An introduction to the DPLA Metadata Model

A. General Introduction

The DPLA metadata application profile (DPLA MAP) is the basis for our application programming interface (API) data structure, built in JSON-LD, which in turn is what powers our web portal. The MAP is based on the Europeana Data Model (EDM). EDM primarily uses properties (sometimes referred to as "elements") from other namespaces, like Dublin Core and the Resource Description Framework (RDF) in addition to a smaller number of unique, locally developed properties. EDM was developed specifically for the aggregation of metadata from diverse digital objects across multiple institutions in the European Union. It only made sense, then, for DPLA to adopt this model.

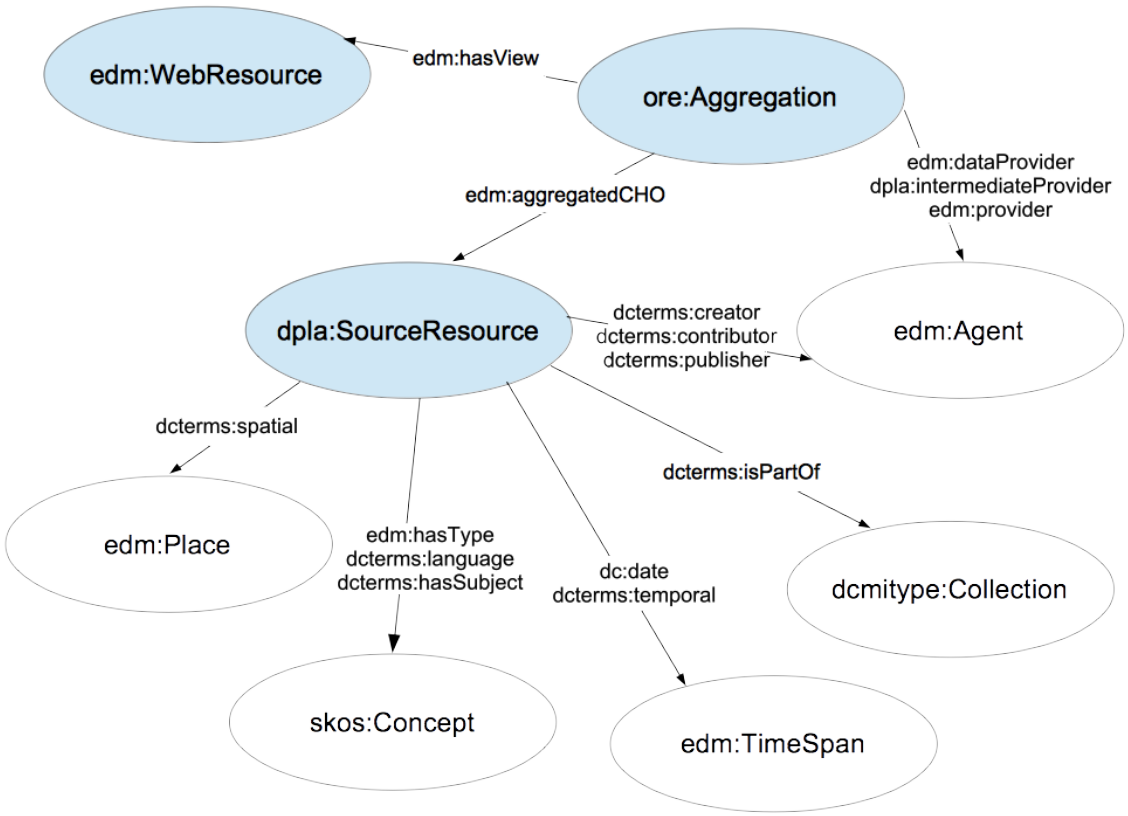
The DPLA MAP was drafted in 2012 and input was gathered during an open, public review period late in that year from metadata specialists. DPLA MAP v3.0 created in response to that review and released in early 2013. It was given a minor update in July 2014 with the release of DPLA MAP v3.1. The current version, DPLA MAP v4.0, was released in March 2015.

Note: An appendix with definitions for some of the terms used in this document can be found on page 6.

