Abstract

After years of planning, the Digital Public Library of America (DPLA) launched in 2013. Institutions from around the United States contribute to the DPLA through regional “service hubs,” entities that aggregate digital collections metadata for harvest by the DPLA. The North Carolina Digital Heritage Center has been one of these service hubs since the end of 2013. This article describes the technological side of being a service hub for the DPLA, from choosing metadata requirements and reviewing software, to the workflow used each month when providing hundreds of metadata feeds for DPLA harvest. The authors hope it will be of interest to those pursuing metadata aggregation, whether for the DPLA or for other purposes.

Introduction

The Digital Public Library of America is a national digital library that aggregates digital collections metadata from around the United States, pulling in content from large institutions like the National Archives and Records Administration and HathiTrust as well as “service hubs” that compile records from small and large collections in a state or geographic area. This aggregated data is then available for search, and for reuse through the DPLA API. To date, over 7 million records are available—this number grows every month.

This article describes the technological side of working with the DPLA, specifically the process used by the North Carolina Digital Heritage Center, which is the DPLA’s service hub in North Carolina. It outlines what the Center discovered while choosing metadata requirements and reviewing software, as well as the workflow for providing hundreds of exposed OAI-PMH metadata feeds as a single stream for the DPLA. The authors hope it will be of interest to those pursuing metadata aggregation, whether for the DPLA or for other purposes.

DPLA Participation

Different states have created DPLA service hubs in different ways. North Carolina’s DPLA service hub is the North Carolina Digital Heritage Center, which was approached by the DPLA because the Center already served as a hub of sorts. Since 2010, the Center has provided digitization and digital publishing services to institutions throughout the state. Content from over 150 partners—from public libraries to local museums to large universities—is available on DigitalNC. In addition, North Carolina boasts a number of institutions hosting their own rich digital collections, and the Center has working relationships with many of these. Finally, the Center is supported by the State Library of North Carolina with funds from the Institute of Museum and Library Services under the provisions of the Library Services and Technology Act, and by the UNC-Chapel Hill University Library. Charged by the State Library with providing statewide services and with statewide connections, the Center was a likely fit as a service hub. After a number of conversations with the DPLA and interested data providers, the Center began working as a hub in September 2013.

The DPLA’s technical requirements for participation are straightforward: it asks that service hubs commit to being a “metadata aggregation point” for their community, and to “maintaining technology or technologies (such as OAI-PMH, Resource Sync, etc.) that allow for metadata to be shared with the DPLA on a regular, consistent basis.” To begin sharing that metadata, the Digital
Heritage Center had to evaluate, choose, and implement metadata schemas and encoding standards, workflows, and software that would fulfill this part of the hub agreement.

In broad strokes, the technical path to actively sharing North Carolina’s metadata with the DPLA looked like this:

1. Work with DPLA to determine metadata schema and “required” fields
2. Assess and set up feed aggregator
3. Review metadata feeds provided by data providers
4. Harvest feeds from data providers on a monthly basis, and expose that metadata to DPLA in a single stream.

Figure 1 presents an overview of the current technical workflow used by the Center to share metadata from data providers around the state to the DPLA as a result of the steps just mentioned. It is this workflow which will be discussed individually for the remainder of this paper.

Choosing a Metadata Schema and Required Fields

After committing to become a hub, one of the first steps involved deciding which fields the Digital Heritage Center would make available to the DPLA for harvest. The DPLA provided detailed help and engaged collaboration during this process. Its Metadata Application Profile (MAP) is based on the model used by Europeana, a digital library aggregating metadata from archives, libraries and other collections throughout Europe. The MAP outlines which fields are required and recommended, which are provided by the DPLA compared to the hub, and the metadata properties of each. It can be daunting for those unused to seeing metadata with this sort of notation. In short, the following fields are required from the hub:

<table>
<thead>
<tr>
<th>Field Name</th>
<th>Property</th>
<th>Requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>dc:title</td>
<td>Required per item</td>
</tr>
<tr>
<td>Rights</td>
<td>dc:rights</td>
<td>Required per item</td>
</tr>
<tr>
<td>Collection Title</td>
<td>dc:title</td>
<td>Required per set</td>
</tr>
<tr>
<td>Collection Description</td>
<td>dc:description</td>
<td>Required per set</td>
</tr>
</tbody>
</table>
The DPLA, however, accepts and encourages a larger group of fields, in part because its user interface exposes and leverages fields like date, subject, and geographic coverage. Further, if provided by hubs, the DPLA in turn exposes fuller metadata records through its API — a service that features prominently in its mission.

