Analyzing and Normalizing Type Metadata

Originally created December 2018
Updated May-June 2019

Joshua D. Lynch, Metadata Services Specialist, Illinois Digital Heritage Hub
Jessica Gibson, Senior Application Support Coordinator, Consortium of Academic and Research Libraries in Illinois

1. Background: Problems with Type Metadata and Existing Quality Controls

1.1. Introduction to the IDHH and Type Metadata

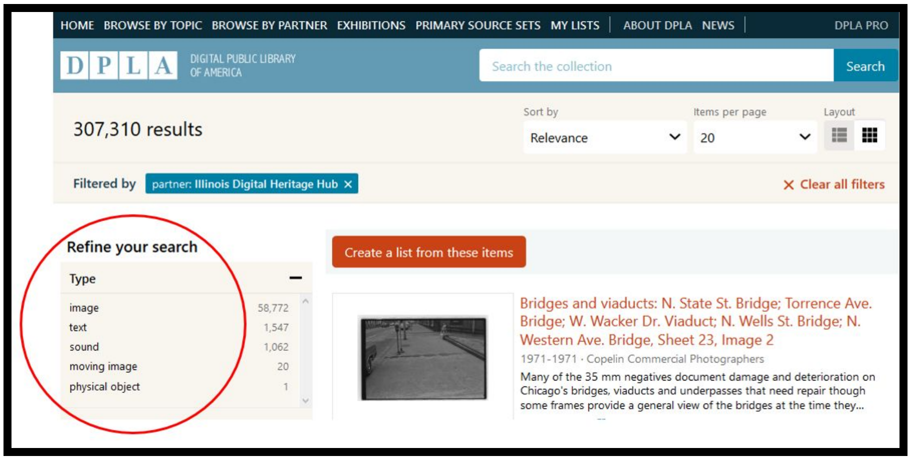
The Illinois Digital Heritage Hub (IDHH) gathers and enhances metadata from contributing institutions from around Illinois and provides this metadata to the Digital Public Library of America (DPLA) for greater access. The IDHH helps contributors shape their metadata to the standards recommended and required by the DPLA. This white paper describes the process for gathering and analyzing contributor's Type metadata, problems seen in the analysis of the metadata, and ways that contributors, the IDHH, and the DPLA can remediate these problems.

The IDHH is the Illinois Service Hub for the Digital Public Library of America (DPLA). The IDHH is a collaboration among the Illinois State Library (ISL), the Consortium of Academic and Research Libraries of Illinois (CARLI), the Chicago Public Library (CPL), and the University of Illinois Urbana-Champaign (UIUC) Library and, as of December 2018, aggregated metadata from more than 140 individual contributing institutions from around the state. The IDHH harvests metadata as Qualified Dublin Core (QDC)[1] and provides QDC records to the DPLA.

As of December 2018, the IDHH had 440 collections and more than 307,000 items. Through metadata assessment[2], a best practice document[3], and training, the IDHH has made significant efforts so that IDHH's metadata conforms to the DPLA's standards and recommendations that ultimately improve discoverability of the IDHH's rich and unique digital resources.

As in any aggregated environment, type metadata has been particularly difficult to standardize across the IDHH's contributing institutions. The data is very diverse and the requirements by the DPLA are strict; only DCMI Type[4] values, such as 'Image', 'Moving Image', 'Physical Object', 'Sound', and 'Text' are used to facilitate faceting in the DPLA's search interfaces. Type is one of several data fields the DPLA uses to create facets by which to narrow search results and to link from an individual record to a list of items of a given type. These features only work when valid DCMI Type values are present in a record.

[1] http://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[2] https://docs.google.com/document/d/1I46jjoehq5KI78VNWRBjR6a6DpLZIN_Xv7IaCD6lB3w/edit
[3] https://docs.google.com/document/d/12NYtWUO9WNzewoXcFuKgpLsCvFA4fj2VCx9fAiiP1PE/edit
[4] http://dublincore.org/documents/dcmi-type-vocabulary/



307,310 results

Sort by: Relevance

Items per page: 20

Filtered by partner: Illinois Digital Heritage Hub

Type

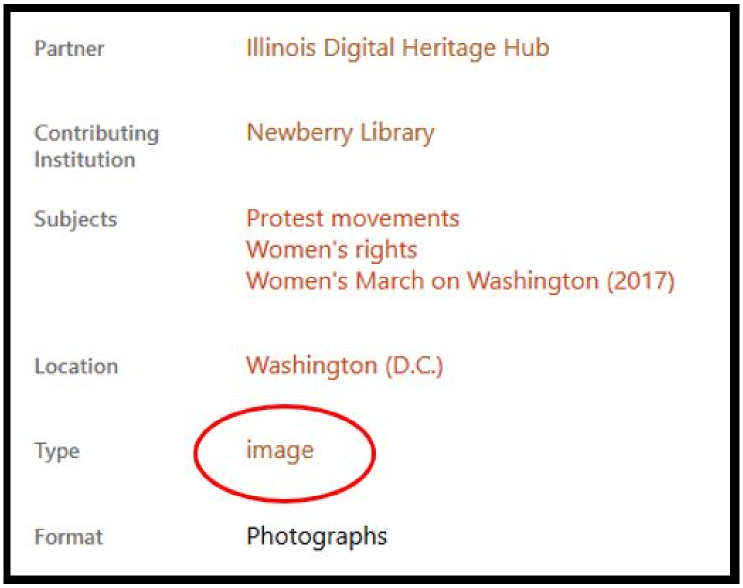
Type	
image	58,772
text	1,547
sound	1,062
moving image	20
physical object	1

Bridges and viaducts: N. State St. Bridge; Torrence Ave. Bridge; W. Wacker Dr. Viaduct; N. Wells St. Bridge; N. Western Ave. Bridge, Sheet 23, Image 2

1971-1971 · Copelin Commercial Photographers

Many of the 35 mm negatives document damage and deterioration on Chicago's bridges, viaducts and underpasses that need repair though some frames provide a general view of the bridges at the time they...