B. Overview of DPLA Map Structure

The DPLA MAP enables the integration of metadata 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. DPLA's MAP is an application profile, or a set of metadata elements, taken from multiple schemas for a particular local use. It is also a semantic metadata model, or an abstract structure that describes the relationships between different types of data about the same thing. This means it is more robust and abstract than a metadata schema like Dublin Core or MODS in that it describes entities and the relationships between them. Entities are things like the resource being described, the record describing it, the person who created it, and the place it is located. By describing these entities independently and then creating relationships between them, we can create a more dynamic and efficient metadata store.

Figure 1. DPLA MAP 4.0 Class Diagram



The description of entities is contained in classes. Each of these classes contains a list of possible properties based on existing metadata element sets such as Dublin Core and the Europeana Data Model, as well as some locally defined elements. By reusing metadata elements from widely adopted schemas where appropriate, we can increase our interoperability and remain loosely coupled with other systems, with the additional benefit of not reinventing the wheel.

The "SourceResource" class contains many of the properties that hold descriptive metadata (title, date, format, etc.) about the original resource, or "source resource." This is where most of the data you see when you search the DPLA web portal is stored. For this reason, when we work with partners to match, or "crosswalk," their elements to the DPLA MAP properties, we like to begin with the SourceResource class. Most of the properties in this class are based on Dublin Core.

The SourceResource class links to other classes in the DPLA MAP that (1) store information about the digital version of the original resource (WebResource), (2) allow for the enhanced description of certain fields (Agent, Concept, Place and TimeSpan), and (3) package all of this information together (Aggregation). 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 a link to a stored version of the original metadata record we received from the Hub. In addition, the Collection class allows us to gather information about locally defined sets or collections to which the original resource belongs. Specific details about these classes and their properties can be in found in Appendix B: DPLA MAP Structure.

B. Data Standards

Since DPLA MAP v4.0 is an abstract metadata model, it therefore can be expressed in any standard computer language. Examples in the DPLA MAPv4.0 description are given in JSON-LD, which is the language used for our API. Natively, DPLA MAPv4.0 metadata is stored as RDF triples.

These two standards were chosen for their ability to handle Linked Open Data (LOD). There are many benefits to LOD including synchronization with authority records. For example, if we store the URI (http://id.loc.gov/authorities/subjects/sh97009149) instead of the string value "snowman" in the Subject property, we can always go back to the URI and retrieve the appropriate representation (no typos!), or find it in another language (*schneemann*), or differentiate from a film of the same name (*The Snowman* animated film from 1982). A fuller description of the benefits of LOD is out of scope for this document, however, if you are interested a good introduction is found in the video *Linked Open Data – What Is It*? found at: http://vimeo.com/36752317.

C. Enrichment

"Enrichment" refers the process DPLA uses to enhance original records from partners with additional data, typically in the form of standardized versions of names and places as well as URIs to LOD vocabularies such as GeoNames or the Virtual Identity Authority File (VIAF). DPLA has developed services to check these authorities for the values in specific partner-supplied metadata fields and record any matches. DPLA currently performs such enrichments only on place names, however, partner-supplied URIs for subject-headings or name authorities are stored within MAP properties. DPLA also performs enrichments that remove extraneous punctuation and whitespaces, normalize date formats to a standard (yyyy-mm-dd), and other clean-up tasks.

When enrichment results in the addition of a URI or other data, the original partner-supplied value is stored alongside the enriched value. This means that if DPLA's enrichment should happen to result in an error, the original data is still retained in the DPLA record. Additionally, the original record provided by the partner is always retained within DPLA's infrastructure. We copy and host the original file and provide a URI back to it from our enriched record.

D. Data feeds and harvesting

Because of the DPLA MAP's robust and flexible 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 DPLA MAP. We are confident that other standards and metadata implementations, such as VRA Core, CDWA, and CIDOC CRM, for example, are compatible. And, because of the underlying alignment with the Europeana Data Model, we know that if a metadata standard has been mapped to EDM, 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. We currently have Hubs that use OAI to supply us with data in the simple or qualified Dublin Core standard or even in MODS. In addition, many institutions are creating locally defined APIs that can provide their data in multiple formats. 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.

