Metadata Best Practices Report

August 10, 2023

Report Authors

DPLA Metadata Working Group Co-Chairs 2022-2023

Nicole Lawrence, Digital Library of Georgia
Elliot Williams, University of Texas at San Antonio

DPLA Metadata Working Group Members 2022-2023

Erica Boudreau, National Archives and Records Administration
Graham Dethmers, HathiTrust Digital Library
Amy Hitchner, Colorado State Library
Molly O’Brien, New York Public Library
Rachel Senese, Georgia State University
Michael L. Stewart, University of Delaware

Purpose

The purpose of this report is to summarize the metadata guidelines and best practices analysis work the Metadata Working Group (MWG) has completed and propose a series of recommendations for future work, to be reviewed by the Digital Public Library of America (DPLA) hub network. The hub network will be asked to review the report and given the chance to vote on the work they would like to see the group tackle going forward.

Method

In August 2022, the MWG began analyzing metadata guidelines and best practices across hubs. The MWG had previously compiled a list of hub metadata guidelines in 2021, which was used as a starting place for this project. An email was sent to the DPLA allhubs email list in September 2022, asking for additions or updates to that list, which resulted in a small number of updates to the list of guidelines documents.

Of the 48 content and service hubs currently listed on the DPLA website, metadata guidelines documentation was found for 33 hubs. The group decided to exclude documentation from hubs that are not currently active within DPLA. We also excluded documentation from large content hubs (such as HathiTrust and Library of Congress) that were focused on the internal practices of those organizations and not representative of the data that is sent to DPLA. In total, documentation from 24 hubs was included, representing 50% of the current hubs in the DPLA network.

To begin analysis, each hub's documentation was analyzed by one member of the working group. We collected data about individual metadata elements, focusing on elements that are sent to DPLA; elements that are not included in the DPLA metadata element set were excluded from analysis. After data was collected for each hub's guidelines, the group standardized the information gathered, particularly related to requirement and content. For each element, the following information was recorded:

- Name of element: Field name used in documentation
- Property: Dublin Core and/or Metadata Object Description Schema (MODS) elements that the element is mapped to
- Requirement: Whether the element is required. Values used are: required, recommended, strongly recommended, optional, unknown (requirement not specified)
- Content rules or requirements: Whether content rules apply to the element. Values used are: unstructured (free text with no guidelines or restrictions), semi-structured (free text with guidelines or strongly suggested vocabularies), fully structured (tightly restricted from a required set of terms)
- Cardinality: How many times the element can be used, or how many values are allowed. Values were not standardized for this.
- Notes: Additional information about the element, including any applicable content requirements or suggestions.

Next, each element was analyzed and compared across the guidelines. One group member wrote a short summary, noting areas of similarity and differences across the hubs. Properties, requirement, and structured content were also counted for each element. These summaries were used as the basis for discussion of each element by the entire working group. Recommendations were made for fields where variability in hub guidelines could impact the DPLA front-end and/or related projects (i.e. Wikimedia) or where field usage needs clarification.

Analysis by field

- ### Title
    - DPLA definition: Primary name given to the described resource.
    - Nearly universally required. Style and formatting recommendations vary across guidelines; much of the variation comes from differences in source systems and descriptive practices (archival, MARC, etc.). Because some differences will always be present in Title guidelines, no further analysis is needed.
    - Recommendation: None
- ### Subject
    - DPLA definition: Topic of described resource.
    - Fairly consistent suggestions for a number of vocabularies, with Library of Congress (LC) headings being the most widely adopted followed by Thesaurus for Graphic Materials (TGM) and Art & Architecture Thesaurus (AAT). Almost all guidelines reference at least

