

http://bit.ly/dpla-metadata-qual

DPLA Metadata Quality Guidelines

This document was created by the DPLA Metadata Quality Working Group, and is now maintained by the DPLA Metadata Working Group. The Metadata Working Group is charged with reviewing the DPLA Metadata Application Profile (MAP) on a regular basis and deciding on the need for updates or revisions; evaluating the need for and developing metadata quality guidelines, documentation, and training to improve the quality of metadata at DPLA and beyond; periodically reviewing and analyzing metadata in DPLA with the goal of providing information back to partners and for the planning and implementation of data quality improvement initiatives; and undertaking initiatives that promote and facilitate the use of DPLA metadata.

Scope of this Documentation

This document includes best practices for creating shareable metadata for the DPLA aggregation. Each property in the DPLA Metadata Application Profile is reviewed with tiered recommendations for minimal, improved, and best quality.

Other Documentation Available:

Please see <u>Appendix 1</u> for additional documentation on specific metadata guidelines such as Introduction to the DPLA Metadata Application Profile, DPLA Standardized Rights Statements Implementation Guidelines, and Metadata Creation Resources.

Introduction

The quality of metadata available through the DPLA search portal, API, and other uses such as Wikimedia is directly proportional to the quality of the metadata it gathers from its partnering Hubs. The guidelines identify best practices for creating shareable metadata for the DPLA aggregation.

Compliance with the recommendations in this document would result in higher quality metadata in virtually any context. There are, however, some specific recommendations that may be incompatible with other systems. These guidelines, therefore, should be taken within the scope of DPLA aggregation only – they are not intended to serve as universally accepted best practices.

The guidelines are developed around three tiers of recommendations for each property in the DPLA Metadata Application Profile (MAP):



Minimal quality requirements

The guidelines in this tier should be considered necessary for data being shared. The guidelines typically relate to correct use of the property or granularity issues.

Each property may or may not have minimal requirements. This does not mean that all properties with a minimal requirement must be included in a record, only that if the property is included these guidelines must be followed.

Recommendations for improved quality

These guidelines recommend methods for improving metadata within the aggregation context, creating better overall consistency through actions such as the use of authorities or content standards.

Recommendations for best quality

These guidelines represent the highest quality benchmarks recommended. They are mostly related to the use of linked data URIs for controlled values. As more data providers add URIs to their data, DPLA will be able to offer enhanced data services. Since linked data is still a relatively new concept in libraries, archives, and museums, the working group has reserved these for the highest quality tier. Over time, as tools for creating linked data become more readily available, the use of URIs should increase.

When using URIs, they should be added in addition to string values.

In addition to the tiers of recommendations, each property is labeled "Required", "Recommended", or "Optional."

In addition to the guidelines for each specific property, the working group recommends the overall strategy of correct and consistent use of properties and the consistent usage of spellings, punctuations, and other formatting considerations. Consistency is the biggest quality issue encountered when aggregating data. Even slight differences in the use of things like attributes or data formats are time consuming to uncover and normalize.

Some examples of typical consistency issues:

- A Hub has aggregated data from two institutions. One uses <dc:publisher> to record the name of the institution itself. Another uses it to record publishing information about the original object. When DPLA receives the data it has no way to tell the two types of data apart.
- A Hub is using <dc:identifier> to record both the URL to the original resource and the thumbnail URL. Some of their data providers also use the field for a call number. The thumbnail URL is always the second instance of <dc:identifier>. However, one institution has no thumbnail images, but does include call numbers. In those records the

second instance of <dc:identifier> is a call number, not a URL.

 A large institution is pulling together MODS records from several units across their distributed library system. While all of them use <mods:dateCreated> for dates, they do not realize that some institutions used the "key date" attribute to distinguish the original item creation date from the date of digitization. Therefore, some records may have

```
<mods:dateCreated keyDate="yes"
begin="1882">1882</mods:dateCreated>
while others have
<mods:dateCreated begin="1882">1882</mods:dateCreated>.
```

If the mapping does not account for the differences in the attributes, errors will result in the mapped data.

