MODULE 20

SHARING ARCHIVAL METADATA

Aaron Rubinstein

The Digital Public Library of America's Application Programming Interface and Metadata Ingestion Process

by Mark A. Matienzo, Stanford University Libraries
(Matienzo served as DPLA's Director of Technology until September 2016)

The Digital Public Library of America (DPLA) is a nonprofit dedicated to providing access to and promoting openly available cultural heritage resources made available digitally from institutions, including libraries, archives, museums, and historical societies, across the United States. DPLA provides access to materials from more than 1,600 contributing institutions made available through a network of thirty-seven Hubs,[55] the primary partners from which DPLA harvests metadata. Hubs are either state-wide or regional digital libraries that provide services for their given community or large institutions that maintain a one-to-one relationship with DPLA. The metadata aggregated by DPLA is also made freely available for access and reuse, through the user-facing portal, as a downloadable dataset, and through a freely available application programming interface, or API. The DPLA API, a web-based service, directly provides the user-facing portal with the ability to access and search the aggregated metadata.[56] This case study describes the workflow wherein DPLA obtains and transforms the metadata from its hubs, as well as the underlying design philosophy for the DPLA API.

The Metadata Ingestion Process

The metadata provided by DPLA's Hubs goes through a set of steps in which DPLA harvests the metadata, maps the metadata to a common format and structure, enriches the metadata to address data quality issues and to add value to it, and indexes it so the metadata can be accessed through the DPLA API. This overall process is referred to as DPLA's *metadata ingestion process*.[57] The first step in the process of

55 Digital Public Library of America, "Hubs," http://dp.la/info/hubs/, captured at https://perma.cc/DJ96-Z3XD.

56 Digital Public Library of America, "API Codex," http://dp.la/info/developers/codex/, captured at https://perma.cc/6YA3-9VM7.

57 More information about the DPLA ingestion process, specifically in terms of known issues with the process, can be found in Mark A. Matienzo and Amy Rudersdorf, "The Digital Public Library of America Ingestion Ecosystem: Lessons Learned After One Year of Large-Scale Collaborative Metadata Aggregation," *Proceedings of the International Conference on Dublin Core and Metadata Applications 2014*, October 2014, http://dcpapers.dublincore.org/pubs/article/view/3700. Full paper captured at https://perma.cc/QM46-2YC8.

bringing metadata into DPLA is *harvesting*. Before harvesting metadata, DPLA staff members work with a Hub to identify the best standardized process by which it will receive their metadata. Generally speaking, DPLA is able to work with nearly any schema in which metadata is expressed and any method used to harvest metadata. Schemas in use vary widely across the Hubs and include MODS, MARCXML, simple and qualified Dublin Core, and a number of system- or institution-specific schemas. Harvesting methods also differ but most often include the OAI Protocol for Metadata Harvesting, site-specific APIs, or downloadable dumps of records.

Given the wide range of schemas in which Hubs provide metadata, DPLA undertakes additional steps to process the metadata into a common form usable by the DPLA API and other applications. This step is *mapping* the incoming metadata to the DPLA Metadata Application Profile, or DPLA MAP.[58] The DPLA MAP is an application profile[59] based on the Europeana Data Model (EDM).[60] Both DPLA MAP and EDM are based on the Resource Description Framework (RDF) and reuse a number of existing vocabularies and ontologies, including Dublin Core Terms, the DCMI Type Vocabulary, OAI-Object Reuse and Exchange (OAI-ORE), and the Simple Knowledge Organization System (SKOS).

The DPLA mapping process identifies elements in the incoming metadata provided by Hubs and maps it into the properties and classes defined by the DPLA MAP. This allows DPLA to ensure that all aggregated metadata has a consistent underlying model that also allows us to integrate that metadata with other linked data sources. Because the DPLA MAP is designed to cover metadata from multiple providers, the mapping process must transform metadata from the elements in each provider's data to the corresponding properties within the MAP. Even though some Hubs use the same schemas as others (e.g., MODS), each

58 Digital Public Library of America, "An Introduction to the DPLA Metadata Model." http://dp.la/info/wp-content/uploads/2015/03/Intro_to_DPLA_metadata_model.pdf, captured at https://perma.cc/96RG-A4A8.

59 See Karen Coyle and Thomas Baker, "Guidelines for Dublin Core Application Profiles," Dublin Core Metadata Initiative, May 18, 2009, http://dublincore.org/documents/profile-guidelines/, captured at https://perma.cc/B9WQ-7HWQ.

60 Antoine Isaac, ed., "Europeana Data Model Primer," Europeana, July 14, 2013, http://pro.europeana.eu/files/Europeana_Professional/Share_your_data/Technical_requirements/EDM_Documentation/EDM_Primer_130714.pdf, captured at https://perma.cc/7UXB-TGUH.

Hub may use a particular schema somewhat differently than another. Accordingly, although DPLA usually indicates a preferred mapping for each major type of schema received in harvesting, it is often necessary to have distinct mappings for each provider.

