DIGITAL PUBLIC LIBRARY OF AMERICA

An introduction to the DPLA metadata model

Note: An appendix with definitions for some of the terms used in this document can be found on page 4.

A. General Introduction

The DPLA metadata application profile (DPLA MAP) is stored, serialized, and made available through our API in JSON-LD. It also powers our web portal. The MAP is based on the Europeana Data Model (EDM). EDM is extensible, and while it uses locally developed properties (sometimes referred to as "elements"), it also employs properties from other namespaces, like Dublin Core and the Resource Description Framework (RDF). OAI-ORE (Open Archives Initiative Object Reuse and Exchange). It only made sense, then, for DPLA to adopt this model.

Version three of the DPLA MAP (the current iteration) was developed in early 2013 by metadata specialists, in collaboration with Europeana staff and public data specialists who provided input during an open review period in late 2012. It is meant to grow to support the sharing of records in a linked data format. For example, if we store the URI (http://id.loc.gov/authorities/subjects/sh97009149) for the subject value "snowmen" along with or instead of the string value it is always represented appropriately (no typos!) and computers can easily identify and validate it.

B. DPLA MAP structure

The DPLA MAP is constructed with classes and properties, which aid in structuring data hierarchy and manage information about data values. The DPLA MAP enables the integration of data created and shared by our Hubs in a variety of metadata standards. This allows us to appropriately represent and share that data, and to enrich it for greater discovery and access. The "SourceResource" class contains many of the properties that hold descriptive metadata (title, date, format, etc.). This is the data you see when you search the DPLA web portal. Most of the properties in this class are based on Dublin Core Metadata Initiative vocabularies. For this reason, when we work with partners to "crosswalk" their elements to the DPLA MAP properties, we like to begin with the SourceResource class.

The SourceResource class links to other classes in the DPLA MAP that (1) store information about the digital version of the object or objects (WebResource), (2) allow for the enrichment of certain fields (Place and TimePeriod), and (3) package all of this information together (Aggregation). In addition, the (4) Collection class allows us to gather information about locally defined sets or collections to which the SourceResource belongs. The Aggregation class stores important information about our direct collaborators (we call them Hubs), the actual providers

of the data (Contributing Institutions), the location of the local record and thumbnail, and the original record.

C. Data feeds and harvesting

Because of the DPLA MAP's structure, it can interface with nearly any metadata standard. To date, simple and qualified Dublin Core, MODS, METS-wrapped MODS, MARC XML, and several local metadata application profiles have been crosswalked to the MAP. We are confident that other standards and metadata implementations, such as VRA Core, CDWA, and CIDOC from the museums domain, for example, are compatible. And, if a metadata standard has been mapped to the Europeana data model, then it can be mapped to the DPLA MAP.

We receive data from our Hubs in a variety of structures, as well. To date, OAI-PMH is the type of feed most utilized by our Hubs. In addition, many institutions are creating locally defined APIs that can provide their data in multiple formats. In some cases, we receive data in an OAI-PMH feed in the simple or qualified Dublin Core standard or even in MODS. In other cases, we receive data via an API in MODS. In a few cases, we download static batches of data from a provider. The data may come to us as tab-delimited text files, XML files, or another format. While this is not our preferred method for receiving data, it is one that can be employed if no other option is available.

D. Where to start?

Prospective partners often ask how to test their standard against the DPLA MAP. It's hard to answer this question definitively, but we suggest that the following points be considered.

