organization had in effecting metadata remediation and local best practice at the partner level, but also seriously investigate what level of remediation can be taken on at the center and the implications that will have on the central aggregations technical staffing.

- *Technical expertise will be a challenge of varying complexity at all levels*
  Technical expertise will impact this project in a variety of ways. It will influence the stack that is selected, the potential institutions that can host, the long-term decisions around metadata remediation. Throughout the symposium, and in conversation with stakeholders, a number of specific concerns have come to the top:
    - Potential host organizations have specific concerns related to expertise necessary to host specific technologies. As noted in the environmental scan, DPLA-OHIO has a wide range of technology options to choose from. Some of these options have significant technical requirements, which would severely limit potential host organizations; while others mediate ease of use and implementation by making specific assumptions related to scope (i.e., Repox assumes minimal metadata remediation from the center, Hydra assumes deep customization, but with a high developer cost)
    - Metadata remediation has clear technical impacts such as:
        - Any solution that requires significant metadata remediation from the center will require steep and long-term technical expertise to remain sustainable.
        - Any solution that requires significant metadata remediation from local participants may put participation in DPLA-OHIO out of reach due to their own local system and staff challenges.
    - The inability to provide an OAI feed, and the need to be able to support multiple metadata ingest streams.

  For contributing partners, technical expertise can be a huge barrier, not only to achieve metadata remediation but also enabling an OAI-PMH feed. This will be a particular challenge for institutions without a OAI-PMH option, for which onboarding methods will need to be developed.

- *OCLC has the potential to be both a valuable partner and a distraction*
  OCLC and Ohio institutions have a special relationship -- and as such, OCLC has the potential to be a valuable partner in any endeavor such as this. The challenge will be ensuring that the partnership makes sense for both organizations and is one that is sustainable over time, regardless

of the technology infrastructure utilized.

- *Sustainability and Budget costs are going to be a moving target, especially during the first 2 years*
  Due to the ramp up costs related to starting a new hub, developing a sustainability model may be difficult if basing costs on the first 2 years of the project. And yet, it is vitally important that we get this right. DPLA-OHIO will cost money and time -- and whether these are allocated from existing resources or through the hiring of new staff, funding for these positions and resources will need to come from somewhere. A major concern of the Working Group is that these funding models must include tiers that don't create undue barriers for smaller organizations to participate.

- *Current DPLA Timeline for implementation*
  The current timelines DPLA provides to future hubs places initial harvest of metadata from an aggregation at 3 months. Given that DPLA-OHIO is starting from nothing, this timeline needs to be moved to reflect reality. A timeline of 6 months before initial metadata harvesting would provide a much more realistic and attainable implementation goal.

Concluding Thoughts

Shared Technology Support

Regardless of the technology stack, the DPLA-OHIO effort will likely require a shared technology model throughout the lifespan of the 3 year pilot project. Whether that group provides hands on support for XSLT development or simply functions as a sounding board for the host organization, the final governance model for the DPLA-OHIO should include a long-standing technology group to ensure that the expertise and responsibilities related to the implementation of a DPLA hub within the state, doesn't stay locked up within a single organization. A standing technology group would also be able to sponsor hackathon events and foster creative projects related to Ohio using the DPLA-API.

Long-term Technical Planning

While the technology working group believes that strategies outlined above represent the best chance for success within each potential scenario, it should be noted that this likely will be a short-term solution. Long-term, the DPLA is working to develop their own software platforms to simplify the management of metadata within statewide hubs. The decisions we make now likely will need to be revisited after the conclusion of the 3-year pilot to determine if they continue to make sense, or if new technology should take its place, and does the new technology significantly shift the resources needed to support the project.

The strategies discussed here were developed primarily with type A institutions in mind, those with OAI-PMH capabilities. If DPLA-OHIO hopes to launch a service hub by first quarter 2017, we will need to be pragmatic about where we start so that we can achieve velocity and get the hub off the ground. The technology stack we choose will affect our ability to onboard other institution types. After the successful launch of our hub, further strategies for onboarding other institution types will need to be developed.