Figure 1: Type faceting options that appear with search results



Partner	Illinois Digital Heritage Hub
Contributing Institution	Newberry Library
Subjects	Protest movements Women's rights Women's March on Washington (2017)
Location	Washington (D.C.)
Type	image
Format	Photographs

Figure 2: Type facet within a record. Clicking 'image' will link to other records in the DPLA with the same Type metadata.

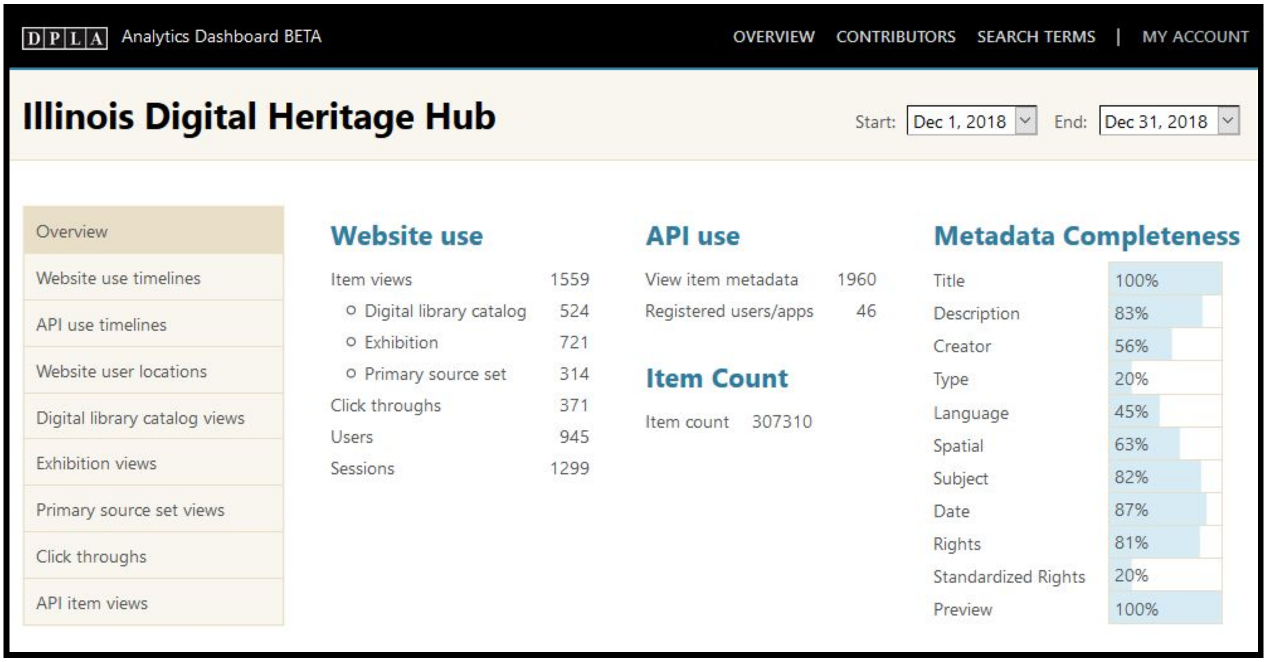
Due to its importance for access, discovery and interactivity, and the low rates of completion of the Type field by IDHH contributors seen in the DPLA's search interface[5], it was determined that work was necessary to try to enhance the Type metadata contributed by IDHH institutions. Analysis revealed that there are more than 550 unique Type values in all IDHH records[6]. The IDHH also learned of limitations of the DPLA's current quality controls, informally known as Ingestion 1[7], although expects that the DPLA's controls will improve in the future. After metadata analysis, a three-pronged approach to enhancing Type metadata was undertaken. Type metadata contributed by IDHH institutions could be improved by:

[5] https://dp.la/search?partner=%22Illinois%20Digital%20Heritage%20Hub%22
[6] https://docs.google.com/spreadsheets/d/1D47FFeizLyAqyr3YnfRl6sU5wFRDnTyN69mGdS65IyY/edit#gid=973851675
[7] https://github.com/dpla/ingestion

1. institutions submitting better quality metadata,
2. more robust XSL transformations by the IDHH, and
3. working with the DPLA to improve their quality controls on IDHH type metadata which may in turn, contribute to improved quality controls on Type and other fields for other hubs contributing metadata to the DPLA.

1.2. Overview of IDHH Type Metadata Problems

In depth analysis discussed in detail below showed that, as of 2018-12-13, there were at least 558 unique Type metadata values provided and only about half of all values (about 176000 of 327000 values examined) perfectly conformed to DCMI Type. Due to limitations in the DPLA's current ingestion system's quality control, only about 61000 metadata values, about one for every five records that IDHH institutions have contributed to the DPLA, display and function as working facets in the DPLA's search interfaces.



Illinois Digital Heritage Hub

Start: Dec 1, 2018 End: Dec 31, 2018

Website use

Item views	1559
○ Digital library catalog	524
○ Exhibition	721
○ Primary source set	314
Click throughs	371
Users	945
Sessions	1299

API use

View item metadata	1960
Registered users/apps	46

Item Count

Item count	307310

Metadata Completeness

Title	100%
Description	83%
Creator	56%
Type	20%
Language	45%
Spatial	63%
Subject	82%
Date	87%
Rights	81%
Standardized Rights	20%
Preview	100%

Figure 3: Low rates of Type completion indicated by DPLA Analytics Dashboard, 2018-12-14.