• The same institution is also pulling together records that record geospatial data differently. Some place names appear in <mods:geographic> in one set of records and others in <mods:hierarchicalGeographic><mods:state>, <mods:city>, <mods:country>, etc. in another set. While this may be fine, the complexity adds to overall quality assessment time, increases the likelihood of duplication of values (which later have to be removed), and introduces more room for error. A similar example could be made of the use of <dc:format>, <dcterms:medium>, and <dc:type>. All have similar, but slightly different definitions and get used interchangeably at times. Inconsistent use can cause major problems during mapping.

Metadata Guidelines by Property

Source Resource

Alternate Title

Optional

Definition: Any alternative title of the described resource including abbreviations and

translations.

Displayed in DPLA Portal: No.

Available in API: No Used in Wikimedia: No

Minimal quality requirements:

- This is not intended to be a repetition of the main title, however translated titles are acceptable.
- Ensure that the alternative title is for the object, not the title of the series or the collection, or for other related objects.

Recommendations for improved quality:

Break multiple alternative titles into multiple instances of the field.

Recommendations for best quality:

• Refer to content standards like CCO, RDA, DCRMs, etc. for the formulation of titles.

Collection Title

Recommended

Definition: Collection or aggregation of which described resource is a part.

Displayed in DPLA Portal: Yes, and displayed in Collection facet

Available in API: Yes

Searched: Yes

Used in Wikimedia: No

Minimal quality requirements:

- Collections in DPLA MAP can contain both a title and a description. Both may be mapped if they can be indicated consistently in source metadata.
- Use collection titles that are understandable outside of the institutional context (examples: no acronyms, institutional names if necessary, not broad topics used for local browsing).
- Use collection titles consistently across all records.

Contributor

Optional

Definition: Entity (individual or corporate body) responsible for making contributions to

described resource.

Displayed in DPLA Portal: No

Available in API: Yes

Searched: Yes

Used in Wikimedia: No

Minimal quality requirements

• There are no minimal quality requirements for contributor.

Recommendations for improved quality

- Use of an authority list of names (LCNAF, VIAF, ULAN, SNAC, etc.) is highly recommended.
- For names not appearing in authorized lists, follow the naming convention as used by your preferred authority:
 - o Last name, First name preferred
 - List individuals separately (List "John Williamson" and "Mary Williamson" as separate values Not "Mr. and Mrs. John Williamson" as a single value)

Recommendations for best quality

- Use a controlled vocabulary (external or local) when indicating roles.
- If you are using an authorized name, recommend the use of the URI in addition to the string value depending on the schema used.

Creator

Recommended

Definition: Entity (individual or corporate body) primarily responsible for making the described resource.

Displayed in DPLA Portal: Yes. Appears in the search result. Appears as "Creator" in item record.

Available in API: Yes

Searched: Yes

Used in Wikimedia: Yes

Minimal quality requirements

• There are no minimal quality requirements for creator.

Recommendations for improved quality

• Use of an authority list of names (LCNAF, VIAF, ULAN, SNAC etc.) is highly recommended.

- For names not appearing in authorized lists, follow the naming same convention as used by your preferred authority:.
 - Last name, First name preferred
 - List individuals separately (List "John Williamson" and "Mary Williamson" as separate values Not "Mr. and Mrs. John Williamson" as a single value)

Recommendations for best quality

- Use a controlled vocabulary (external or local) when indicating roles.
- If you are using an authorized name, recommend the use of the URI in addition to the string value depending on the schema used.

Date

Recommended

Definition: Date value as supplied by Data Provider.

Displayed in DPLA Portal: Yes. Appears in search result. Appears as "Date" in item record.

Used for Date facet.

Searched: Yes Wikimedia: Yes Available in API:Yes

Minimal quality requirements

- An institution may want to record several different dates in a record. However, the creation date of the original item and the date of digitization must be distinguishable from one another.
- Many schemas have multiple properties related to temporal elements. The date of creation of the original item must consistently be found in a discrete, easy to map property. It must not be mixed with other dates.

Recommendations for improved quality

• Create dates in a consistent format.

Recommendations for best quality

- To support date range searching in the DPLA portal, hubs should provide specific begin
 and end dates that can be mapped to begin and end date fields in the EdmTimeSpan
 class.
- Use of the EDTF is recommended to avoid ambiguity and to normalize the date format.
- There are many ways to express uncertainty about dates. Recommend the use of the standard ways compatible with EDTF found in DPLA's <u>Geographic and Temporal</u> <u>Guidelines</u>. If those are not possible, recommend use of internally consistent methods.