    one vocabulary as a starting point. Of note, only two hubs mention URIs or de-coordinated strings. These are two areas that greatly impact Wikimedia usage and are a priority for DPLA at the moment.
    - Recommendation: Explore the existing URI and de-coordinated recommendations and work with the Wikimedia staff at DPLA to develop a general set of Wikimedia specific subject guidelines.
- Rights (URI)
    - DPLA 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.
    - Very consistent across guidelines. Most hubs point to rightsstatements.org and creativecommons.org for URIs. A few hubs include a reference to using an additional free text rights field for additional, non-uri information. Since hubs and DPLA already provide clear recommendations for rights URIs, no further analysis is needed.
    - Recommendation: None
- Alternate title
    - DPLA definition: Any alternative title of the described resource including abbreviations and translations.
    - Variation in guideline interpretation. Can be a foreign language translation, subtitle, or alternate title. Not published on DPLA front end; no further analysis is needed.
    - Recommendation: None
- Collection title
    - DPLA definition: Collection or aggregation of which described resource is a part.
    - Variability in how it is used across guidelines but they do fit within DPLA's broad definition. Includes information either about the parent collection from which the digital object originates *and/or* the name of the institution which contributed the object.
    - Recommendation: Further analyze the field and clarify the definition in DPLA guidelines to better facilitate use in DPLA.
- Contributor
    - DPLA description: Entity (individual or corporate body) responsible for making contributions to a described resource.
    - Mostly a semi-structured and optional or recommended field. Hubs suggest Library of Congress Name Authority File (LCNAF); other suggested controlled vocabularies include Virtual International Authority File (VIAF) or Union List of Artist Names (ULAN).
    - Recommendation: Follow the recommendation for Creator.
- Creator
    - DPLA description: Entity (individual or corporate body) primarily responsible for making the described resource.
    - Mostly a semi-structured and recommended/strongly recommended field. Hubs suggest LCNAF or to follow LCNAF formatting if the creator has no official entry. Other suggested

    controlled vocabularies include VIAF, ULAN, and Social Networks and Archival Context (SNAC).
  - Recommendation: Check with the Wikimedia team to see if they will be tackling Creator within the project in the near future. If so, get clarification of what they need. If not, no recommendations.
- Date
  - DPLA description: Date value as supplied by Data Provider.
  - Required or recommended in majority of guidelines, and most hubs recommend a standard such as ISO 8601 or Extended Date/Time Format (EDTF). DPLA can map from most date fields, but generating begin and end dates can be difficult. DPLA's ideal format would be a combination of date label, begin date, and end date, but that is not supported by most metadata schemas. Since hubs and DPLA already provide clear recommendations for date formatting, no further analysis is needed.
  - Recommendation: None
- Temporal coverage
  - DPLA description: Temporal characteristics of the described resource. Captures aboutness.
  - Many hubs and institutions do not use this field. Among those that do, recommendations for values differ: some use a date range, some use a named period. Not published on DPLA front end; no further analysis is needed.
  - Recommendation: None
- Description
  - DPLA description: Includes but is not limited to: an abstract, a table of contents, or a free-text account of described resource.
  - Primarily a free text field with similar formatting recommendations across hubs. Not accepted are OCR, transcripts, or descriptions of related resources. Since there is no control of content, no further analysis is needed.
  - Recommendation: None
- Extent
  - DPLA definition: Size, dimensions, or duration of described resource.
  - Primarily formatted as an unstructured free-text field, most hubs use the Dublin Core Metadata Initiative (DCMI) and DPLA MAP definition of "size or duration" though some specifically mention AV length, dimensions, and/or page numbers. Since there is no control of content, no further analysis is needed.
  - Recommendation: None
- Format
  - DPLA definition: Physical medium of described resource.
  - Though the DPLA definition notes that this field should describe the physical medium of the resource, hub usage varies, with some using it as recommended, some using it to describe the material's digital properties (e.g. MIME type), and some using the field to distinguish between born-digital and digitized materials. DPLA already employs filters