Some examples of problems in the type metadata originally provided by contributors included a wide variation of capitalizations ("Still Image" vs. "Still image" vs. "still image"), highly granular values with slight variations of the same essential type or format ("Letter - carbon typescript", "Letter - typescript", "letter", etc.), large numbers of multiple delimited values in a field (e.g., "Text; Image", "text; sound", "Text; Still Image"), superfluous delimiters appearing at the end of values or value sets ("Still Image;", "Text; Image;"), a mix of semicolon- and comma-separated values, pluralizations (Books vs. Book), and values that are otherwise misspelled or are better expressed in fields other than Type field.

1.3. Existing IDHH Type Metadata Quality Controls

Before work could be done improving IDHH quality controls on Type metadata, already-existing controls needed to be examined. The IDHH conducts normalization using Extensible Stylesheet Language (XSL) transformations on QDC metadata formatted in XML harvested from contributing institutions via the Open Access Initiative Protocol for Metadata Harvesting version 2.0 (OAI-PMH 2.0). Harvesting QDC metadata from contributors, transforming to the DPLA's standards, and exposing QDC metadata to the DPLA are facilitated by a REPOX version 2.3.7 aggregation server.

Among the XSL templates used to normalize IDHH metadata was a template used through the end of 2018 and until the implementation of the Type XSL templates that came out of this project, described below. The older template matched to specific Type values, especially digital format information like file types, and deleted them as they do not describe the DPLA-recommended analog or physical Type of the item.

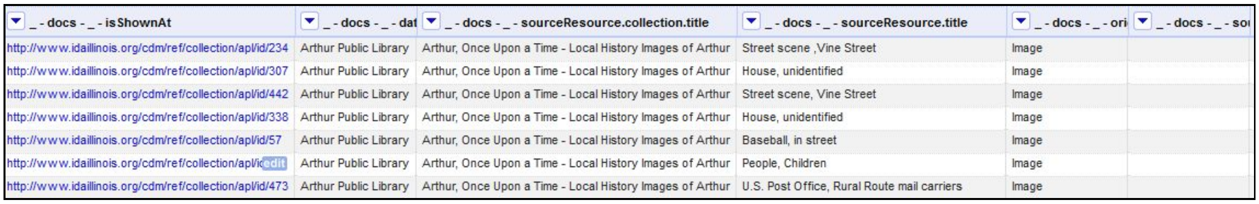
1.4. DPLA Metadata Quality Controls

1.4.1. DPLA Ingestion 1

At the time of the IDHH Type metadata project, the DPLA used Ingestion 1 for harvesting XML from service hubs like the IDHH, transforming it into JSON-LD, and enhancing it. Ingestion 1's python code includes modules for creating and enriching Type metadata based on providers' dc:type[8] and dcterms:medium and dc:format[9] fields values. It was not entirely clear to the authors or to the DPLA staff consulted at the time of the initial type metadata analysis what in Ingestion 1's code may contribute to the low rates of Type completion by IDHH collections.

There were several unexpected ways in which Ingestion 1 was handling Type values. These included:

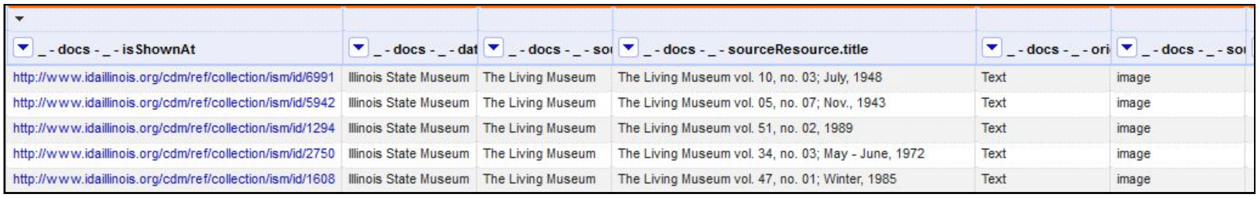
1. not mapping values (even those that seemed to perfectly conform to DCMI Type or only had very slight variations in letter-case) resulting in no Type values in the DPLA catalog:



▼ _ - docs - _ - isShownAt	▼ _ - docs - _ - dat	▼ _ - docs - _ - sourceResource.collection.title	▼ _ - docs - _ - sourceResource.title	▼ _ - docs - _ - ori	▼ _ - docs - _ - so
http://www.idaillinois.org/cdm/ref/collection/apl/id/234	Arthur Public Library	Arthur, Once Upon a Time - Local History Images of Arthur	Street scene ,Vine Street	Image	
http://www.idaillinois.org/cdm/ref/collection/apl/id/307	Arthur Public Library	Arthur, Once Upon a Time - Local History Images of Arthur	House, unidentified	Image	
http://www.idaillinois.org/cdm/ref/collection/apl/id/442	Arthur Public Library	Arthur, Once Upon a Time - Local History Images of Arthur	Street scene, Vine Street	Image	
http://www.idaillinois.org/cdm/ref/collection/apl/id/338	Arthur Public Library	Arthur, Once Upon a Time - Local History Images of Arthur	House, unidentified	Image	
http://www.idaillinois.org/cdm/ref/collection/apl/id/57	Arthur Public Library	Arthur, Once Upon a Time - Local History Images of Arthur	Baseball, in street	Image	
http://www.idaillinois.org/cdm/ref/collection/apl/icedit	Arthur Public Library	Arthur, Once Upon a Time - Local History Images of Arthur	People, Children	Image	
http://www.idaillinois.org/cdm/ref/collection/apl/id/473	Arthur Public Library	Arthur, Once Upon a Time - Local History Images of Arthur	U.S. Post Office, Rural Route mail carriers	Image	