Description

Recommended

Definition: Includes but is not limited to: an abstract, a table of contents, or a free-text account of described resource.

Displayed in DPLA Portal: Yes

Portal Label: Description

Record Preview: Yes

Searched: Yes Wikimedia: Yes Available in API: Yes

Minimal quality requirements

- No raw OCR output.
- No full-text transcription of a textual or linguistic resource.

Recommendations for improved quality

- Do not include HTML markup or formatting in descriptive text. This will be stripped out by the ingestion system.
- Ensure that the description is of the object being described and not a collection to which
 it belongs or any other parent or child entity. The Collection class within the DPLA MAP
 exists to record collection descriptions.

Extent

Optional

Definition: Size, dimensions, or duration of described resource.

Please see the <u>ingestion3 README documentation</u> for information about how the extent field can be filtered or how another field can be filtered to identify values which belong in the extent field.

Displayed in DPLA Portal: Yes

Portal Label: Extent Searched: Yes Wikimedia: No

Available in API: Yes

Minimal quality requirements

• Make sure content is appropriate to the field, not a broader format term.

Recommendations for improved quality

- Recommend use of consistency in handling measurement terminology for maximum understandability. This could include spelling out abbreviations.
- Recommend the use of a current content standard like CCO, RDA, DCRMs, etc.

Format

Recommended

Definition: Physical medium of described resource.

Please see the <u>ingestion3 README documentation</u> for information about how the format field can be filtered or how another field can be filtered to identify values which belong in the format field.

Displayed in DPLA Portal: Yes

Portal Label: Format

Searched: Yes Wikimedia: No Available in API: Yes

Available III AFI. 165

Minimal quality requirements

• Format is a more granular description of the type of object described than the simple vocabulary used in dc:type. It can encompass description of the medium, materials, genre, or other similar terms. Be sure that format is used appropriately and consistently, especially when properties like type and genre are also being used.

Recommendations for improved quality

• Use of a controlled vocabulary (TGM, AAT, etc.) is highly recommended.

Recommendations for best quality

• The use of a URI, in addition to the string value, is recommended when a controlled vocabulary is implemented.

Identifier

Optional

Definition: ID of described resource designated by the contributing institution.

Displayed in DPLA Portal: No

Searched: Yes Wikimedia: Yes Available in API: Yes

No requirements or recommendations.

Language

Recommended

Definition: Language(s) of described resource. Strongly recommended for text materials.

Displayed in DPLA Portal: Yes

Portal Label: Language

Searched: No Wikimedia: No Available in API: Yes

Minimal quality requirements

• Use for objects that contain text or speech only. If used for non-linguistic objects, use a code that indicates no linguistic content.

Recommendations for improved quality

• Use a controlled vocabulary for language names.

Recommendations for best quality

• Use the ISO 639-3 standard for best results.

Place

Recommended

Definition: Spatial characteristics of described resource, such as a country, city, region,

address or other geographical term. Captures aboutness.

Displayed in DPLA Portal: Yes

Portal Label: Location

Searched: Yes **Wikimedia**: No

Available in API: Yes

Minimal quality requirements

Please see DPLA's Temporal and Geographic Guidelines.

 Use only for spatial topics that a resource is about. Not intended to capture the place of publication or digitization.

Publisher

Optional

Definition: Entity responsible for making the described resource available, typically the

publisher of a text (not edm:dataProvider or edm:provider).

Displayed in DPLA Portal: Yes

Portal Label: Publisher

Searched: Yes Wikimedia: No

Available in API: Yes

Minimal quality requirements

- The field is intended to contain the publisher of the original item.
- Not intended to capture the institutions involved in digitization or hosting.
- Use the Creator field to capture the entity primarily responsible for making the described resource.

Relation

Optional

Definition: Related resource. **Displayed in DPLA Portal**: No.

Searched: Yes Wikimedia: No. Available in API: Yes

Minimal quality requirements

Relation is intended for use with other items that have some relationship with the
content. However, if the relationship is specifically that of collection membership or
replacing another item, those more specific fields should be used.

Recommendation for improved quality

 Recommend only using information that is intelligible outside of the original institution context. For example, call numbers or identifiers that do not make sense out of context should not be included.