The Digital Heritage Center began evaluating which fields it would require from data providers in light of the DPLA’s requirements and recommendations. In making this decision, the Center also factored in its previous experience taking part in a local metadata aggregation project, NC ECHO. During that project, which took place in 2012, a handful of partners created a single search of digital collections metadata from across North Carolina. While describing that experience is outside of the scope of this paper, Center staff learned during that process that despite seeing high quality metadata, for each additional field included, aggregation, end-user display and data provider participation became exponentially more challenging. Balancing this lesson with the DPLA’s requirements and interface features, the Center decided to keep required and recommended fields to a minimum, and to make additional core fields optional.

Next came the evaluation of metadata schemas to determine what type of metadata the Center would accept from data providers, as well as what it would pass on to the DPLA. The DPLA can and does accept a variety of schemas — Simple Dublin Core, Qualified Dublin Core, and MODS are the most frequently used to date. In addition, information about what other hubs had already chosen and their crosswalks was made freely available during planning. Through informal conversations, reviewing collections online, and due to work with NC ECHO, the Center knew that many of the digital collections around North Carolina provide metadata in Dublin Core (DC). CONTENTdm, OCLC’s digital collections software, which supports Qualified DC, has a large market share in North Carolina. To help keep the barrier for participation low, Dublin Core was considered the best first choice for the schema the Center would ask institutions around the state to provide.

Digital Heritage Center staff then had to determine whether or not it would pass DC on to the DPLA, or transform DC to an alternative. There was some concern whether DC would be extensible enough if the Center ever decided to accept other formats. MODS, the Library of Congress’ Metadata Object Description Schema, allows more detailed encoding of metadata. The Library of Congress’ MODS site also provides a convenient crosswalk for MODS to DC. The Center was also able to take advantage of the work already done by the Boston Public Library, which was also providing metadata via MODS and whose crosswalk was shared with Center staff. For all of these reasons, it was decided that MODS instead of DC would be the schema used for the outgoing metadata stream from North Carolina.

Below is a table listing the crosswalk from Dublin Core to MODS, with all of the per-item metadata currently provided to the DPLA from North Carolina.

<table>
<thead>
<tr>
<th>DPLA Label</th>
<th>Dublin Core Element</th>
<th>MODS Element(s)</th>
<th>Requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contributor</td>
<td>contributor</td>
<td>&lt;name&gt;&lt;namePart&gt; where &lt;name&gt; also contains &lt;role&gt;&lt;roleTerm&gt;contributor&lt;/roleTerm&gt;&lt;role&gt;</td>
<td>Optional</td>
</tr>
<tr>
<td>Coverage</td>
<td>coverage</td>
<td>&lt;subject&gt;&lt;geographic&gt;</td>
<td>Recommended</td>
</tr>
<tr>
<td>Creator</td>
<td>creator</td>
<td>&lt;name&gt;&lt;namePart&gt; where &lt;name&gt; also contains &lt;role&gt;&lt;roleTerm&gt;creator&lt;/roleTerm&gt;&lt;role&gt;</td>
<td>Optional</td>
</tr>
<tr>
<td>Date</td>
<td>date</td>
<td>&lt;originInfo&gt;&lt;dateCreated keyDate=&quot;yes&quot;&gt;</td>
<td>Optional</td>
</tr>
</tbody>
</table>
Table 2: Dublin Core to MODS crosswalk for metadata contributed to the DPLA.

These few additional required fields are either automatically derived from the provided metadata feeds or hard-coded by the Center.

<table>
<thead>
<tr>
<th>DPLA Label</th>
<th>MODS Element(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference URL</td>
<td>&lt;location&gt;&lt;url usage=&quot;primary display&quot; access=&quot;object in context&quot;&gt;</td>
</tr>
<tr>
<td>Thumbnail URL</td>
<td>&lt;location&gt;&lt;url access=&quot;preview&quot;&gt;</td>
</tr>
<tr>
<td>Owning Institution</td>
<td>&lt;note type=&quot;ownership&quot;&gt;</td>
</tr>
</tbody>
</table>

Table 3: Additional MODS elements provided for each metadata record contributed to the DPLA.