    which transform Format data for population of both the Format field itself and associated fields like Genre.
    - Recommendation: The Metadata Working Group will encourage hubs to communicate their use of the Format field to DPLA to help optimize filter usage and further analyze the data being submitted to clarify the relationship to other fields like genre and subtype.
- Type
    - DPLA definition: Nature or genre of described resource.
    - Primarily a structured or semi-structured field, the vast majority of hubs already control their type data to the currently recommended DCMI type vocabulary (or the MODS typeOfResource value which maps easily to it).
    - Recommendation: None
- Genre
    - DPLA definition: Not currently defined in the DPLA Metadata Quality Guidelines
    - Information provided in Genre may be used to populate associated fields such as Subtype and Format. Many hubs and institutions do not use this field.
    - Recommendation: Further analysis by the Metadata Working Group will be undertaken to help clarify the use of filters by DPLA during ingest and the relationship between Genre, Format, and Subtype (see also: Format).
- Identifier
    - DPLA definition: ID of described resource designated by the contributing institution.
    - Majority of the hubs are looking for a URI though some would like a specific file name or location. This field is primarily unstructured or semi-structured. Identifier is one of the fields that is included in metadata uploaded to Wikimedia.
    - Recommendation: Work with the Wikimedia team to incorporate training that emphasizes the role of identifiers in helping people identify items across discovery platforms.
- Language
    - DPLA definition: Language(s) of described resource. Strongly recommended for text materials.
    - This field is either recommended or strongly recommended by over half of the evaluated hubs. Almost all hubs recommend ISO 639 though some more specifically recommend 639-3 or 639-2. For best quality, DPLA already recommends 639-3. Since hubs and DPLA already provide clear recommendations for language formatting, no further analysis is needed.
    - Recommendation: None
- Place
    - DPLA definition: Spatial characteristics of described resource, such as a country, city, region, address, or other geographical term. Captures aboutness.

    - Primarily structured or semi-structured, with a variety of vocabularies and formatting practices used. There is enough guidance already provided in the DPLA Geographic and Temporal Guidelines for MAP 3.1.
    - Recommendation: None
- Publisher
    - DPLA definition: Entity responsible for making the described resource available, typically the publisher of a text.
    - Hub definitions vary to include publisher of original, publisher of digital surrogate, or both. Most hubs recommend using LCNAF or the same format. Free text. Since there is no control of content, no further analysis is needed.
    - Recommendation: None
- Relation
    - DPLA definition: Related resource
    - Relation is intended for related resources other than the Collection title. Hubs may want to check that they are using the Collection field appropriately instead of Relation. Otherwise this field is used for whatever makes sense in the context of the institution. Many hubs and institutions do not use this field. Not published on DPLA front end; no further analysis is needed.
    - Recommendation: The Metadata Working Group will encourage hubs to evaluate their use of the relation field to ensure it does not contain collection titles.
    -
- Rights (Free text)
    - DPLA 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.
    - Most hubs use this field to supplement the Rights (URI) field. Contains a variety of content including access rights, copyright holder, usage restrictions, guidelines, contact info, and/or local rights statement. It most commonly contains free text but may also contain URIs. Since there is no control of content, no further analysis is needed.
    - Recommendation: None
- Rights Holder
    - DPLA definition: A person or organization owning or managing rights over the resource.
    - Used in seven of the twenty four hubs. Mostly unstructured data with a few hubs noting it can be a free text string or URI. Since there is no control of content, no further analysis is needed.
    - Recommendations: None
- Subtype
    - DPLA definition: Captures categories of a described resource in a given field. Does not capture aboutness.
    - This field was eliminated from analysis as only one hub uses the field and their documentation points to DPLA's MAP v5 rather than outlining their own guidelines.

    - Recommendations: None

Additional recommendations

- Standard formatting
    - In doing this research, the group discovered there is no format consistency between guidelines. This made it difficult to compare recommendations across hubs even with the group composed of metadata professionals. This poses an even higher barrier for non-professional contributors who are not familiar with metadata terminology or the various guideline formats. We propose developing a standard format that can be easily interpreted no matter the level of metadata training.
- Inclusive vocabularies

    While discussing controlled vocabularies used for fields such as Subject and Creator, the group discussed the growing use of alternative, more inclusive vocabularies and standards. Use of those vocabularies can have a positive impact on discoverability within DPLA, as well as downstream uses of metadata, such as inclusion in Umbra Search. We also recognize that the political climate in some states makes it difficult for some hubs to promote inclusive metadata practices as openly as they would like. We suggest further exploration of the use and impact of these vocabularies, and the possible creation of resources that contributing institutions can use as justification for implementing them.
- Reparative description practices & policies
    - The working group is interested in knowing more about how reparative description practices are being implemented across the DPLA hub network. We know that many hubs and contributing institutions are undertaking reparative description work; some are publicly discussing their efforts, and others are engaging in it more quietly. We recommend that the working group explore how hubs and contributors are engaging in reparative description and consider ways to support that work across the DPLA network.