•

Rights (Free text)

Required (If no URI Rights)

Definition: Information about rights held in and over the described resource. Typically, rights information includes a statement about various property rights associated with the described resource, including intellectual property rights.

Displayed in DPLA Portal: Yes

Searched: Yes Wikimedia: No. Available in API: Yes

Minimal quality requirements

 The value should not contradict or repeat information found in the other rights fields, such as that of a standardized URI from RightsStatements.org. See http://bit.ly/dpla-rights-quidelines for more details.

Recommendations for improved quality

 Use an affirmative statement that provides additional value and/or context to the URI Rights field. Do not use a boilerplate statement such as "Please contact X institution for information about rights."

Rights Holder

Optional

Definition: A person or organization owning or managing rights over the resource.

Displayed in DPLA Portal: No.

Searched: No Wikimedia: No. Available in API: No

Minimal quality requirements

Make sure this is not an institution contact, but the actual rights holder.

Subject

Recommended

Definition: Topic of described resource.

Displayed in DPLA Portal: Yes, also appears in Subject facet.

Searched: Yes Wikimedia: No

Available in API: Yes

Minimal quality requirements

• There are no minimal quality requirements for subject.

Recommendations for improved quality

 Use of a controlled vocabulary (LCSH, TGM, FAST, etc.) or name authority (LCNAF, VIAF, etc.) is highly recommended. Please see <u>this document</u> for additional vocabulary resources.

Recommendations for best quality

- If you are using a controlled vocabulary, we recommend providing either a label or URI
 that identifies the authority (e.g. <untl:subject <u>qualifier="LCSH"</u>>Texas -- Annexation to
 the United States.</untl:subject>)
- If you are using a controlled vocabulary, we recommend the use of the URI in addition to the string value depending on the schema used.
- Recommend uncoordinated subject heading if headings are being newly created. For example, in a Dublin Core record you might use:

```
<dc:subject>Civil rights movements</dc:subject>
<dc:subject>Mississippi</dc:subject>
<dc:subject>Jackson</dc:subject>
Instead of
<dc:subject>Civil rights movements -- Mississippi --
Jackson</dc:subject>
```

This suggestion is made to increase matching of terms in the aggregated data set and facets. Not all providers will use the same controlled vocabulary lists. Even among those who do, the granular nuance of the coordinated subject headings makes it impossible to bring together records based on the larger concepts present in the heading.

Subtype

Not yet implemented, DPLA-supplied

Definition: Captures categories of a described resource in a given field. Does not capture aboutness. See Appendix C in the <u>DPLA Metadata Application Profile, version 5.0</u> for a list of values. These values may be supplied by the provider, or added by DPLA based on existing format and type terms during mapping.

Displayed in DPLA Portal: No.

Minimal quality requirements

 Use of the values found in Appendix C in the <u>DPLA Metadata Application Profile</u>, version 5.0

Recommendations for best quality

 Use of the URIs for values found in Appendix C in the <u>DPLA Metadata</u> <u>Application Profile, version 5.0</u>

Temporal Coverage

Optional

Definition: Temporal characteristics of the described resource. Captures aboutness.

Displayed in DPLA Portal: No

Searched: No Wikimedia: No Available in API: Yes

Minimal quality requirements

• Temporal is intended to describe the temporal characteristics of an object, not the specific date of creation.

Recommendations for best quality

• Temporal values are best captured when they are separated (i.e., not combined with spatial information in a generic "coverage" element).

Title

Required

Definition: Primary name given to the described resource.

Displayed in DPLA Portal: Yes. Appears in search result. Appears as "Title" in item record.

Searched: Yes Wikimedia: Yes Available in API: Yes

Minimal quality requirements

• There are no minimum requirements for title other than the presence of a value.

Recommendations for improved quality

- Descriptive and informative titles are preferred whenever possible (as opposed to things like "unknown" or an id number). Not all materials can or should be titled uniquely; but this recommendation exists to encourage data creators to create unique and informative titles when possible.
- When titles are created for works, they should be concise. The description field should be used for more detail.
- Consider how titles will be interpreted outside of local context.

Recommendations for best quality

- Refer to content standards like CCO, RDA, DCRMs, as well as any <u>local guidelines or</u> documentation.
- Recommend minimal but appropriate use of punctuation (i.e., not unnecessary quotation marks, brackets or periods).
- DPLA will remove additional punctuation to help normalize all values. For additional information please refer to https://qithub.com/dpla/ingestion3/#text-normalizations]

Type

Recommended

Definition: Nature or genre of described resource.