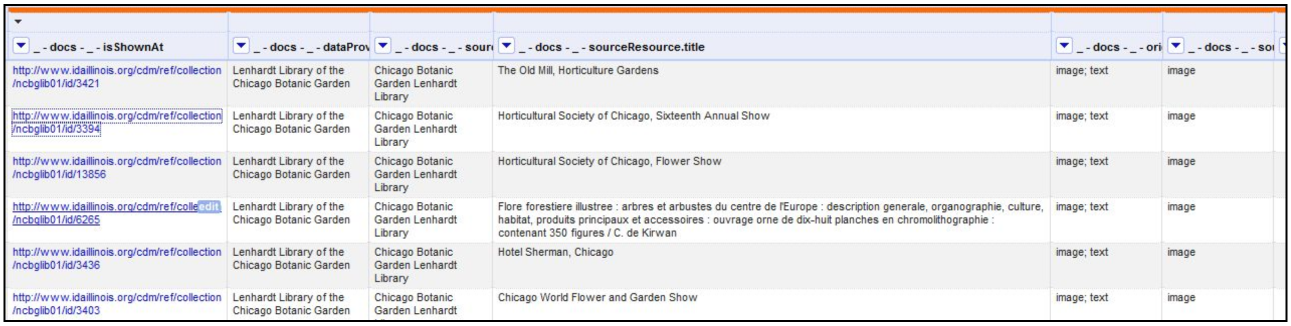
[8] https://github.com/dpla/ingestion/blob/develop/lib/akamod/scdl_format_to_type.py
[9] https://github.com/dpla/ingestion/blob/develop/lib/akamod/enrich-format.py

2. transforming provided DCMI Type values into different values (see for instance, examples of items provided with an original Type value of ‘Text’ mapped to ‘Image’):



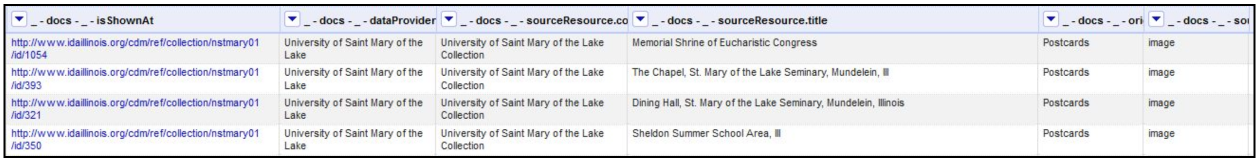
_ - docs - _ - isShownAt	_ - docs - _ - dat	_ - docs - _ - sou	_ - docs - _ - sourceResource.title	_ - docs - _ - ori	_ - docs - _ - sou
http://www.idaillinois.org/cdm/ref/collection/ism/id/6991	Illinois State Museum	The Living Museum	The Living Museum vol. 10, no. 03; July, 1948	Text	image
http://www.idaillinois.org/cdm/ref/collection/ism/id/5942	Illinois State Museum	The Living Museum	The Living Museum vol. 05, no. 07; Nov., 1943	Text	image
http://www.idaillinois.org/cdm/ref/collection/ism/id/1294	Illinois State Museum	The Living Museum	The Living Museum vol. 51, no. 02, 1989	Text	image
http://www.idaillinois.org/cdm/ref/collection/ism/id/2750	Illinois State Museum	The Living Museum	The Living Museum vol. 34, no. 03; May - June, 1972	Text	image
http://www.idaillinois.org/cdm/ref/collection/ism/id/1608	Illinois State Museum	The Living Museum	The Living Museum vol. 47, no. 01; Winter, 1985	Text	image

3. and transforming only one value out of several provided:



_ - docs - _ - isShownAt	_ - docs - _ - dataProv	_ - docs - _ - sour	_ - docs - _ - sourceResource.title	_ - docs - _ - ori	_ - docs - _ - sou
http://www.idaillinois.org/cdm/ref/collection /ncbglib01/id/3421	Lenhardt Library of the Chicago Botanic Garden	Chicago Botanic Garden Lenhardt Library	The Old Mill, Horticulture Gardens	image; text	image
http://www.idaillinois.org/cdm/ref/collection /ncbglib01/id/3394	Lenhardt Library of the Chicago Botanic Garden	Chicago Botanic Garden Lenhardt Library	Horticultural Society of Chicago, Sixteenth Annual Show	image; text	image
http://www.idaillinois.org/cdm/ref/collection /ncbglib01/id/13856	Lenhardt Library of the Chicago Botanic Garden	Chicago Botanic Garden Lenhardt Library	Horticultural Society of Chicago, Flower Show	image; text	image
http://www.idaillinois.org/cdm/ref/colleedit /ncbglib01/id/6265	Lenhardt Library of the Chicago Botanic Garden	Chicago Botanic Garden Lenhardt Library	Flore forestiere illustree : arbres et arbustes du centre de l'Europe : description generale, organographie, culture, habitat, produits principaux et accessoires : ouvrage orne de dix-huit planches en chromolithographie : contenant 350 figures / C. de Kirwan	image; text	image
http://www.idaillinois.org/cdm/ref/collection /ncbglib01/id/3436	Lenhardt Library of the Chicago Botanic Garden	Chicago Botanic Garden Lenhardt Library	Hotel Sherman, Chicago	image; text	image
http://www.idaillinois.org/cdm/ref/collection /ncbglib01/id/3403	Lenhardt Library of the Chicago Botanic Garden	Chicago Botanic Garden Lenhardt	Chicago World Flower and Garden Show	image; text	image

At other times and in fairness, Ingestion 1 seemed to recognize records with Type values of, for example, “Postcards” or “Still Image” with an appropriate DCMI Type:



_ - docs - _ - isShownAt	_ - docs - _ - dataProvider	_ - docs - _ - sourceResource.co	_ - docs - _ - sourceResource.title	_ - docs - _ - ori	_ - docs - _ - sou
http://www.idaillinois.org/cdm/ref/collection/nstmary01 /id/1054	University of Saint Mary of the Lake	University of Saint Mary of the Lake Collection	Memorial Shrine of Eucharistic Congress	Postcards	image
http://www.idaillinois.org/cdm/ref/collection/nstmary01 /id/393	University of Saint Mary of the Lake	University of Saint Mary of the Lake Collection	The Chapel, St. Mary of the Lake Seminary, Mundelein, Ill	Postcards	image
http://www.idaillinois.org/cdm/ref/collection/nstmary01 /id/321	University of Saint Mary of the Lake	University of Saint Mary of the Lake Collection	Dining Hall, St. Mary of the Lake Seminary, Mundelein, Illinois	Postcards	image
http://www.idaillinois.org/cdm/ref/collection/nstmary01 /id/350	University of Saint Mary of the Lake	University of Saint Mary of the Lake Collection	Sheldon Summer School Area, Ill	Postcards	image

The issues may be explained in part by conflicts between Ingestion 1 modules that map to DCMI Type values based on the provided dc:type values and modules that map DCMI Type values based on dc:format or dcterms:medium. In several thousand records, an institution provided dc:format values such as ‘JPEG’ that should map to ‘Image.’ However, they provided Type values, such as ‘Text’, which are, of course, DCMI Types. However, the value ‘Image’ ended up in the DPLA catalog, suggesting that, in many cases, a format value transformation took precedence over a Type value transformation, even when the Type value was more accurate.

1.4.2. DPLA Ingestion 3

Since the completion of the IDHH Type metadata project, the DPLA has completed the development of a new ingestion system, Ingestion 3[10] that will likely be more effective in

[10] https://github.com/dpla/ingestion3

transforming Type values and, as of June 2019, was in the process of rolling it out and switching hubs from Ingestion 1 to this new platform. Ingestion 3 includes a module[11] with a lengthy list of provided values that will be converted to DCMI Type values based on exact matches.

In late 2018, there was discussion among IDHH staff on whether to wait for the release of Ingestion 3 to conduct metadata analysis and refinement. Ingestion 3's code has been in open development and publicly available on GitHub since 2017 and it was clear that the system provided for more robust transformations and enhancements on Type and many other fields than Ingestion 1. However, the IDHH decided to move forward with the Type project for several reasons:

1) At the time of the IDHH Type metadata project, it was not known for certain when Ingestion 3 would be deployed; a rough timeline was provided within the first half of 2019.
2) The IDHH planned to wrap up the metadata analysis and enhancement phase of the project and begin work toward better access of IDHH resources and outreach to potential users by the end of 2018, long before Ingestion 3 was to rollout and the initial results for IDHH records could be examined
3) Moreover, as powerful as Ingestion 3 seemed, there was concern that XSL transformations by the IDHH could be even more thorough. In particular, the exact matches required in Ingestion 3's module for transforming to DCMI Type may miss numerous values. For instance, there is a transformation for all values converted to lowercase equal to "engraving" but not "engravings", a pluralization that is among the most common values in IDHH-contributed data.

Both Ingestion 1 and Ingestion 3 are capable of parsing multiple delimited values in individual fields and de-duplicate fields. However, these processes must be enabled for particular institutions and the field that needs de-duplicating must be specified. Although it was ultimately decided to use XSL to refine and separate delimited values into multiple instances of the dc:type field before providing them to the DPLA (discussed below), informing the DPLA of certain properties of contributed metadata is important for the best performance of the DPLA's ingestion systems and for a hub's records to be transformed and display as expected in the DPLA catalog.

2. Inspecting IDHH Type Metadata

Initial analysis using DPLA Analytics and OAI-PMH metadata analysis tools[12], originally developed by North Carolina Digital Heritage Center and modified by the IDHH, scratched the surface of what appeared to be widespread issues with IDHH-contributed Type metadata. It was

[11] https://github.com/dpla/ingestion3/blob/develop/src/main/scala/dpla/ingestion3/enrichments/TypeEnrichment.scala
[12] https://github.com/ncdhc/dpla-aggregation-tools

then determined that more expansive and detailed analysis of the IDHH Type metadata was necessary.

Due to their out-of-the-box limitation of not being able to evaluate more than one record set at a time, large-scale data analysis could not be easily accomplished with the metadata analysis tools. Moreover, several attempts from 2017 to 2018 by IDHH harvesting coordinator Jessica Gibson, along with DPLA's former IDHH contact and metadata coordinator, Gretchen Gueguen, did not succeed in analyzing metadata via the IDHH entire record set. These attempts included gathering all of XML in a single dataset and attempting to analyze it with OpenRefine. Due to the size of the set and possibly insufficient available memory on the machines facilitating this effort, OpenRefine failed to parse the dataset after multiple attempts. Moreover, although OpenRefine supports gathering and analyzing XML through OAI-PMH queries, it was unknown to IDHH staff even after extensive research, if and how OpenRefine can handle resumption tokens; otherwise, the only method known to IDHH staff for dealing with huge OAI-PMH sets with OpenRefine is manually entering resumption tokens, an incredibly tedious process.

Several alternative methods were investigated in order to gather and analyze all the IDHH Type metadata, including bulk download[13] of JSON-LD data created by the DPLA from the XML harvested from the IDHH, or calling the DPLA's API[14] for smaller samples of JSON-LD data. The last option was selected primarily because data analysis tools like OpenRefine (discussed below) work better with JSON data than with XML and gathering data via the API would provide the most up-to-date data available from the DPLA in a fairly small package (around 100MB compared to the 1GB bulk download of all IDHH metadata, the latest of which as of 2018-12-17 was dated 2018-06).