Strategically, thinking about how DPLA-OHIO could eventually integrate content from organizations who do not have OAI-PMH capabilities may provide a clear avenue for partnership and experimentation with a group like OCLC. While the OCLC Gateway tool has been discussed above as a potential technology stack for hosting the entire DPLA-OHIO aggregation, not using them as the central hub doesn't exclude OCLC from being a potentially valuable partner in this process. OCLC's Gateway tool supports a wide range of metadata formats, opening up a wider range of data import mechanisms to partners. Assuming OCLC and DPLA can come to terms with the licensing issues, DPLA-OHIO could

potentially utilize OCLC's Gateway export as a shim between partners without OAI-PMH support, and the central aggregation. It would provide OCLC an opportunity to explore using the Gateway to support DPLA hubs in the future, and allow DPLA-OHIO to continue forward at their own pace, knowing that any partnership done with OCLC could potentially lower the barriers for participation.

Appendix A: Technology Matrix

	Subscription Cost	Hardware Cost	Technical Expertise	Customizability	Time Required	Total Anticipated Staff Time (in FTE)
Hydra/Fedora	optional*	high	high	fully	long	1.00 Developer 0.50 DevOps Administrator 0.25 Project Manager
Repox	no	low	medium	somewhat	medium	1.00 Developer (4 mos.) 0.15 Developer 0.25 Sys Admin.
OCLC Gateway	yes**	none	low	limited	medium****	0.50 Project Manager (4 mos.) 0.25 Project Manager
Primo / Encore/ Summon/ EDS	yes***	none	low	limited	medium	0.25 Project Manager
Local Development	no	high	high	fully	long	1.00 Developer 0.50 DevOps Administrator 0.50 Project Manager

* Hydra/Fedora development, management, and hosting can be contracted with company Data Curation Experts.

** Subscription costs can be paid at the center (DPLA-OHIO Hub) for all members of the project, or by individual members.

*** Subscription costs to vendor would be paid at the center (DPLA-OHIO Hub)

**** Gateway functionality does not currently exist. Would require both OCLC development, and stakeholder development of functionality list and user stories.

Appendix B: Repox Interview Notes

[THESE NOTES ARE FOR THE STEERING COMMITTEE ONLY AND WILL BE REPLACED WITH SUMMARIES BEFORE PUBLIC RELEASE.]

Repox Discussion with Lisa Gregory and Stephanie Williams; North Carolina Digital Heritage Center

Lisa Gregory; Interim Director
Stephanie Williams; Programmer
Interviewer: Terry Reese
Date: Feb. 3, 2016

I had the opportunity to speak to members of the North Carolina Digital Heritage Center on 2/2/2016 – specifically around their implementation of Repox, the time it took initially for them to setup the project, and their long-term support with the project. Here are the highlights:

- As an organization, the North Carolina Digital Heritage Center's DPLA operations are run on a skeleton budget. When the NCDH joined the DPLA, they did it as an extension of their already existing program. They are funded by the State Library of North Carolina, have a mandate to support digitization and collection hosting, and are funded via the State Library and LSTA funds. When Jenn Riley began working with the DPLA as part of the pilot, the organization took the project on without additional funding (and still, to this day, runs in this capacity). This meant that the NCDH has made some very specific program decisions:
    a. They don't recruit content into their portal. They work with the people that are highly motivated and interested in the project. They also work with folks that are technically capable of working with them. If an institution doesn't want to share their metadata, they won't try to convince them. If an institution doesn't have the technical capacity to share their collections, they may see if that institution would like the NCDH to take on hosting responsibilities – but generally, they accept content only from organizations that can provide them with a valid, easy to understand, OAI-PMH feed.
    b. As an organization, they do zero (well, almost zero…they strip some data and add a field to identify the organization an item came from) metadata remediation, and require no remediation from their partners. There are two main areas of thoughts behind this:
        - DPLA isn't providing any funding to support metadata remediation at the Hub level – and given that this is something being done essentially by part-time staff, there just aren't the resources.
        - Repox does what it does well (creating a data aggregation), and doesn't do much else. Organizations that have struggled with Repox have struggled because they have tried to shoehorn functionality like metadata remediation into their process. Repox doesn't do that…easily.
        - As a group, the NCDH felt that asking partners to change their metadata from past collections wasn't sustainable. They provide their partners best practices, allow organizations to see the results of not having specific metadata when rendered in DPLA (i.e., if geographic headings are not standard, your content doesn't show up in a geographic search) – but as an organization, they are a hands off metadata shop.
        - They have no formal agreements with data partners. The NCDH works with members that ask to have their metadata aggregated. They do no education around what that means (CC0), and they assume that their partners understand