Summary of recommendations

- Explore the existing **Subject** URI and de-coordinated recommendations and work with the Wikimedia staff at DPLA to develop a general set of Wikimedia specific subject guidelines.
- Further analyze **Collection Title** across guidelines and clarify DPLA definition to facilitate improved usage.
- Check with the Wikimedia team to see if they will be tackling **Creator** and by extension, **Contributor**, within the project in the near future. If so, get clarification of what they need and plan next steps.
- Encourage hubs to communicate their use of **Format** to DPLA to help optimize filter usage and further analyze the data being submitted to clarify the relationship to other fields like Genre and Subtype.

- Undertake further analysis to help clarify the use of filters by DPLA during ingest and the relationship between **Genre**, **Format**, and **Subtype**.
- Work with the Wikimedia team to incorporate training that emphasizes the role of **Identifier** in helping people identify items across discovery platforms.
- Develop and propose a standard format for metadata guidelines.
- Explore the use and impact of alternative, inclusive vocabularies, and possibly create resources for implementation.
- Explore how hubs and contributors are engaging in reparative description and consider ways to support that work across the DPLA network.

Appendix A – Data, Guidelines, and additional data

Data collection spreadsheet

DPLA Metadata Quality Guidelines

Appendix B – Field summaries

Field Name	**Required**	**Content**	**Key Standards**
Title	Required: 22 Strongly recommended: 0 Recommended: 0 Optional: 0 Unknown: 2 Not applicable: 0	Fully structured: 0 Semi-structured: 2 Unstructured: 21 Unknown: 1 Not applicable: 0	
Subject	Required: 5 Strongly recommended: 6 Recommended: 7 Optional: 3 Unknown: 3 Not applicable: 0	Fully structured: 4 Semi-structured: 19 Unstructured: 1 Unknown: 0 Not applicable: 0	LCSH, TGM, AAT, LCNAF, FAST, MeSH, VIAF, ULAN
Rights (URI)	Required: 13 Strongly recommended: 1 Recommended: 1 Optional: 1 Unknown: 0 Not applicable: 8	Fully structured: 12 Semi-structured: 3 Unstructured: 1 Unknown: 0 Not applicable: 8	RightsStatements.org, Creative Commons

Alternate title	Required: 0 Strongly recommended: 0 Recommended: 0 Optional: 10 Unknown: 1 Not applicable: 13	Fully structured: 0 Semi-structured: 1 Unstructured: 10 Unknown: 0 Not applicable: 13	
Collection title	Required: 8 Strongly recommended: 1 Recommended: 2 Optional: 2 Unknown: 2 Not applicable: 9	Fully structured: 2 Semi-structured: 3 Unstructured: 9 Unknown: 1 Not applicable: 9	
Contributor	Required: 0 Strongly recommended: 1 Recommended: 5 Optional: 12 Unknown: 2 Not applicable: 4	Structured: 2 Semi-structured: 16 Unstructured: 2 Unknown: 0 Not applicable: 4	LCNAF, VIAF, ULAN
Creator	Required: 3 Strongly recommended: 4 Recommended: 10 Optional: 4 Unknown: 3 Not applicable: 0	Fully structured: 1 Semi-structured: 22 Unstructured: 1 Unknown: 0 Not applicable: 0	LCNAF, VIAF, ULAN
Date	Required: 8 Strongly recommended: 6 Recommended: 5 Optional: 3 Unknown: 2 Not applicable: 0	Fully structured: 8 Semi-structured: 16 Unstructured: 0 Unknown: 0 Not applicable: 0	ISO 8601, EDTF, W3CDTF
Temporal coverage	Required: 0 Strongly recommended: 1 Recommended: 1 Optional: 10 Unknown: 1 Not applicable: 10	Fully structured: 2 Semi-structured: 5 Unstructured: 6 Unknown: 0 Not applicable: 10	