E. Access and the DPLA API

DPLA provides access to its transformed aggregated metadata three ways: a single download, a searchable portal, and a web-based application programming interface (API). An extensive overview of the API is available at http://dp.la/info/developers/codex/, however, it should be noted that the data that the API structure is based on the previous version of the DPLA MAP, v3.1. There are multiple reasons for this, but the most important is providing uninterrupted support for applications built around the existing v3.1 API. This means that metadata stored internally within DPLA as MAP v4.0 is mapped to v3.1 for publication through the API. An additional API endpoint that publishes v4.0 metadata will be developed in the near future.

F. 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 as error-free as possible and that elements and properties are consistently implemented according to the rules of the metadata standard you are using.
2. Consider how your data will work in a global context 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 when your metadata is mapped to DPLA's metadata structure. This means that the data for those fields will not be harvested by or 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 field 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 (NCDHC) 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' individual metadata implementation to MODS occurs at the Hub level by NCDHC. This approach provides an on-ramp for smaller and under-funded institutions and ensures greater sustainability for the contributing institutions, the Hubs, and DPLA.

For further questions about metadata or data feeds, please contact Gretchen Gueguen, Data Services Coordinator, gretchen@dp.la.

Appendix A: 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. [6]

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

Contributing Institutions are the institutions at which the original resources are held. They provide metadata to a Hub, which then aggregates metadata from multiple contributing institutions and shares it all via a single feed or otherwise exposes it to DPLA.

Crosswalks are tables that map the relationships and equivalencies between two or more metadata schemas. Crosswalks--or metadata mapping--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 data ***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 World Wide Web. [7]

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.

JSON-LD a lightweight Linked Data format. It is easy for humans to read and write.[9]

Linked Open Data (LOD) The term Linked Data refers to a set of best practices for publishing and connecting structured data on the Web. [8]

Mapping (see *Crosswalks*)

Metadata Application Profile is a defined set of metadata properties that combines selected elements from several standard schemas along with locally defined ones. Policies, and guidelines are also defined for a particular application. [2]

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

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 resources. Properties can be seen as the attributes or characteristics of a resource.[3]

Resource Description Format, or ***RDF*** is a standard model for data interchange on the Web…. RDF extends the linking structure of the Web to use URIs to name the relationship between things as well as the two ends of the link (this is usually referred to as a "triple").[10]

Service Hubs aggregate metadata from multiple contributing institutions (libraries and other cultural heritage institutions) that resolves to digital resources `(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 strings that uniquely identify things on the Web.[5]

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.w3.org/2003/glossary/
6. http://www.webopedia.com/TERM/A/API.html
7. https://www.wikipedia.org/
8. http://linkeddata.org/faq
9. http://json-ld.org
10. http://www.w3.org/RDF/

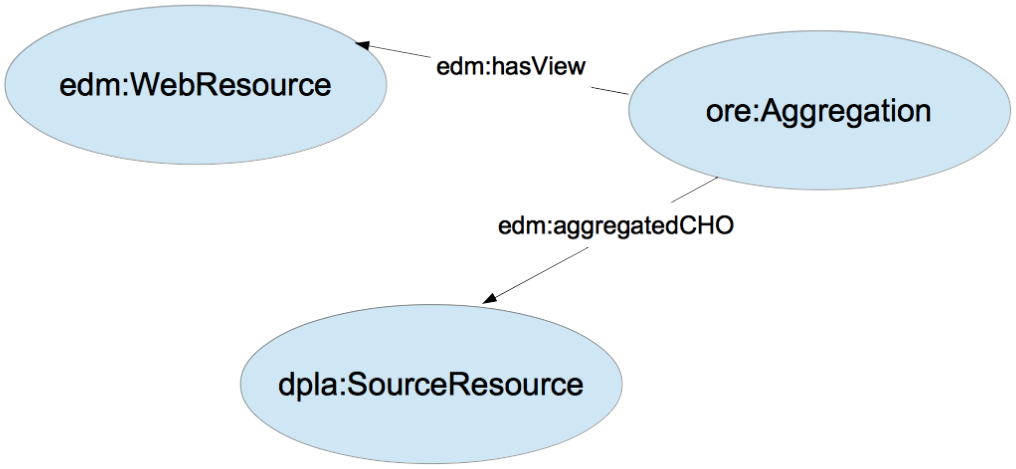
Appendix B: DPLA MAP Structure

The MAP

Core classes

The DPLA MAP model consists of an aggregation of classes. The following are present for every DPLA record.

