DPLA Prototype Platform Overview [*excerpts*]

Who is the DPLA platform for?

The DPLA prototype platform is a technical infrastructure designed to enable developers to build applications that use the metadata aggregated by the DPLA. With this platform, (i) developers (including beta sprinters) will be able to create innovative applications, (ii) the information and services of the DPLA will be integrated into existing sites, and (iii) an end-user portal/home for the DPLA will be supported.

Overview of how it will work

The DPLA prototype platform is a metadata server. It will aggregate and manage metadata about content available online and in local physical institutions (libraries, archives, museums, etc.).

The prototype platform's success depends to a large degree on how much metadata it can gather, and how useful it can make that metadata. It will compile a meta-catalog, and it will provide developer access to what mappings it can to central schemas, but will also permit access through the API to the submitted metadata in its native structure.

Some scenarios

* External developer builds an app that lets children browse the DPLA for appropriate content
* External developer creates a commercial library analytics package
* Public library wants to share metadata with the DPLA but lacks technical resources
* Public library enhances its OPAC with data, including multi-library usage data, from the DPLA
* Public library system enhances its library’s efficiency by analyzing data
* A major collection of historic material wants its content to be found when users of the DPLA search for related items
* Local historical society wants a tool to mount a small online exhibition, and have it become part of the DPLA
* Museum enhances the experience of real-world visitors by displaying information on mobile devices that is relevant to items on display
* Wikipedia creates a page for each item cited, using some metadata from the DPLA
* External developer builds an OPAC that incorporates social network elements
* University enhances its user-generated content with UGC from other libraries
* Independent researcher performs computationally complex analysis using DPLA data

Metadata

The sources of metadata will include (pending decisions by DPLA):

* Collections of online books, music, videos, etc. from libraries and other curators
* Curated online collections of multiple media types (documents, photos, recordings, etc.)
* Bibliographic, holdings and usage data from participating libraries, archives, and museums
* Digitizations sponsored by the DPLA
* Select Web content collections (e.g., TED Talks, Open Courseware)

The types of metadata will include:

* *Items*: metadata about particular items in collections. This includes books, photos, multi-media, etc., as well as any Web content cataloged by the prototype platform.
* *Collections*: metadata about collections, including user-generated
* *Events*: Circulation and other usage information from affiliated libraries, usage of the DP.LA site, registered collections, indicators of the use of Web content, etc.
* *Creators*: Pointers to authority files to identify the creators of works
* *Contributors*: Contributing libraries, archives, museums, Web services, etc.
* *Users*: Possibly retain information about the preference, activities, and created content (e.g. reviews, ratings) of users of DPLA platform-based services, for users who opt-in. Possibly track social graphs also.

As re-deployable open source software

The DPLA prototype platform will be available as open source software for use by local institutions that import their own metadata and content. For example, a state library system might use it to pull together metadata from across its multiple libraries, creating local services and performing cross-library analytics.

Any node using the DPLA prototype platform locally will find it very easy to keep their own metadata updated at the central version of the platform, and will also find it easy to take advantage of the metadata and services offered by that central platform.

Value and trade-offs

The prototype platform wants to lower the hurdle to importing many different types of metadata from libraries, archives, museums, and other curating institutions. When possible, it will harvest data from aggregation nodes. Whenever feasible, the metadata will be put into a slightly extended version of Dublin Core. This will enable cross-collection and cross-type querying, although of in many cases there may be just a few shreds of metadata. The API will provide direct and easy access to the Dublin Core values.

The metadata that does not fit into the Dublin Core will be maintained by the platform, and will be accessible to developers who know the particularities of the original schema. Those schemas will themselves be available via the API.

The platform prototype’s metadata will be unevenly distributed. The platform will include it all, in its raw state, on the grounds that some developer might find it useful. But this disparity in richness of metadata creates issues, such as skewing query results in unhelpful ways.

At the same time, the platform will do what it can to enhance the utility of the metadata it’s ingested. It will take as a continuing project the use of algorithms to semantically associate its content so that queries through the API return more of the value. This might involve more inclusive normalizing techniques, vocabulary matching, entity grouping at various abstraction levels, uniform title matching, and using linked data to build “metadata clouds” around materials that have little native metadata associated with them. These techniques will, we hope, incrementally enrich the metadata so that queries through the API will return increasingly useful results. But we also understand it is an enormously difficult problem.

How will it get built?

The scope of this project is, of course, gigantic. We will build it incrementally and agilely, with frequent builds, engaging a community of developers, using existing software whenever possible.