Assessing, Choosing and Setting Up a Feed Aggregator

Identifying Requirements

While the Digital Heritage Center currently has its own collections OAI-PMH enabled and, in fact, shares its metadata through the aforementioned NC ECHO project and through WorldCat, for the DPLA the Center needed software that could ingest and output not only its own metadata, but also that of other data providers around the state, transforming it into a single MODS stream. Center staff began the process of assessing and choosing a feed aggregator by identifying functional requirements. The following were deemed required:

- Software at low or no cost
- Harvest OAI-PMH in multiple metadata formats
- Provide OAI-PMH in multiple metadata formats
- Harvest non-OAI-PMH data sets (static XML, Atom/RSS, etc.)
- Transform records (either as harvested or provided) with custom XSLT
- Group and manage data sets by data provider, for ease of administration
Schedule harvesting and record updates on a per-set or per-data-provider basis.

- Rename sets as they're harvested, to avoid confusion between similarly-named sets from different data providers.
- Schedule harvesting and record updates.

In addition, the Center targeted standalone software solutions for evaluation. Aggregation solutions exist for Drupal and Wordpress — two content management systems used by the Center — but were less attractive because the Center was in the process of moving between content management systems both for web and repository content at the time it became a hub. A standalone system could remain divorced from the Center’s own collections and web presence, and any changes those might undergo.

---

### Reviewing Aggregator Options

Finding available options was a complicated process; many content management systems can approximate an OAI-PMH harvester by allowing bulk import, then acting as an OAI provider. There are libraries for OAI-PMH harvesting alone available for a large number of programming languages, including PHP, Java and Perl, and a corresponding set of helpers for building an OAI-PMH provider on top of existing data. Some solutions are intended for use as a repository in their own right.

The list of requirements provided in the previous section was used to narrow down a collection of available standalone software solutions based on what could be understood from documentation available online. Staff also filtered the list of available solutions to ones using technologies the Center’s server environment could already support (Windows or Linux operating systems, written in PHP, Java, or Python). The Center evaluated the following software tools:

- Public Knowledge Project’s [Open Harvester Systems](#)
- [REPOX](#)
- [MOAI](#)

<table>
<thead>
<tr>
<th>Operating Systems</th>
<th>Language</th>
<th>Data Storage</th>
</tr>
</thead>
<tbody>
<tr>
<td>PKP-OHS</td>
<td>PHP</td>
<td>MySQL</td>
</tr>
<tr>
<td>REPOX</td>
<td>Java 6 (or higher)</td>
<td>Apache Derby, extensible</td>
</tr>
<tr>
<td>MOAI</td>
<td>Python 2.5 or 2.6</td>
<td>MySQL, SQLite, extensible</td>
</tr>
</tbody>
</table>

**Table 4: Aggregators reviewed**

---

### Choosing REPOX

It was determined that REPOX, originally developed by the Technical University of Lisbon to help institutions aggregate data for the [Europeana Digital Library](#), was the “best bet.” As a self-contained Java application, REPOX presented a good fit for existing server architecture and in-house expertise and offered functionality most in line with requirements.

<table>
<thead>
<tr>
<th>Requirements</th>
<th>REPOX 2.3.6</th>
<th>PKP-OHS 2.X</th>
<th>MOAI 2.0.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free and/or Open Source</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Harvest OAI-PMH in multiple metadata formats</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Provide OAI-PMH in multiple metadata formats</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Harvest non-OAI-PMH data sets (static XML, Atom/RSS, etc.)</td>
<td>✓ Partial, extensible</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>
Transform records (either as they're harvested or as they're provided) with custom XSLT or another crosswalking mechanism | ✓ | Partial, extensible[^3] | ✓

Group/manage data sets by data provider, for ease of administration | ✓ | ✓ | Unknown