Figure 2. DPLA MAP diagram showing only the core classes.



SourceResource

The dpla:SourceResource class contains the main descriptive information about the original resource (the book, photograph, painting, digital file, etc). Therefore, it contains most of the information that will be mapped from an original record and searched in the DPLA portal or API.

Partner institutions are encouraged to provide as much of this information as they can, but are asked to provide, at a minimum a title and a rights statement.

The properties in the sourceResource class are:

DPLA Label	Equivalent Element	Requirement
Alternative title	dcterms:alternative	Optional
Collection	dcterms:isPartOf	Recommended
Contributor	dcterms:contributor	Optional
Creator	dcterms:creator	Recommended
Date	dc:date	Recommended
Description	dcterms:description	Recommended
Extent	dcterms:extent	Optional
Format	dc:format	Recommended
Genre	edm:hasType	Optional
Identifier	dcterms:identifier	Optional
Language	dcterms:language	Recommended
Place	dcterms:spatial	Recommended
Publisher	dc:publisher	Recommended
Relation	dc:relation	Optional
Replaced by	dcterms:isReplacedBy	Optional
Replaces	dcterms:replaces	Optional
Rights	dc:rights	Required
Rights Holder	dcterms:rightsholder	Optional
Subject	dcterms:subject	Optional
Temporal Coverage	dcterms:temporal	Optional
Title	dcterms:title	Required
Type	dcterms:type	Recommended

Ore:WebResource

This class contains information about the digital resource that is representing the original source material. In the case of a photograph, the ore:WebResource class could record whether or not the digital representation was a jpeg or a tiff. This class allows the information about the original resource that a user might be interested in to be separate from the information about the digital representation(s) that the system may need.

DPLA Label	Equivalent Element	Requirement
File Format	dc:format	Optional
Rights	dc.rights	Optional
Standardized Rights Statement	edm:rights	Required

Ore:Aggregation

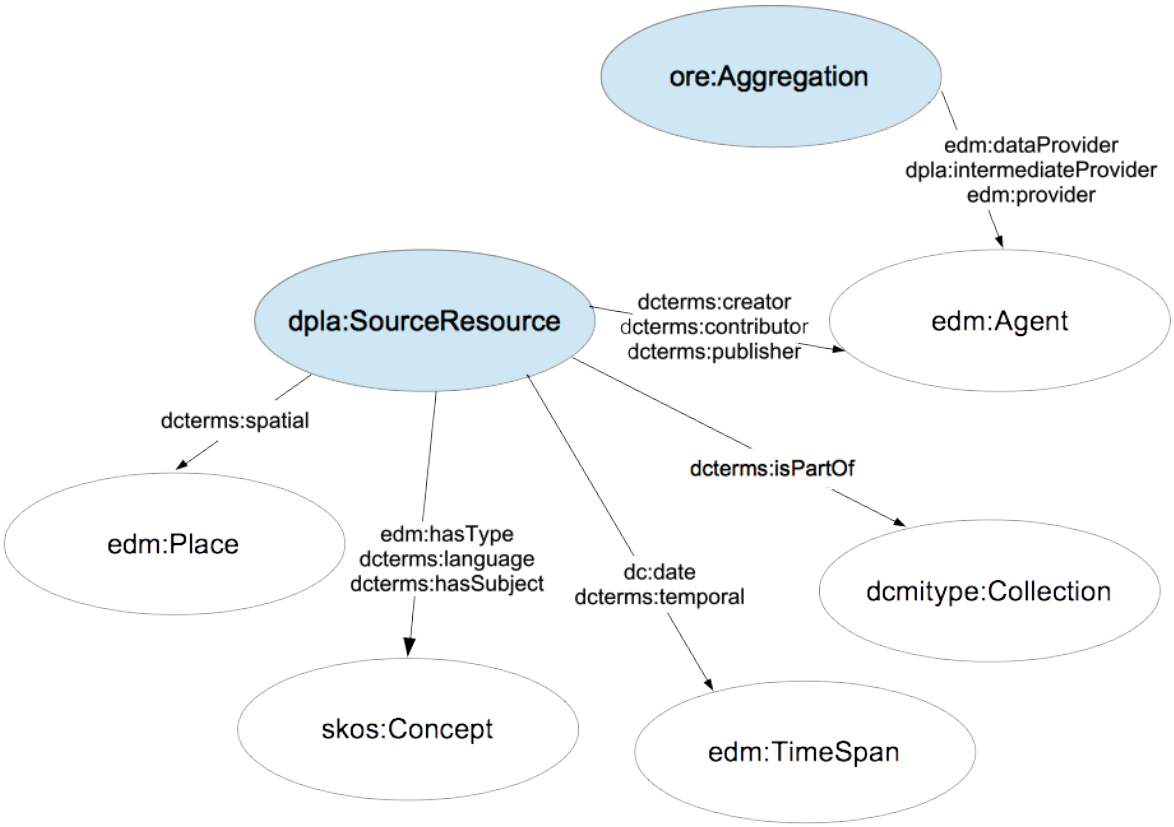
Finally, the aggregation class brings together the core classes, along with the context classes that will be described later, as a single object within the DPLA data store.