In order to gather a comprehensive sample of data, a Python script[15] was created to generate DPLA API calls and to download all the JSON-LD data for each DPLA data provider in the IDHH Registry[16], as read from a UTF-8 csv file. It may seem more obvious to use the API to harvest all records that contain the name or ID of the hub in the provider metadata. The problem with this approach is that the API limits calls to 100 pages of results or a maximum of 40,000 records. As only one of the IDHH's providers currently contributes more than 40,000 records and its metadata already well conforms to DCMI Type, it was decided to gather collections by provider in order to circumvent the API's limit. The following fields were selected from each record in each collection:

Field Name	***Description***
dataProvider	Individual institution that provided record

[13] http://dpla-provider-export.s3.amazonaws.com/
[14] https://pro.dp.la/developers/api-codex
[15] https://github.com/jlynch2121/dpla-api-json-downloader
[16] https://docs.google.com/spreadsheets/d/1OcID6-Ha8c5-ih0MZT8gOgsv-qixDd9cBxCi5xQd1ls/edit#gid=911519317

sourceResource.collection.title	Title of collection of which a given record is a part
sourceResource.title	Title of the item an individual record describes
isShownAt	URL linking to the local instance of the record
originalRecord.type	Original Type value provided by IDHH
sourceResource.type	DCMI Type value created by the DPLA

The dataProvider and sourceResource.collection.title fields were selected in order to identify the contributing institutions and collections of which individual records were a part. This allowed for faceting in OpenRefine by the institution that provided a record and by a record's collection and thus, checking if problems with particular records represented more widespread issues at the collection or institution level. Similarly, sourceResource.title was chosen to identify an individual record and isShownAt was selected for easily linking to a record's source instance. These fields allowed for greater refinement, especially in terms of faceting, grouping, and filtering the two type metadata fields available in the JSON-LD sample: the originalRecord.type, or the value provided by the IDHH to the DPLA, and the sourceResource.type, the value created by the DPLA and available in its search interface if and when provider data has passed through the DPLA's quality controls, discussed above.

Once the JSON-LD data was downloaded, it was then plugged into OpenRefine 3.0. The initial OpenRefine table revealed that many records had two instances of the Type metadata field, each with a corresponding value. Further refinement was required in order to prioritize action items from the overwhelming results and focused on gathering the facets and corresponding counts into another OpenRefine sheet, removing duplicates between the two instances of Type fields, and summing the total counts of the duplicate values. The final product was a spreadsheet of 558 type values and their respective counts from all IDHH collections.

The number of unique Type values revealed problems that could not be solved by the IDHH through XSL transformation alone. In addition to mass-normalization of common values by the IDHH, two other categories of issues were identified: those that would need to be redressed by the DPLA and those that would simply need to be tackled by individual contributors refining the metadata they provide.

3. Solutions for Data Clean Up

3.1. IDHH XSL Enhancements

Much of the work of enhancing Type metadata can be handled automatically through XSL stylesheet transformations applied in the IDHH REPOX aggregation server. An XSL template was developed[17] for matching on the dc:type element and performing operations on the text node, e.g., the Type metadata value the dc:type element contains.

3.1.1. Problems that can be addressed with XSL

Normalizing Different Delimiter Combinations
Many providers use commas or a combination of comma-separated values and, the more common delimiter, semicolons. Occasionally, commas and semicolons appear in the same field. In the XSL templates for transforming type, a line of code utilizing the XPath replace() function converts commas to semicolons.

Normalizing Each Value in a Field Containing Several Delimited Values
The tokenize() XPath2 expression is used to split multiple values and then, each token is inspected separately using a for-each loop which calls up a template for matching a token to the correct DCMI Type. While the DPLA is also able to parse values from delimited lists and attempts to match each separate term to its DCMI Type, it will likely miss many more unusual values as discussed below, especially before Ingestion 3 is deployed.

Catching Many Similarly-Named Values
The DPLA's Ingestion 3 uses exact matching and thus, may miss many idiosyncratic values, especially pluralizations, odd spellings, and letter case combinations. Thus, even if the DPLA is capable of identifying and converting the term "Photograph", values like "Photographs", of which there are nearly 30,000 among IDHH contributions, may be missed.To address this, the IDHH XSL converts all strings to lowercase. Instead of relying on an exact match on a certain string value, the XSL transforms based on if this lowercase string contains a certain keyword, using the XPath lower-case() and contains() functions. When the lower-case version of a record's dc:type value contains a keyword such as "still," as in the half dozen or so common variations of "still image", this text node is to be transformed into 'Image'.

3.1.2. XSLT Limitations

Duplications

[17] https://github.com/jlynch2121/idhh-xsl

XSL works on individual nodes. Thus, if multiple dc:type elements are present, each will be passed through the normalization function. Some records will have Type values that will be transformed into the same value: for example,

```
<dc:type>Poster</dc:type>
<dc:type>War Poster</dc:type>
```

Will be transformed into:

```
<dc:type>Image</dc:type>
<dc:type>Image</dc:type>
```

Fortunately, the DPLA can de-duplicate specific fields and has been asked to do so for the IDHH's Type metadata.

Inadvertent Value Conversion

The keyword matches utilizing contains() described above will require continued careful attention to contributors' type metadata to insure that the keywords that the type values contain are transformed properly. Periodic tests, the latest of which at the writing of this paper was in 2019-05, show that the transformation has worked well for nearly half a year. However, there may eventually be newly-added type values that contain a certain keyword which will be converted to incorrect Type values.

3.2. Clean up by Individual Institutions

Many problems boil down to unusual values appearing in the Type metadata field. Examples include phrases that appear to be more suited for an item description (dc:description) field or subject (dc:subject), URLs that seem to be duplicate record identifiers, rights metadata, and odd misspellings of common type values. None of these issues can be efficiently dealt with via XSL transformations as there is either no clear Type value to transform, or due to the inefficiency of checking numerous variant misspellings of even the most common Type values.

Instead, the IDHH respectfully engaged contributors and provided encouragement and instruction on remediating Type and other metadata values. This message included the purpose and significance of Type metadata in the DPLA, an overview of the DPLA's selected controlled vocabulary, DCMI Type, as well as the steps the IDHH is taking to ease at least some of the burden on contributors in conforming to the standards of DCMI Type.