Once the metadata is mapped, it passes through an additional set of steps referred to as *enrichments*, which allow us to both address consistency issues in the metadata and to provide targeted enhancements to specific properties. Many of these enrichments can be categorized as global cleanup of values to address minor differences in capitalization, punctuation, or whitespace. For specific properties, DPLA also aligns terms received in Hub metadata against small controlled vocabularies such as the DCMI Type Vocabulary or controlled lists of language names. The DPLA ingestion system also undertakes more complex transformations, such as normalizing dates to standardized formats when possible and splitting strings based on a given delimiter (e.g., a semicolon) to yield multiple values. Finally, these enrichments include a geocoding process that takes place-names identified in Hub metadata and compares them against the Geonames dataset, allowing Geonames URIs to be associated with those places when they match. After the enrichment processes are complete, the resulting DPLA MAP records are indexed into a search engine used by the DPLA API.

DPLA's Technical Design Philosophy

The underlying design philosophy for the DPLA API is to emphasize its ease of use and adoption, allowing beginning API users to become productive quickly. In the DPLA API, the MAP-compliant metadata is stored and presented as JSON-LD,[61] a representation of RDF that makes it easier for developers to work with RDF data natively or to ignore the complexity of the underlying model if they choose. The use of JSON-LD is another intentional design choice made for the DPLA API to make it easier for developers to reuse our data without requiring them to be experts in cultural heritage metadata. This is particularly important for DPLA, as it allows the organization to more easily encourage experimentation and use of the DPLA API and the metadata it contains.

61 "JSON for Linking Data," http://json-ld.org/, captured at https://perma.cc/8XFH-6UBV.

This design philosophy also extends to the DPLA MAP. Because the MAP is based on EDM and reuses a number of widely used vocabularies, API users with even a passing familiarity with other metadata standards will find it straightforward to understand. Vocabulary reuse in many cases should be a conscious decision, and in the case of DPLA, this ensures that the MAP is broadly interoperable with other communities of practice and reusable by other communities. In addition, DPLA has seen other institutions, such as the University of British Columbia,[62] begin to reuse and adapt the DPLA MAP for their own purposes. As such, DPLA aspires to follow existing best practices for linked data by ensuring that the DPLA MAP both reuses existing properties and classes from other vocabularies whenever possible and is reusable by others.[63]

Although seemingly complex, the DPLA MAP distinguishes the cultural heritage object itself (the dpla:SourceResource) from the digital representations of that object (the edm:WebResource). In addition, these are further distinguished from the abstract object, which aggregates the object and its representations (the ore:Aggregation).[64] This makes it easier to distinguish between the metadata about each of these three types of resources and allows users of the API to adjust their queries accordingly if they care about filtering on certain aspects of each kind of resource.

Most important, DPLA's commitment to openness has been a strong guiding philosophy in the development of DPLA overall, as well as the API. Although the DPLA currently requires all users of the API to register for an API key to make requests, we ensure that the registration process remains simple. Furthermore, DPLA presumes all users of the API have good intentions and do not enforce rate limiting. In the nearly three years since the public launch of DPLA, there has been no intentional abuse of the API, and as such, no users have been blocked from accessing it. DPLA is also interested in reducing the need for API keys when making single-item requests, ensuring that users just

62 "Open Collections: Metadata Terms," University of British Columbia Library, https://open.library.ubc.ca/terms, captured at https://perma.cc/B5BR-J82M.

63 See Bernadette Hyland, Ghislain Atemezing, and Boris Villazón-Terrazas, eds., "Standard Vocabularies," in *Best Practices for Publishing Linked Data,* W3C Working Group Note, January 9, 2014.

64 Digital Public Library of America, "Metadata Application Profile, version 4.0," http://dp.la/info/wp-content/uploads/2015/03/MAPv4.pdf, captured at https://perma.cc/6GKQ-PFZ8.

getting started with the API will have minimal barriers. In addition, the simple design of the API, which uses basic HTTP requests with parameters, also makes it easy for beginning API users to make queries using just a web browser once they have an API key. Finally, as part of the contributor agreement, all Hubs whose metadata is aggregated by DPLA provide that metadata under either a Creative Commons CC0 license[65] or place it in the public domain. This ensures that the metadata is both freely reusable and can be enhanced by DPLA as well as anyone interested in reusing the metadata.

Conclusion

Overall, DPLA remains dedicated to ensuring that the metadata it aggregates and that is provided by its network of Hubs remains broadly reusable and that its API serves as a platform for enabling new and transformative uses of digital cultural heritage. As an organization, DPLA encourages the users of its APIs to report on the projects and applications built on it or that reuse our metadata. DPLA also appreciates additional feedback on the API itself, as well as its supporting documentation, to ensure that it remains straightforward to use and that it meets the needs of software developers and other users.

65 Creative Commons, "About CC0: No Rights Reserved," https://creativecommons.org/share-your-work/public-domain/cc0/, captured at https://perma.cc/N3FZ-33NS.