Displayed in DPLA Portal: Yes. Appears in search result. Appears as "Type" in item record.

Appears in Type facet.

Searched: Yes Wikimedia: No

Available in API: Yes

Minimal quality requirements

• There are no recommendations for minimum quality.

Recommendations for improved quality

 Type is intended as a broad categorization, not a more granular term like format or genre. Types need to be distinguishable from these other terms.

Recommendations for best quality

- Use the DCMI type vocabulary if possible, or an internal standard that can be mapped to DCMI and is consistent. Some examples of other vocabularies that DPLA could easily match to DCMI terms are the MODS type of resource values and the Library of Congress's Content Types list.
- Learn more about how DPLA normalizes and enriches provided type values to DCMIType values in the <u>ingestion3 README</u>.

ORE:Aggregation

AggregatedSR

DPLA-supplied

Definition: Unambiguous ID to SR. **Displayed in DPLA Portal**: No.

This property is supplied by DPLA at time of metadata ingest.

Data Provider

Required

Definition: The organization or entity that supplies metadata about digital content through a provider or hub to DPLA.

Displayed in DPLA Portal: Yes. Appears in search results. Appears as "Contributing Institution" in item record. Appears in Contributing Institution facet (along with Intermediate Provider).

Minimal quality requirements

- Ensure that the institutional name is used correctly and consistently throughout.
- The data must be in a discrete property that is easily mapped.

For Content Hubs the values in this property and the field Provider will be identical.

Digital Resource Original Record

DPLA-supplied

Definition: Complete original record provided by the provider or hub.

Displayed in DPLA Portal: No.

This property is created by DPLA at time of metadata ingest.

Intermediate Provider

Optional

Definition: An organization that enables the data supply chain between Data Provider and Provider. The relationship between these three organizations can vary. The Intermediate Provider must be distinct from both the Data Provider and the Provider.

Displayed in DPLA Portal: Yes. Appears as "Supporting Institution" in item record. Appears in Contributing Institution facet (along with Data Provider).

Minimal quality requirements

Ensure that the institutional name is used correctly and consistently throughout.

Is Shown At

Required

Definition: Unambiguous URL reference to digital object in its full information context.

Displayed in DPLA Portal: Yes. Appears as "URL" in item record.

Minimal quality requirements

- The URL must resolve to a web page showing the item and its accompanying metadata.
- The data must be in a discrete property that is easily mapped.

Object

Optional

Definition: The URL of a suitable source object in the best resolution available on the website

of the Data Provider.

Displayed in DPLA Portal: No.

Available in API: Yes

Searched: No. Wikimedia: No

Minimal quality requirements

- Not to be confused with Preview, this url leads to a high resolution object.
- Images should represent the item and not generic placeholders.

Preview

Recommended

Definition: The URL of a thumbnail, extract, or other type of resource representing the digital object for the purposes of providing a preview.

Displayed in DPLA Portal: No (the image that is linked to appears, but the text of the URL here does not appear).

Available in API: Yes

Searched: No Wikimedia: No

Minimal quality requirements

- The URL must resolve to a digital object.
- Images should represent the item and not generic placeholders.

Recommended for improved quality

• Images should be at least 300px on the long side

Provider

DPLA-supplied, this property is created by DPLA at time of metadata ingest.

Required

Definition: The name of the Service Hub or Content Hub.

Displayed in DPLA Portal: Yes. Appears as "Partner" in item record. Appears in the Partner

facet.

Available in API: Yes

Searched: Yes Wikimedia: Yes

Rights (URI)

Required (If no Freetext Rights)

Information about rights held in and over the described resource. Typically, rights information includes a statement about various property rights associated with the described resource, including intellectual property rights.

Displayed in DPLA Portal: Yes. Appears as "Rights" in item record. Values are clustered and grouped into "How Can I Use It?" facet. For more information about how Rights Statements are grouped in the facet, please see the FAQ on Copyright Status Filter

https://dp.la/about/rights-categories

DPLA displays labels, descriptive text and badges for these URI values based on lookups against this file (https://github.com/dpla/dpla-frontend/blob/main/constants/rights.js). Please note that not all URIs have a description.