the terms DPLA requires when making metadata available. As such, they don't ask data providers to sign formal agreements.
        - Finally, I was interested in DPLA's response to the lack of metadata remediation. Apparently, DPLA has spoken with them a number of times, but their attitude is that as long as they are maintaining the aggregation, DPLA will just have to live with what they get.
    c. Technically, the NCDH utilizes Repox – specifically version 2.2.7 (this is the old development branch).
        - They did note some trepidation around using repox. Apparently, there were a couple of years when no one was supporting it, and the project website disappeared. The new github page is new, and also based on a different architecture. This concerns them a little, but not enough to consider a technology switch. The minimal support work is what keeps them on the version they are using.
        - Information about the Repox Versions:
            i. Version 2.2.7 – Java client/application with a LAMP backend.
            ii. Version 3.x – Java web application utilizing a LAMP backend with Jersey as the interface framework
        - Initial setup:
            - NCDH noted that the initial up and running time for the project was approximately 4 weeks. This included developer time to work on the initial set of 6 XSLTs for the initial metadata harvests, and a system admin to get Repox running within their environment.
            - They run Repox on windows. They did this because they had significant trouble getting Repox secured on their Linux infrastructure. They didn't elaborate – but the issue threatened to derail them, so they run Repox on a standalone windows server, only accessible by IP address by NCDH staff and the single DPLA harvester.

- Long-term:
    a. Repox is managed as part of their normal infrastructure. The software is managed by their programmer, which said she spends ~1-2 hours a month with the program. A system admin just maintains the Windows Server.
    b. New members take ~1 week programmer time to get ingested into the aggregation. This time is spent creating the XSLT that generates the MODs feed DPLA requests.
        - New members take ~2 weeks for the program manager, as she does the initial metadata discussions and profiling with the interested member.

Things that they had for us to think about:

1. If DPLA-OHIO wants to do anything beyond simply creating an aggregation, Repox may not be the best fix. It certainly has the lowest barrier to entry – but it does one thing very well, and that is about it. If we need to do more than that, we'll find very quickly that we will be fighting with the toolset.

Repox long-term support is still up in the air. It's definitely supported by Europeana and others – but in the time they have used it, 2/3 of that time, the project website simply disappeared. They're biggest worry is that support may not be available long-term; especially for the development branch that they are most comfortable with as the 3.x branch would require more hands on treatment. What they like about the 2.2.7 branch is that it's a completely self-contained application.

2.

Appendix C: OCLC Gateway Interview Notes

[THESE NOTES ARE FOR THE STEERING COMMITTEE ONLY AND WILL BE REPLACED WITH SUMMARIES BEFORE PUBLIC RELEASE.]

OCLC Meeting Notes
Taylor Surface/CDM Staff (Seattle)
Interviewer: Terry Reese
Date: Jan. 21, 2016

I had the opportunity to spend a few hours talking a bit more to Taylor about the OCLC Gateway and OCLC's interest in potentially pursuing a partnership with the DPLA-OHIO effort. I'm not sure that I learned anything new during the meeting, but I do have a better idea of how we could move forward if we were interesting in exploring this as an option, as well as a few other potential thoughts for discussion.