To help contributing institutions understand the issue and best practices for the Type metadata, an online Type metadata guide[18] for contributors was created and linked to on several other online platforms, including the IDHH LibGuide[19].

[18] https://ildplametadatawrkgrp.wordpress.com/documentation/type/

[19] http://guides.library.illinois.edu/idhh

3.3. DPLA Ingestion 3

Lastly, values that were likely to be transformed by the DPLA's Ingestion 3 were identified and noted. For now, XSL will transform most non-DCMI Type-conforming metadata and is likely to map more values than the DPLA's module.

4. Deploying XSLT and Type Metadata Guide

4.1. Testing and Deploying XSLT

The XSLT was developed in test-and-go fashion on small XML data sets in Oxygen Editor 18.0. On 2019-01-24, immediately prior to the first DPLA ingestion of 2019, XSLT was run for the first time on the REPOX 2.3.7 server. Key features of the transforms were tested by the harvest coordinator and metadata services specialist on several large datasets, including the basic mapping from a provided metadata value to a DCMI type, its effects on pluralized values or alternative spellings, spacings, etc. and its handling of delimited values on two different delimiters (comma and semicolon and various combinations thereof). Tests confirmed the duplication of values predicted in the development of the XSL. However, no unexpected transformations occurred and the XSL was deployed for transformations on all records to be provided to the DPLA on 2019-01-28.

4.2. Disseminating Type Metadata Guide

The Type metadata guide for contributing institutions was disseminated through the first half of 2019. The IDHH Metadata Best Practices was updated to include the guide[20] and it has also been posted on the IDHH metadata working group website[21] along with a list of values the IDHH transforms to DCMI Type. Institutions are encouraged to contact the metadata services specialist at the IDHH if they wish to recommend additions or other changes to the current mapping. On 2019-04-18, the metadata services specialist provided a lightning talk at DPLAFest 2019 on the Type Metadata cleanup project and included links to the Type Metadata Guide and to a draft of this report.

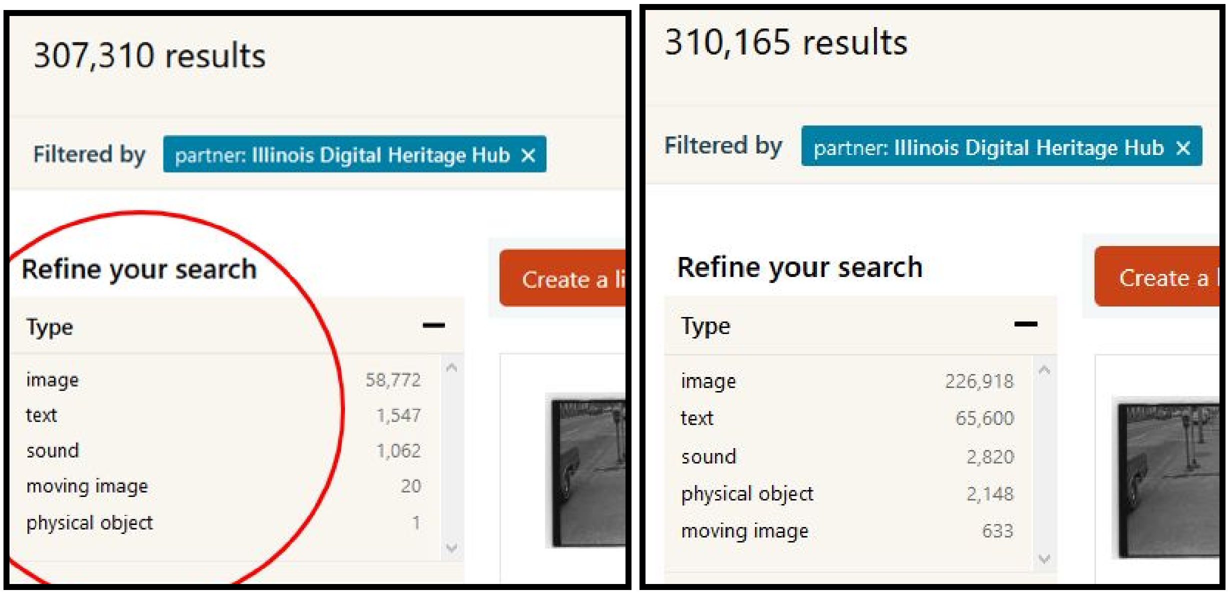
5. Measuring Impact of the Type Metadata Project

Type metadata was rechecked on 2019-02-15 and on 2019-05-20 using the Python metadata downloader tool and OpenRefine. The February check-up revealed that the number of unique

[20] https://docs.google.com/document/d/1q1AORHoa0ey0fUGOTYMHLvZNCm6Wq1Qe9DDvFZSRPT0/edit?usp=sharing

[21] https://ildplametadatawrkgrp.wordpress.com/documentation/type/

metadata fields was reduced to 117, down from the more than 550 before the type metadata project began. The number of DCMI conforming values as of 2019-02-15 was 300,985, whereas there were only 918 non-conforming types, meaning that 99.7% of type metadata values conformed to DCMI Type as of 2019-02. This was up from only 176000 of 327000 values (53.8%) perfectly conforming to DCMI type before the project began and where only about 20% of records, only 61,000 out of the more than 300,000 provided by the IDHH, were showing up with Type values in the DPLA catalog.