Rename sets as they're harvested, to avoid confusion between similarly-named sets from different data providers | ✓ | No | ✓

| Table 5: Fulfillment of functional requirements by NCDHC-evaluated software. |

Feed Evaluation and Stylesheet Setup

As institutions around the state began providing their OAI-PMH feed addresses, it became clear that evaluating metadata for minimum requirements using the raw OAI-PMH output would be tedious and time consuming. Digital Heritage Center staff were interested in a way to quickly assess (1) the presence of required, recommended and optional fields and (2) the baseline quality of different metadata collections. In response, staff developed a “DC Mapping Checker” tool (See Figure 2). Upon inputting an OAI feed address and selecting an available set, the tool parses field data into columns for easy perusal. This helped staff determine if institutions had mapped fields correctly (for example, was dc:type incorrectly showing dc:format terms, a common discovery), if any anomalies like inappropriate special characters or markup were present in the metadata, and whether or not required fields were blank. Metadata quality beyond mapping accuracy and absence or presence of required fields was not rigorously reviewed. In some instances, basic observations were noted (such as missing optional fields) but only for future discussion and not as a barrier to initial ingest. Missing required metadata or anomalies were reported to the different data providers for evaluation and correction.

![Figure 2: Screenshot of DC Mapping Checker created for metadata review.](larger image.)

Once an institution had at least one feed that fulfilled all requirements, Center staff created appropriate XSLT stylesheet(s). To help cut down on the number of stylesheets that would need to be created, a “two-tier” stylesheet system for transforming incoming DC to MODS was devised:

[1] A base stylesheet with transforms needed for all feeds OR [2] Base stylesheet(s) with transforms for all
feeds from a specific content management system

AND


For example, the Center's own feed, provided by CONTENTdm, would make use of the base CONTENTdm stylesheet [2] as part of its own “ncdhc” file [3]. Figure 3 shows this concept graphically.

![Figure 3: The two-tier stylesheet system, with examples from different types of data providers.](image)

Base stylesheets [1] simply map the starting data format to the end-use data format (MODS). A simple DC (oai_dc) to MODS base stylesheet would, for each simple DC field, designate which MODS field to output — i.e., "when I come across a 'date' field, output a MODS 'dateCreated' field with the corresponding data."

Base stylesheets for specific content management systems [2] take the base stylesheet one step further, attempting to account for any platform-specific idiosyncrasies in output. For example, in CONTENTdm OAI-PMH feeds, multiple values for the same field are concatenated into one XML entry with semicolons in-between. The base oai_dc CONTENTdm stylesheet would say, "when I come across a 'date' field, separate semicolon-delimited values and output a MODS 'dateCreated' field for each."

A data provider's specific stylesheet [3] is used to insert data into all of that data provider's MODS records. This covers data that either doesn't exist in the original metadata records for good reason, or is stored in the original metadata record in a way that is particular to that provider. A good example of this is the DPLA-required "Owning Institution" field, as most institutions either do not provide this as a part of their records, or provide it in a field specific to the nature of their relationship to the physical materials. Once the appropriate stylesheets were created, they were added to REPOX.

Using REPOX

REPOX has an in-browser graphical user interface. The view is organized by “Data Providers” (OAI-PMH feeds), under each of which is listed the data set(s) that have been added for aggregation. (See Figure 4.) Other details about each set are included: a user-defined set name, which schemas are applied to the set, how that set is ingested, the date and time of last ingest, the number of ingested records, and an icon that shows at-a-glance if the last ingest was successful. If ingests are scheduled, this date and time also appears.
Data providers are added first, followed by each set that the provider would like harvested. To do this, the OAI-PMH URL is included and the appropriate set(s) chosen. Each set is given a unique Record Set name and description. Finally, the appropriate stylesheet is applied.
Once all of the data provider’s feeds have been added, Digital Heritage Center staff ingest the records. REPOX harvests the DC metadata and transforms it using the assigned MODS stylesheet. The MODS output is reviewed when the ingest is complete, to make sure all of the Dublin Core fields are represented correctly.

About a week before the DPLA wants to harvest the single stream of metadata from the North Carolina service hub, Center staff go through and re-ingest each set to make sure the hub provides the most up-to-date metadata. Any new collections requested by the data providers are also added. When the DPLA has finished its harvest, Center staff spot check the results on http://dp.la to make sure newly harvested content shows up correctly. This process occurs regularly on a monthly basis.