Available in API: Yes.

Required for Wikimedia: Yes **Minimal quality requirements**

- The value must be a single URI from the RightsStatements.org or Creative Commons vocabularies. See http://bit.ly/dpla-rights-guidelines for more details.
- DPLA performs a series of normalizations on provided values to bring them into alignment but we can only accept values which match valid RightsStatements.org or Creative Commons URIs. See

https://github.com/dpla/ingestion3#edmrights-normalization-and-validation for descriptions of validations and

https://github.com/dpla/ingestion3/blob/6a4e1e38152da480e5b33070df2996fedd3ea51f/src/main/scala/dpla/ingestion3/model/DplaMapData.scala#L153-L745 for the full list of valid URIs.

IIIF Manifest

Optional

Definition: A resource that references or otherwise points to the described resource. This field is used for a manifest URI for a IIIF resource. This resource will be the source for image uploads to Wikimedia Commons.

Displayed in DPLA Portal: No

Available in API: Yes

Required for Wikimedia: Yes [either IIIF Manifest or Media Master is required]

Minimal quality requirements

Must resolve to an actionable URI for the presentation manifest of a IIIF object.

Media Master

Optional

Definition: URLs to full frame image resources to be uploaded to Wikimedia Commons.

Displayed in DPLA Portal: No

Available in API: Yes

Required for Wikimedia: Yes [either IIIF Manifest or Media Master is required]

Minimal quality requirements

Must resolve to a valid image asset that can be uploaded to Wikimedia Commons.

Appendix 1

An Introduction to the DPLA Metadata Application Profile

https://dp.la/info/developers/map/

A general introduction to DPLA's metadata standards, workflows, and processes and introduces the model for how data is stored in DPLA.

Geographic & Temporal Guidelines for MAP

http://bit.ly/dpla-geo-styleguide-3 1

Recommendations for the formatting of geographic and temporal data in records that will be shared with DPLA.

DPLA Metadata Application Profile

https://dp.la/info/developers/map/

A definition of the metadata properties and classes used in DPLA.

DPLA Standardized Rights Statements Implementation Guidelines

http://bit.ly/dpla-rights-guidelines

A description of DPLA's implementation of standardized rights statements and recommendations for the use of statements in records that will be shared with DPLA.

DPLA ingestion3 Text Enrichments and Normalizations

https://github.com/dpla/ingestion3#enrichment-and-normalization

A summary of the specific normalizations and enrichments that are applied to each field during metadata ingestion.

Hub Metadata Guidelines

https://docs.google.com/spreadsheets/d/1CWbYIXu3urH7OHwQ1w5ugL_p7Folfs36E0yB3 DLOwkc/edit#gid=550922422&range=A1

A compilation of hub specific metadata guidelines and documentation

Metadata Creation Resources

https://docs.google.com/spreadsheets/d/1CWbYIXu3urH7OHwQ1w5ugL_p7Folfs36E0yB3 DLOwkc/edit?usp=sharing

A compilation of metadata creation guidelines and documentation

DPLA's Statement on Potentially Harmful Content

https://dp.la/about/harmful-language-statement

A statement about the nature of DPLA content, its origins, and language used in description.

Appendix 2

Original authors

Robin Dean, Michigan State University
Leanne Finnigan, Temple University
Gretchen Gueguen, DPLA
Christina Harlow, Cornell University
Ann Lally, University of Washington
Susan Lynch, The New York Botanical Garden
Catherine Phan, University of Wisconsin-Madison
Mark Phillips, University of North Texas
Danielle Pucci, Boston Public Library
Chris Stanton, Metropolitan New York Library Council (METRO)

Rev. Jun 1, 2022

Greta Bahnemann, Minnesota Digital Library
Jeremy Bartczak, Digital Virginias
Erica Boudreau, National Archives & Records Administration
Graham Dethmers, HathiTrust Digital Library
Leanne Finnigan, PA Digital
Teresa Hebron, Mountain West Digital Library
Nicole Lawrence, Digital Library of Georgia
Amelia Mowry, Wayne State University
Rachel Senese, Georgia State University
Penelope Schumaker, Ohio Digital Network
Sarah Beth Seymore, University of Oregon
Elliot Williams, Texas Digital Library
Scott Williams, DPLA