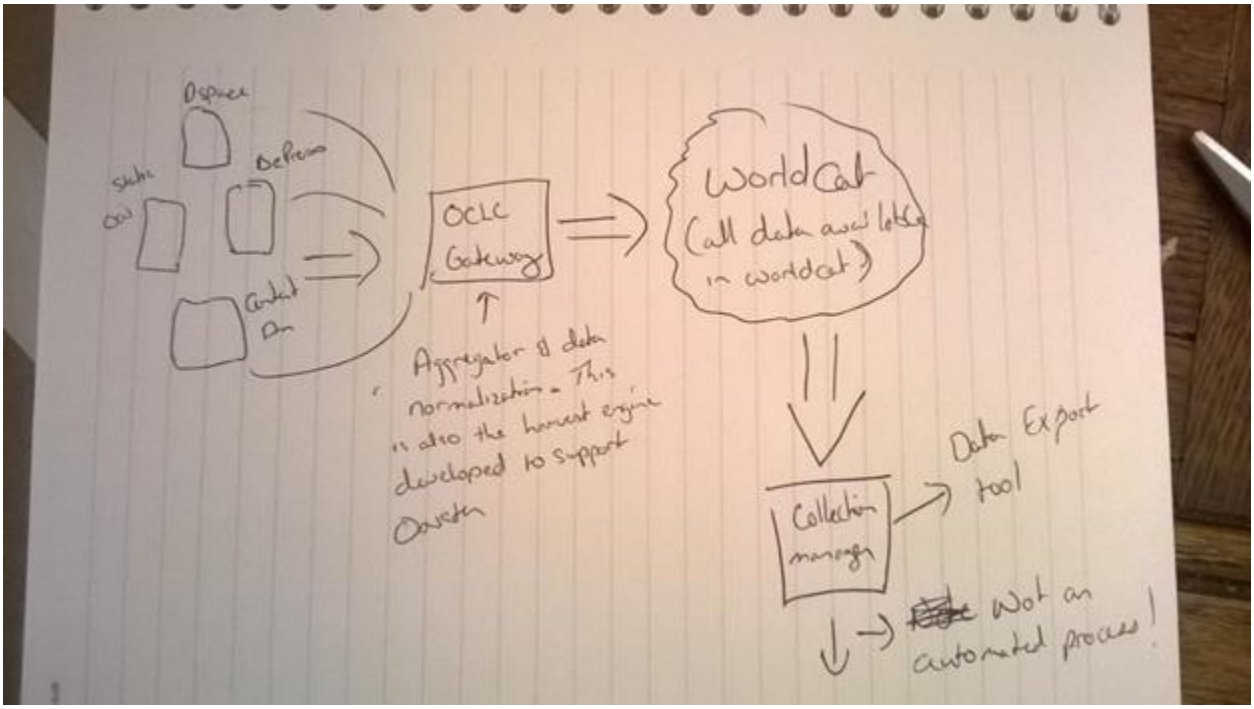
Background

A little background on why OCLC is interested in this. OCLC's gateway export infrastructure was developed when they took over OAISTER from the University of Michigan. The gateway tool was developed as a way to simplify metadata harvest, provide some simplified metadata normalization tools, and encourage organizations to register their digital collections with OCLC. What OCLC got out of this project was more content to add to WorldCat. All data harvested through the gateway tool is loaded into WorldCat. By loading them into WorldCat, these records become available through OCLC's other management tools, including metadata export using their collection manager tool.

So why is OCLC interested in this now? Well, this isn't altruistic. OCLC is looking to sell a service and is looking to make sure that they are not left out of the digital collections space. The development of Hydra in a Box and the push by DPLA to have a presence in each state raises the very real possibility DPLA, and note WorldCat, would become the preferred location for exposing digital collections in aggregate. Given that, they are looking to develop a service offering that they could take to other states looking to join DPLA by providing the infrastructure necessary to handle the data aggregation, allowing the hubs to focus on education, outreach, recruitment, and sustainability.