Description	Required: 3 Strongly recommended: 1 Recommended: 11 Optional: 6 Unknown: 2 Not applicable: 1	Fully structured: 0 Semi-structured: 5 Unstructured: 18 Unknown: 0 Not applicable: 1	
Extent	Required: 0 Strongly recommended: 0 Recommended: 3 Optional: 10 Unknown: 0 Not applicable: 11	Fully structured: 0 Semi-structured: 3 Unstructured: 10 Unknown: 0 Not applicable: 11	
Format	Required: 7 Strongly recommended: 0 Recommended: 8 Optional: 5 Unknown: 2 Not applicable: 2	Fully structured: 11 Semi-structured: 8 Unstructured: 3 Unknown: 0 Not applicable: 2	Media Types, AAT, TGM
Type	Required: 13 Strongly recommended: 4 Recommended: 2 Optional: 2 Unknown: 3 Not applicable: 0	Fully structured: 16 Semi-structured: 8 Unstructured: 0 Unknown: 0 Not applicable: 0	DCMI Type, MODS Type
Genre	Required: 2 Strongly recommended: 0 Recommended: 2 Optional: 1 Unknown: 0 Not applicable: 19	Fully structured: 1 Semi-structured: 4 Unstructured: 0 Unknown: 0 Not applicable: 19	AAT, TGM
Identifier	Required: 9 Strongly recommended: 0 Recommended: 2 Optional: 9 Unknown: 2 Not applicable: 2	Fully structured: 6 Semi-structured: 8 Unstructured: 8 Unknown: 0 Not applicable: 2	

Language	Required: 2 Strongly recommended: 7 Recommended: 7 Optional: 4 Unknown: 2 Not applicable: 2	Fully structured: 11 Semi-structured: 8 Unstructured: 3 Unknown: 0 Not applicable: 2	ISO 639-3, ISO 639-2
Place	Required: 1 Strongly recommended: 5 Recommended: 4 Optional: 7 Unknown: 1 Not applicable: 6	Fully structured: 3 Semi-structured: 14 Unstructured: 1 Unknown: 0 Not applicable: 6	GeoNames, LCSH, TGN, FAST, LCNAF
Publisher	Required: 2 Strongly recommended: 2 Recommended: 5 Optional: 9 Unknown: 2 Not applicable: 4	Fully structured: 1 Semi-structured: 8 Unstructured: 11 Unknown: 0 Not applicable: 4	
Relation	Required: 1 Recommended: 2 Optional: 10 Unknown: 1 Not applicable: 10	Fully structured: 1 Semi-structured: 6 Unstructured: 7 Unknown: 0 Not applicable: 10	
Rights (Free text)	Required: 6 Strongly recommended: 1 Recommended: 1 Optional: 6 Unknown: 2 Not applicable: 8	Fully structured: 1 Semi-structured: 4 Unstructured: 11 Unknown: 0 Not applicable: 8	
Rights Holder	Required: 0 Strongly recommended: 0 Recommended: 2 Optional: 5 Unknown: 0 Not applicable: 17	Fully structured: 0 Semi-structured: 2 Unstructured: 0 Unknown: 0 Not applicable: 17	

Subtype	Required: 0 Strongly recommended: 1 Recommended: 0 Optional: 1 Unknown: 0 Not applicable: 22	Fully structured: 1 Semi-structured: 1 Unstructured: 0 Unknown: 0 Not applicable: 22	AAT

Appendix C - Acronyms

AAT - Art & Architecture Thesaurus
AV - Audio/Visual
DC - Dublin Core
DCMI - Dublin Core Metadata Initiative
DPLA - Digital Public Library of America
EDTF - Extended Date/Time Format
FAST - Faceted Application of Subject Terminology
ISO - International Organization for Standardization
LC - Library of Congress
LCNAF - Library of Congress Name Authority File
LCSH - Library of Congress Subject Headings
MAP - Metadata Application Profile
MARC - Machine-Readable Cataloging
MeSH - Medical Subject Headings
MIME - Multipurpose Internet Mail Extensions
MODS - Metadata Object Description Schema
MWG - Metadata Working Group
OCR - Optical Character Recognition
SNAC - Social Networks and Archival Context
TGM - Thesaurus for Graphic Materials
TGN - Thesaurus of Geographic Names
ULAN - Union List of Artist Names
URI - Uniform Resource Identifier
VIAF - Virtual International Authority File
W3CDTF - World Wide Web Consortium Date and Time Formats