DPLA Label	Equivalent Element	Requirement
Aggregated SR	edm:aggregatedCHO	Required
Data Provider	edm:dataProvider	Required
Digital Resource Original Record	dpla:originalRecord	Required
Has View	edm:hasView	Optional
Intermediate Provider	dpla:intermediateProvider	Optional
Is Shown at	edm:isShownAt	Required
Object	edm:object	Optional
Preview	edm:preview	Required
Provider	edm:provider	Required
Standardized Rights Statement	edm:rights	Required

Some of the information that populates the ore:Aggregation class will be created by DPLA upon ingest (Aggregated SR, Digital Resource Original Record). Properties that are supplied by Data Providers include URLs to digital representations of the original resource (edm:preview) and the Data Provider's name (edm:dataProvider).

Context Classes

In addition to the SourceResource record, the DPLA MAP contains classes to hold further contextual information referenced in the description such as agents, collections, concepts, places, and time spans.



An overview of each context class follows.

Agent

An agent is any person or organization referenced in the descriptive record, particularly those who engage in some activity related to the record (the creator, contributor, or publisher). Partners may supply the URI to an identity in this or another authority file exposed as an LOD vocabulary.

The Agent class is very simple in structure, containing only the Preferred Name as found in an authority file, a Provided Label as supplied by the Data Provider, and the URI stored in the Close or Exact Match properties. The Close Match property (skos:closeMatch) can be used to designate the URI of a very similar term when an exact match can't be found (such as in the case of a possible typo).

DPLA Label	Equivalent Element	Requirement
Preferred Name	skos:prefLabel	Optional
Provided Label	dpla:providedLabel	Optional

Exact Match	skos:exactMatch	Optional
Close Match	skos:closeMatch	Optional

Collection

The collection class contains information about any collection of DPLA objects. Typically this information can be derived from the OAI data set the original resource belongs to or from any appropriate field in the metadata record. An individual object may point to more than one collection class (i.e. may be a member of more than one collection). The collection elements are as follows:

DPLA Label	Equivalent	Requirement
Collection Title	dcterms:title	Optional
Collection Decription	dcterms:description	Optional

Concept

The concept class is used to define concepts or subject headings as defined by authoritative controlled vocabularies. As with the agent class, it can handle references to LOD vocabularies as exact or close matches. It also stores the original partner-supplied value as Provided Label. DPLA does not currently attempt to match partner-supplied subjects to controlled vocabularies, but will store any information about LOD references to concepts directly supplied by partners. The Concept class can also accommodate notes as well as the name of the term's originating vocabulary (i.e. LCSH or TGM).