Infrastructure model

OCLC's aggregation infrastructure currently looks like the following:



This is the infrastructure as it exists today. OCLC manages something like 2100 collections and approximately 40 million items through its gateway. The gateway service provides user the ability to select and profile their records – offers some limited normalization functionality, and passes information into WorldCat. As envisioned now, the hub would then gather all the metadata together on a particular schedule, and submit that data for ingest into DPLA.

So what would need to be done? The Gateway export tool is pretty limited in what it does – but there is the potential to create a DPLA profile that would enable OCLC and a partner to define a set of metadata normalizations so that data could be enhanced, with enhancements potentially being passed back to the local user. This enhancement work would need to be identified and done. Likewise, the collection manager – DPLA envisions a service that isn't a data dump but more of an interactive harvest. This would need to be enhanced to support that type of use case.

Of course, I learned some interesting things – obviously, in order to use the harvester and infrastructure, everyone would need to be an OCLC member and have an OCLC symbol. Is this a barrier? I'm not sure. One area of discussion that we had was a configuration that looked like the Orbis-Cascade catalog in Oregon. There, the consortia had its own symbol, and consortia records would have that symbol attached. If individual membership was a barrier, then something like that might be possible. Of course, I also learned that this would be a subscription service. If there was interest in this as a potential technical solution, I'd recommend the governance group having early conversations about what the subscription cost might look like, and what type of subscription model DPLA-OHIO would want to enter into. There are a couple different options, ranging from individual organizations subscribing to the service, to the hub contracting for all members. Would this be cost prohibitive? How does this compare to developing and running a local solution? All questions that would need to be sorted out.

Pitfalls

There are a few – but the biggest is the licensing issue. Probably the most important thing to come out of this meeting is that Taylor will take this question to OCLC's legal council today. Unless OCLC was

willing to allow data to be contributed to DPLA as CC0, it's a non-starter. This has been relayed as part of the conversations with Taylor and DPLA folks, it's what Emily told this group during the symposia…this is the price to play if you want to be in the DPLA. The answer to this question though is important regardless of the option that we might utilize. Unless OCLC makes this allowance, DPLA-OHIO couldn't use as part of our aggregation any metadata that was obtained via OCLC's gateway harvester or pulled out of WorldCat. So, if we had members using the gateway export themselves – we'd still have to require them to profile their data into the DPLA aggregator in order to contribute their data to DPLA. Without a clear statement from OCLC, the metadata couldn't be used due to the existing license. However, as I say, this is something Taylor recognizes and is taking up with their legal department and will be hoping to have with DPLA-OHIO and DPLA in general.

Where could we go from here

So, after talking to Taylor, I asked him how he envisioned this moving forward. The reality is that DPLA-OHIO doesn't have a hub at this point, so there isn't an organization that can take on a contract for this project, and, honestly, we can't go forward discussing this in a proposal to DPLA until the licensing issues are sorted out.

So – here's what we came up with. They are interested in planning out some work, the idea being that they would like to put developers on enhancing the gateway project by this summer. Also, knowing that DPLA-OHIO couldn't move forward that quickly, we discussed a focus pilot. Working just with OCLC members with digital content, OCLC, DPLA-OHIO and DPLA would discuss and test a set of workflows that could facilitate this process for Ohio organizations. At the same time, governance or the steering committee could begin to engage OCLC around potential subscription scenarios to determine potential cost and sustainability options.

Why might we want to do this

It sounds like OCLC is motivated. I think that they are trying to find a way to stay relevant here – and this offers them an opportunity to potentially develop a new long-term subscription service offering. I've often found that when OCLC is motivated, good things happen. Obviously, there are a lot of unknowns, but a limited pilot would allow DPLA-OHIO to look at another potential option, and begin profiling metadata (which we need to do anyway).