### Lessons Learned and Next Steps

The Digital Heritage Center’s success as a service hub to date is due in part because its feed is built expressly for and supplied only to the DPLA; it isn’t a multi-purpose or publicly available feed at this time. Having a single consumer in mind made it easier to scope requirements. Also contributing to its success as a hub is the high quality and uniform nature of the largest digital collections in North Carolina. The decision was made to treat being a hub less like a professional sports team and more like intramurals: metadata requirements were set relatively low to encourage participation by all types of institutions. The idea was that representation should be as robust across the state as possible, leaving conversations about metadata enhancement for later dates.

Requiring few fields was a good decision for encouraging participation, however going forward it remains to be seen how much data providers would be interested in improving their metadata in more substantial ways (normalizing vocabulary terms across sets or institutions, clarifying rights statements, adding location data, etc.). Sharing metadata side-by-side with other institutions can lead to a bit of metadata anxiety; several institutions have held back sharing their “legacy” collections until they can make sure it’s up to their current standards. Some have even used DPLA participation as an impetus for correcting metadata, like remapping fields to other Dublin Core elements or updating rights statements. In the end, MODS and Dublin Core were good choices because they are extensible and popular in North Carolina. States interested in accommodating a broader variety of metadata schemas or in requiring a larger number of fields would have to invest more time in creating stylesheets, and may also face reticence in sharing as dirty data comes to light.

REPOX has worked well for the Center. For those who have a working knowledge of OAI-PMH, simply adding and managing feed ingest is simple. Creating stylesheets requires expertise, however it should be noted that REPOX includes an XSLT “mapper” tool called XMApper, that purports to allow a drag-and-drop capability for creating stylesheets. Added functionality facilitating batch ingests and edits to sets would be helpful; right now most activities like scheduling harvests have to be done on a set-by-set basis. Finally, REPOX appears to only add new and remove deleted records on harvest. If a data provider edits an existing metadata record in their content management system and exposes it in the OAI-PMH feed, REPOX does not seem to pick up those changes. It’s necessary to dump existing data and do a complete reharvest each month. The Center has approached REPOX developers about implementing some of these changes.

As a hub with 12 data providers, representing over 150 institutions, the Center is now both maintaining its current responsibilities as well as beginning to work with those in North Carolina who may not have an OAI-PMH feed exposing their metadata. In several cases these are large institutions with homegrown repositories and the technical expertise to create feeds from scratch. In other cases Center staff are talking with smaller institutions about the possibility of adding their metadata in other ways: REPOX can harvest data from local file systems or Z39.50.

The Center is looking forward to exploring some of the promise the DPLA offers with big data, both using North Carolina content as well as content from the DPLA as a whole. It is also looking to the DPLA to help drive conversations about metadata improvement and enhancement, in situations that seem appropriate and worthwhile for data reuse and consistency. In some ways, providing metadata as a hub may only be the beginning.

### Notes

1. See DPLA, [Become a Hub](#).

2. May require a plugin; plugin support for OAI-PMH and OAI PMH Static Repository XML is provided.

3. PKP-OHS natively supports crosswalking from arbitrary formats into Dublin Core; other crosswalks are possible with a
4 The REPOX UI has options that suggest that scheduling record updates is possible, but it doesn't seem to work in the Center's installation. Deleting and re-ingesting a data set refreshes the data (see 'lessons learned' section).

5 A command-line harvesting utility is provided and may be run via cron job.

About the Authors

Lisa Gregory is the Digital Projects Librarian for the North Carolina Digital Heritage Center, where she oversees digitization, metadata creation, and online publication of original materials contributed by cultural heritage institutions throughout North Carolina. She also serves as the Center's liaison to the Digital Public Library of America. Prior to the Digital Heritage Center, Gregory worked for the State Library of North Carolina's Digital Information Management Program.

Stephanie Williams is the Digital Project Programmer for the North Carolina Digital Heritage Center, housed at the University of North Carolina at Chapel Hill. She maintains the statewide digitization program's web site and digital object repositories and builds tools to aid digitization, description, and public service workflows. She completed an MSLS from the School of Information and Library Science at the University of North Carolina at Chapel Hill in 2008, with a focus on digital collections.