DPLA Label	Equivalent Element	Requirement
Preferred Name	skos:prefLabel	Optional
Provided Label	dpla:providedLabel	Optional
Note	skos:note	Optional
Scheme	skos:inScheme	Optional
Exact Match	skos:exactMatch	Optional
Close Match	skos:closeMatch	Optional

Place

Authoritative geographic information relating to the source resource is recorded in the Place class. DPLA provides and populates this class based on supplied partner data by searching relevant geographic LOD vocabularies for matches. When a match is found DPLA can add the URI as well as other geographic features like coordinates and country codes. The data in this class, particularly the longitude and latitude properties, are used in DPLA's online geographic browse feature (http://dp.la/map). We therefore encourage partners to supply as much geographic information as possible so that an exact match can be made. The Place class includes

DPLA Label	Equivalent Element	Requirement
Preferred Name	skos:prefLabel	Optional
Provided Label	dpla:providedLabel	Optional
Latitude	wgs84_pos:lat	Optional
Longitude	wgs84_pos:long	Optional
Altitude	wgs84_pos:alt	Optional
Geometry	geojson:geometry	Optional
Parent Feature	gn:parentFeature	Optional
Country Code	gn:countryCode	Optional
Exact Match	skos:exactMatch	Optional
Close Match	skos:closeMatch	Optional

TimeSpan

The TimeSpan class is used to record information regarding the Date or Temporal Coverage properties in a normalized format (yyyy-mm-dd). Date information may be supplied in multiple formats, but DPLA will normalize and store the date in the Begin and End properties of this class. If a single exact date is provided, it will populate both the Begin and End properties.

DPLA Label	Equivalent Element	Requirement
Original source date	skos:prefLabel	Optional
Begin	edm:begin	Optional
End	edm:end	Optional

Appendix C: Suggested Crosswalks to dpla:SourceResource, edm:WebResource, and ore:Aggregation

Simple and Qualified Dublin Core

dpla:SourceResource

DPLA LABEL	DPLA PROPERTY	QDC ELEMENT	DC ELEMENT
Alternative Title	dcterms:alternative	dcterms:alternative	*No suggested mapping**
Collection	dcterms:isPartOf	dcterms:isPartOf	*No suggested mapping**
Contributor	dcterms:contributor	dcterms:contributor	dc:contributor
Creator	dcterms:creator	dcterms:creator	dc:creator
Date	dc:date	dcterms:created	dc:date
Description	dcterms:description	dcterms:description	dc:description
Extent	dcterms:extent	dcterms:extent	*No suggested mapping**
Format	dc:format	dcterms:medium	dc:format
Genre	edm:hasType	*No suggested mapping*	*No suggested mapping**
Identifier	dcterms:identifier	dcterms:identifier	dc:identifier
Language	dcterms:language	dcterms:language	dc:language
Place	dcterms:spatial	dcterms:spation	dc:coverage
Publisher	dcterms:publisher	dcterms:publisher	dc:publisher
Relation	dc:relation	dcterms:relation	dc:relation
ReplacedBy	dpla:replacedBy	dcterms:replacedBy	*No suggested mapping**
Replaces	dpla:replaces	dcterms:replaces	*No suggested mapping**
Rights	dc:rights	dcterms:rights	dc:rights
Rights Holder	dcterms:rights	Dcterms:rightsholder	*No suggested mapping**
Subject	dcterms:subject	dcterms:subject	dc:subject
Temporal Coverage	dcterms:temporal	dcterms:temporal	*dc:coverage*
Title	dcterms:title	dcterms:title	dc:title
Type	dcterms:type	dcterms:type	dc:type

*By using the phrase "no suggested mapping," DPLA is not stating that a mapping to this element is impossible, but that we have no preferred mapping suggestion.

edm:WebResource

DPLA LABEL	DPLA PROPERTY	QDC ELEMENT	DC ELEMENT
File Format	dc:format	dcterms:format	*No suggested mapping**
Rights	dc:rights	*No suggested mapping*	*No suggested mapping**
Standardized Rights Statement	edm:rights	*No suggested mapping*	*No suggested mapping**