However, the technology is likely the smallest reason to potentially participate in a pilot. As of right now, finding a technical home for this project is one of the hang-ups. We have some potential technical solutions available to us – but these will involve technical staff and infrastructure. This is going to have a high upfront cost, and will ask one institution to shoulder infrastructure for the state. I think we have a lot of folks excited about doing this – I don't think we have anyone excited about hosting the infrastructure. Does it make sense from a resources and sustainability perspective to focus the hub on education and outreach, and partner on technology? Would that make it easier to find a central hub?

Why wouldn't we want to do this

OCLC doesn't always have a great record of moving things from pilot to production. While they seem motivated, we may be asking them to do a lot of unfunded work – and even after the pilot, potentially not go with that option. My guess is that any pilot likely would be just using the current infrastructure, with potential enhancements identified but not implemented unless this went beyond this phase, which may be a harder sell when making a bid to DPLA. Of course, there is also the license issue. Unless this can be

resolved, DPLA won't touch this information. By not pushing forward with a pilot and continuing to push for resolution on the licensing issue, we may be able to influence OCLC's thinking on this issue.

Appendix D: OCLC Proposal for Support Services

[THESE NOTES ARE FOR THE STEERING COMMITTEE ONLY AND WILL BE REPLACED WITH SUMMARIES BEFORE PUBLIC RELEASE.]

OCLC Proposed Pilot Project to Provide Services to Support the Ohio DPLA Service Hub

March 4, 2016

Purpose

Ohio libraries want to create better visibility for their unique digital collections on the Web. Through a pilot, libraries from Ohio, the Digital Public Library of America, and OCLC will create a workflow for libraries that want to share the metadata for their unique digital collections with DPLA. In addition, this metadata will become a part of WorldCat, so that libraries will have increased visibility of their digital collections through Worldcat.org.