1. Make sure your data is clean and that elements and properties are implemented according to the rules of the metadata standard you are using.
2. Consider how your data will work on a global scale next to the data of thousands of other institutions in DPLA. Will John Brown in Australia understand that your geographic location "Washington" is different than the State of Washington or Washington County in Wisconsin?
3. Are your descriptions useful to users that may not be familiar with your subject area/field/institution type?
4. If the implementation of metadata standards varies among your collections, the fields that contain the conflicting information may not be included in the crosswalk. This means that the data for that field will not appear in DPLA. For example, if one collection about books uses a "contributor" field to describe secondary authors (Smith, Jane), and another collection about photographs uses the "contributor" field to describe the institution that provided access to the photographs (Acme Museum of Art), that field cannot be crosswalked to the DPLA MAP. This type of issue must be identified and resolved at the Hub and/or Contributing Institution before the data in that property can be shared with the DPLA.
5. *For Service Hubs (the institutions that aggregate data from other contributors)*: The metadata for all collaborators at a *Service Hub* must be aggregated and shared with

DPLA through a single feed. For example, the North Carolina Digital Heritage Center Hub brings together all of the metadata from their approximately 150 institutions and makes it available to DPLA in one OAI-PMH feed in the MODS metadata standard. The work of crosswalking those 150 institutions occurs at the Hub level. This approach provides an on-ramp for smaller and under-funded institutions and ensures greater sustainability for the contributing institutions, the Hubs, and DPLA.

E. The Future

Work is already underway to upgrade to the next version of the DPLA MAP. Additionally, our plans include an application that prospective partners can use to easily test their metadata against the DPLA MAP. As information about these activities becomes available, it will be shared by DPLA staff.

For further questions about metadata or DPLA data feeds, please contact Amy Rudersdorf, Assistant Director for Content, at amy@dp.la.

Appendix: Definitions

Aggregation is the gathering of metadata records from multiple data providers and then, in the case of DPLA, making those records available for gathering by others.[4]

API, an abbreviation of Application Program Interface, is a set of routines, protocols, and tools for building software applications. APIs specify how software components should interact.[7]

Classes are groupings of related properties, e.g., SourceResource in the DPLA MAP.

Contributing Institutions provide data to a Hub, which then aggregates data from multiple Contributing Institutions and shares it via a single feed or otherwise exposes it to DPLA.

Crosswalks are tables that map the relationships and equivalencies between two or more metadata formats. Crosswalks support the ability of search engines to search effectively across heterogeneous databases, i.e., crosswalks help promote interoperability.[2]

EDM, or the Europeana Data Model*, is the metadata profile upon which the DPLA MAP is based. Many of the classes and properties in the DPLA MAP are taken directly from EDM.[3]

A *feed* is a mechanism for users to receive updated data from data sources (think "news feed"). It is commonly used by real-time applications in point-to-point settings as well as on the Web.[8]

Harvesting refers to the gathering together of metadata from a number of distributed repositories into a combined data store.[4]

Hubs are institutions that share data directly with DPLA.

A *metadata application profile* is a set of metadata elements, policies, and guidelines defined for a particular application.[2]

Namespaces are qualifiers added to an XML tag to ensure uniqueness among XML elements.[4]

OAI-ORE (Open Archives Initiative Object Reuse and Exchange) defines standards for the description and exchange of aggregations of digital objects.[5]

OAI-PMH is the Open Archives Initiative Protocol for Metadata Harvesting, a low-barrier mechanism for repository interoperability. Many library repository or digital collections systems use this communication method to share sets of records with the DPLA.[4]

A *Property* is an element that expresses the relationship between two digital objects. Properties can be seen as the attributes or characteristics of a digital object.[3]

Service Hubs† aggregate metadata that resolves to digital objects (online texts, photographs, manuscript material, art work, etc.) from local libraries and other cultural heritage institutions.[1]

URI is the acronym for Universal Resource Identifier. URIs are unique Web identifiers.[6]

Definitions derived from those available at:

1. http://dp.la/info/get-involved/partnerships/
2. http://dublincore.org/documents/2001/04/12/usageguide/glossary.shtml
3. http://pro.europeana.eu/edm-documentation
4. http://www.oaforum.org/tutorial/english/page6.htm#section2
5. http://www.openarchives.org/ore/
6. http://www.w3.org/2003/glossary/
7. http://www.webopedia.com/TERM/A/API.html
8. https://www.wikipedia.org/

---

* http://pro.europeana.eu/edm-documentation

† http://dp.la/info/get-involved/partnerships/