*By using the phrase "no suggested mapping," DPLA is not stating that a mapping to this element is impossible, but that we have no preferred mapping suggestion.

ore:Aggregation

DPLA LABEL	DPLA PROPERTY	QDC ELEMENT	DC ELEMENT
Aggregated Source Resource	edm:aggregatedCHO	*Provided by DPLA*	*Provided by DPLA*
Data Provider	edm:dataProvider	Last dcterms:contributor	Last dc:contributor
Digital Source Resource Record	dpla:originalRecord	*Provided by DPLA*	*Provided by DPLA*
Has View	edm:hasView	*No suggested mapping**	*No suggested mapping**
Is Shown At	edm:isShownAt	Last dcterms:identifier	Last dc:identifier in the record
Object	edm:object	*No suggested mapping**	*No suggested mapping**
Preview	edm:preview	*No suggested mapping**	*No suggested mapping**
Provider	edm:provider	*Static Value*	*Static Value*
Standardized Rights Statement	edm:rights	*No suggested mapping**	*No suggested mapping**

*By using the phrase "no suggested mapping," DPLA is not stating that a mapping to this element is impossible, but that we have no preferred mapping suggestion.

MODS

dpla:SourceResource

DPLA LABEL	DPLA PROPERTY	MODS ELEMENT
Alternative Title	dcterms:alternative	<titleInfo type="alternative"><title>
Collection	dcterms:isPartOf	<relatedItem type="series">
Contributor	dcterms:contributor	<name><namePart> where <name> also contains <role><roleTerm>contributor</roleTerm></role>
Creator	dcterms:creator	<name><namePart> where <name> also contains <role><roleTerm>creator</roleTerm></role>
Date	dc:date	<originInfo> <dateCreated keyDate="yes">
Description	dcterms:description	<note type="content">
Extent	dcterms:extent	<physicalDescription><extent>
Format	dc:format	<physicalDescription><form>
Genre	edm:hasType	<genre>
Identifier	dcterms:identifier	<identifier>
Language	dcterms:language	<language><languageTerm>
Place	dcterms:spatial	<subject><geographic>
Publisher	dcterms:publisher	<originInfo><publisher>
Relation	dc:relation	<relatedItem><location><url> and/or <relatedItem><titleInfo><title>
ReplacedBy	dpla:replacedBy	<relatedItem type="enumerated: succeeding">
Replaces	dpla:replaces	<relatedItem type="enumerated: preceding">
Rights	dc:rights	<accessCondition>
Rights Holder	dcterms:rightsholder	*No Suggested Mapping**
Subject	dcterms:subject	<subject><topic>
Temporal Coverage	dcterms:temporal	<subject><temporal>
Title	dcterms:title	<titleInfo><title>
Type	dcterms:type	<typeOfResource>

*By using the phrase "no suggested mapping," DPLA is not stating that a mapping to this element is impossible, but that we have no preferred mapping suggestion.

edm:WebResource

DPLA LABEL	DPLA PROPERTY	MODS ELEMENT
File Format	dc:format	<physicalDescription><internetMediaType>
Rights	dc:rights	*No suggested mapping**
Standardized Rights Statement	edm:rights	*No suggested mapping**

*By using the phrase "no suggested mapping," DPLA is not stating that a mapping to this element is impossible, but that we have no preferred mapping suggestion.

ore:Aggregation

DPLA LABEL	DPLA PROPERTY	MODS ELEMENT
Aggregated Source Resource	edm:aggregatedCHO	*Provided by DPLA*
Data Provider	edm:dataProvider	<note type="ownership">
Digital Source Resource Record	dpla:originalRecord	*Provided by DPLA*
Has View	edm:hasView	*No suggested mapping**
Is Shown At	edm:isShownAt	<location><url usage="primary display" access="object in context">
Object	edm:object	*No suggested mapping**
Preview	edm:preview	<location><url access="preview">
Provider	edm:provider	*Static Value*
Standardized Rights Statement	edm:rights	*No suggested mapping**

*By using the phrase "no suggested mapping," DPLA is not stating that a mapping to this element is impossible, but that we have no preferred mapping suggestion.