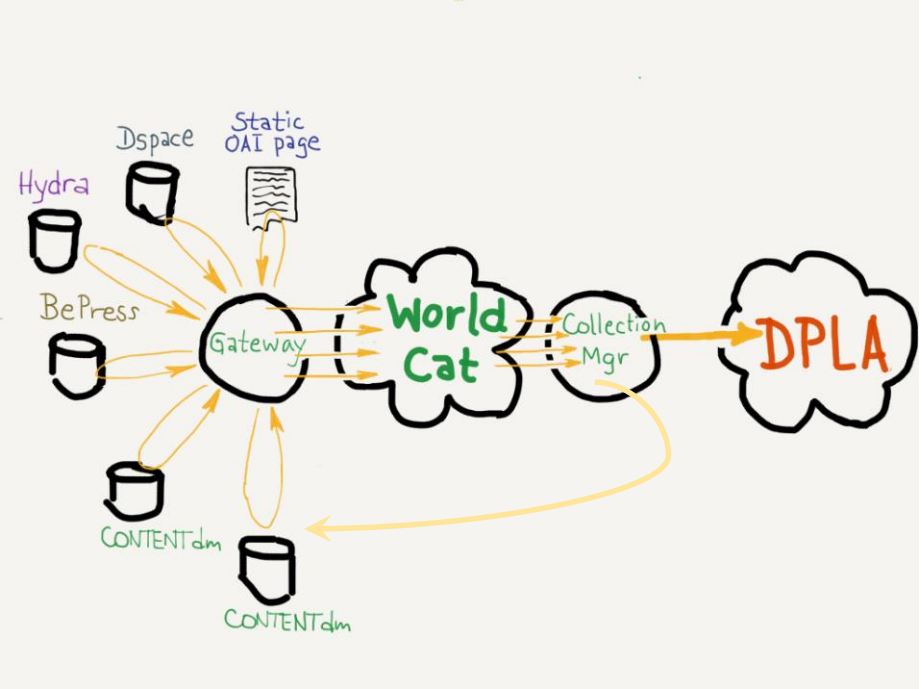
Project description

The project participants will design and create a workflow infrastructure to support sharing metadata from a library's digital repository with the Digital Public Library of America and WorldCat. The metadata will be triaged, corrected, and enhanced by OCLC according to DPLA's best practice recommendations for shareable metadata. The metadata will be enhanced (for an incremental cost) as it is processed in the OCLC workflow infrastructure to add linked open data references. Triaged, corrected and enhanced output from the workflow will deliver metadata for digital items to the Digital Public Library of America with the CC0 license required by DPLA. Enhanced linked open data records will be added to WorldCat.org with a CC-BY license and a copy will be delivered to the individual library for use in their local repository.

The workflow will provide capabilities for quality control review by collection curators. They will be able to ensure the consistency of metadata practice in their local repository, and to triage and correct the coverage of strongly recommended DPLA metadata elements:

- Date
- Place
- Subject
- Thumbnail

Workflow conceptual diagram



The workflow is built for the collection curators of Ohio's existing digital repositories, and is supported with the services provided by the WorldCat Digital Collection Gateway, WorldShare Metadata Services, and the DPLA aggregation system.

Metadata is harvested by the Gateway from any repository that supports OAI-PMH, the open archives initiative protocol for metadata harvesting. Within the Gateway, the collection curator uses the workflow to triage, correct, and enhance metadata and commit the metadata to the WorldCat data network. Using Collection Manager, the collection curator can feed their enhanced, normalized metadata back to their repository. The final link in the workflow is a single, statewide configuration of Collection Manager that draws the triage, corrected and enhanced metadata from the WorldCat data network and delivers this metadata to DPLA in DPLA MAP format.

What exists today in this workflow?

- **WorldCat Digital Collection Gateway** (http://www.oclc.org/en-US/digital-gateway.html) - The WorldCat Digital Collection Gateway is a self-service tool for a collection curator to get more visibility for their digital items by syndicating metadata with WorldCat. The Gateway works with any OAI-PMH compliant repository and provides tools for creating profiles for collections within a repository so they can added to WorldCat and re-synchronized over time to reflect changes (additions, edits, and deletions) in the local repository.

  As the collection curator sets up the profile for their repository and collections with the Gateway they can review metadata quality using data analysis tools that highlight metadata consistency, accuracy, and compliance with local metadata creation policies. In addition, the

Gateway provides a variety of tools that help the curator make profiled adjustments to metadata for sharing in a context outside the local repository such as field splitting, field merging, adding constant data, providing thumbnail links, and material type mapping.

The Gateway puts control of metadata management in the hands of the collection curator. For example, if the curator no longer wants to synchronize a particular collection or entire repository with WorldCat the Gateway gives them the ability to remove the collection or repository profile and remove all records from WorldCat that were present in that repository.

More than two thousand libraries (http://www.oclc.org/oaister/contributors.en.html) around the world using the Gateway today to get more visibility for their digital collections in WorldCat and OAIster. These libraries combined manage nearly 45 million records in WorldCat using the Gateway.

Additional Feature Required for this Project: OCLC will add additional profiling adjustments for metadata to comply with DPLA MAP format.

- **WorldShare Collection Manager** (http://www.oclc.org/en-US/worldshare-collection-manager.html) - WorldShare Collection Manager is primarily used today by technical services managers to configure MARC record exports from WorldCat to their library's public access catalog. While exports are typically configured to deliver MARC records there are a variety of export metadata formats available today including Qualified Dublin Core, MODs, Onyx, and MARC XML. Collection Manager is also configured for groups of libraries to export MARC records to other partner services.

Additional Features Required for this Project: OCLC will add the DPLA MAP format as an export option.

Proposed project timeline

- **Start-up and use of existing systems, Pilot phase 1** (April – June 2016) – Start-up and use of existing systems for initial aggregation and delivery to DPLA using Qualified Dublin Core metadata. Gap analysis for existing system on metadata enhancement needs & DPLA delivery needs.
- **Enhancing existing systems & review by libraries, Pilot phase 2** (July – September 2016) – OCLC enhances existing system based on gap analysis and libraries review the enhancements. Another aggregation and delivery to DPLA using DPLA MAP metadata.
- **Additional library on-boarding, Pilot phase 3** (October – December 2016) – On-board the balance of libraries from Ohio and establish regular delivery to DPLA.
- **"Go Live"** (January 2017) – Ohio libraries achieve a regular operation to harvest, enhance, normalize, aggregate, and deliver metadata from Ohio's digital repositories to DPLA.

Unique strengths

- Ohio's libraries each bring curatorial acumen and the unique collections from their special collections & archives. Additionally, the Ohio DPLA Service Hub provides the community organization including outreach, training, and socialization of metadata best practice.

- The Digital Public Library of America provides a focal point for aggregating metadata from statewide hubs and large, single data providers thus creating a rich environment for developing conventions for best practice in metadata sharing for digital objects.
- OCLC has a robust infrastructure for harvesting, enhancing, and disseminating metadata on behalf of individual libraries. OCLC's harvesting and enhancing capabilities for individual libraries were built from experience with the thousands of libraries participating in OAIster, and capabilities for disseminating metadata on behalf of individual libraries from the cooperatively built database of metadata, WorldCat.

What happens if Ohio libraries decide not to move forward using these services after the pilot?
Ohio libraries are under no obligation to continue to use the services OCLC is developing. In addition, should libraries no longer want to use these services the metadata they've shared can be removed from OCLC's systems, including WorldCat, either by using the tools in the Gateway or by request to have OCLC staff remove the metadata.

What's the difference between metadata the collection curator can export to their repository and the Ohio-wide aggregation delivered to DPLA?
A collection curator can export metadata from WorldCat to their own repository that includes all enhancements created either through community efforts or automated processes in WorldCat. This metadata is provided under OCLC's typical CCBy license.

The Ohio-wide aggregation is delivered to DPLA in the DPLA MAP format and will only include enhancements necessary to supply that required format. This metadata is provided to DPLA under a CC0 license.

Does an Ohio library have to be an OCLC member to participate in the pilot or subsequent service?
For the pilot project there is no prerequisite for OCLC membership. When the service goes live, then any organization that subscribes to the service becomes an OCLC member. Paying fees for OCLC services, ie., fiscally supporting the cooperative, is the definition of OCLC membership.

What is the business model/pricing for the pilot with OCLC?
The Business Model/Pricing for OCLC services for the pilot are yet to be determined. OCLC would like to work collaboratively with the pilot libraries and OhioLink to determine a workable model for all parties. Whatever the final business model, OCLC anticipates that costs for the first year of the pilot will be significantly reduced due to the need to build out the services and understand the overall library need.

Appendix E: The Tactical Strategy for Technical Infrastructure Working Group

Working Group Charge

The Tactical Strategy for Technical Infrastructure Working Group was charged to develop a technical strategy that will support Ohio's participation in DPLA; identify specific system requirements (software and hardware) needed to support the service center aggregation site; identify what harvesting standards will be supported based on DPLA requirements and what Ohio's cultural heritage organizations currently support; develop a budget estimate for the initial implementation and for 3-year operation of the aggregation site; identify technical barriers to contribution at the digitization hubs, Ohio Memory, and other major metadata content contributors.

Members of the Working Group

- Terry Reese (co-chair), The Ohio State University
- Nathan Tallman (co-chair), University of Cincinnati
- Meghan Frazer, OhioLINK
- Bryan Harris, Stark County District Library
- Marcus Ladd, Miami University
- Raymond Rozman, Cleveland Public Library
- Arjun Sabharwal, The University of Toledo
- Derek Zoladz, OhioNET