307,310 results

Filtered by partner: Illinois Digital Heritage Hub ×

Refine your search

Type	—
image	58,772
text	1,547
sound	1,062
moving image	20
physical object	1

310,165 results

Filtered by partner: Illinois Digital Heritage Hub ×

Refine your search

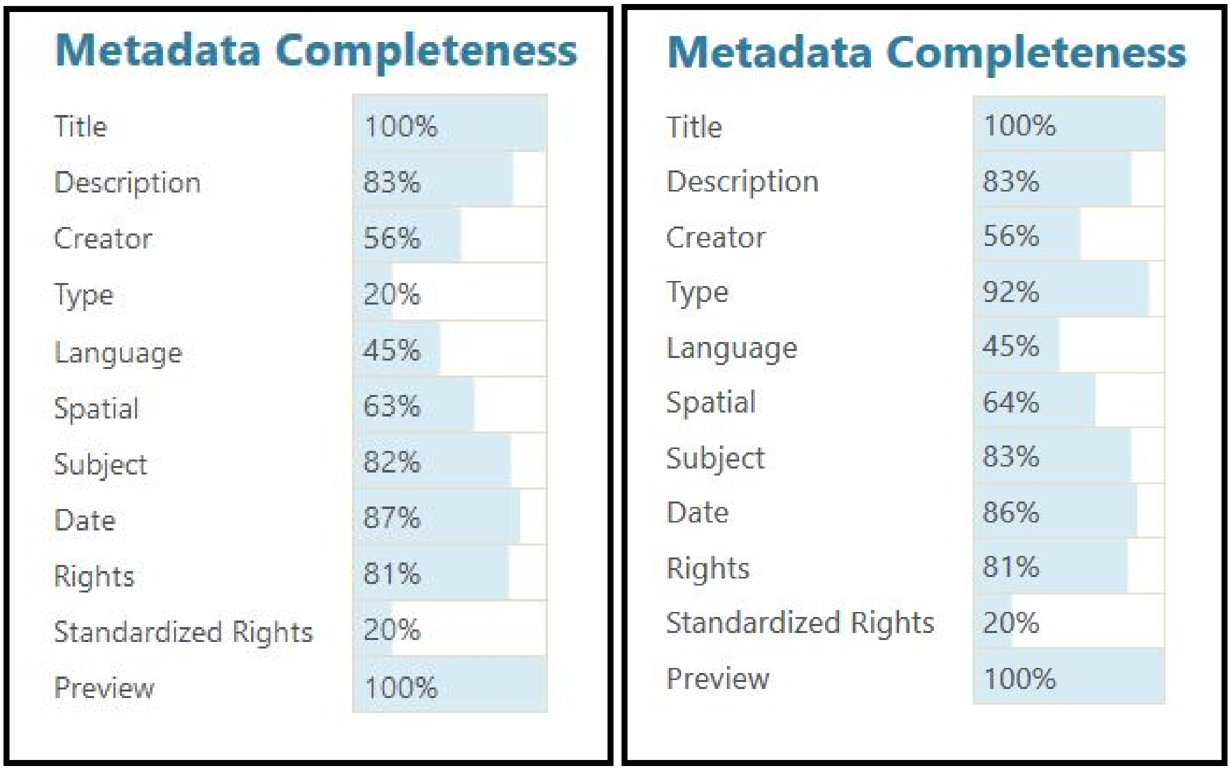
Type	—
image	226,918
text	65,600
sound	2,820
physical object	2,148
moving image	633

Figure 4: Screen captures of type facets for all IDHH items in the DPLA, before (left) and after (right) the type metadata project.

The February check-up also showed there are still a few values that do not conform to DCMI Type leftover from the original list developed in 2018-11. These were generally facets which number less than 10 values each and thus were not made a priority in the project. In addition, there were a few more new values that must have slipped in due to changes provider institutions have made to their metadata since it was initially downloaded and checked in OpenRefine in 2018-11 and 2018-12 and which had not been accounted for by the XSL. Only two of these fields had more than 100 instances, at around 300 and 110, respectively.

The May 2019 check-up revealed once again high percentages of conformity to DCMI Type. Of approximately 308,000 originally provided Type metadata values, 307,000 conformed to DCMI Type, maintaining the 99.7% conformity to DCMI Type seen in the February check-up. There are now only 109 unique fields showing up in the analysis. There are only three non-conforming values with counts greater than 100 and the vast majority of these (over 86.5%) are being transformed by the DPLA's ingestion system.

Both check-ups showed high percentages of completion of the dc:type metadata field and conformity to DCMI Type. This may be allowing the DPLA's own quality control measures to do their best work. Most of the values not transformed by the IDHH's XSL are being transformed by the DPLA's Ingestion system and, therefore, are no cause for concern. It is not recommended at this time that the IDHH change the XSL in order to transform them. As of 2019-05-20, DPLA Analytics reports completion rates of the Type field at 92%, a far cry from the 20% rates seen in December 2018. This would likely be even higher if not for a number of contributors who simply do not provide Type metadata.



Metadata Completeness

Title	100%
Description	83%
Creator	56%
Type	20%
Language	45%
Spatial	63%
Subject	82%
Date	87%
Rights	81%
Standardized Rights	20%
Preview	100%

Metadata Completeness

Title	100%
Description	83%
Creator	56%
Type	92%
Language	45%
Spatial	64%
Subject	83%
Date	86%
Rights	81%
Standardized Rights	20%
Preview	100%

Figure #: Type Metadata Completion Rates, December 2018 (left) and May 2019 (right). Note that Type now has the highest rates of completion of any non-required field.

It should also be noted that none of the problems foreseen by tokenizing multiple semicolon-separated type values manifested in either the February or May check-up. The DPLA was able to de-duplicate Type values, thus making normalizing delimiters and tokenizing them worthwhile. Nevertheless, the IDHH will continue to check up on the Type metadata and transformations from time to time.

6. What's Next

6.1. The Future of Analyzing Provider Data

It is no longer possible to analyze large samples of Type metadata originally provided by contributors using JSON-LD data from the DPLA. Before the XSL was deployed, the DPLA received primarily original values provided by IDHH contributors and therefore, these same values were accessible in the API and through bulk download in the originalRecord.Type field. After the XSL template for transforming Type metadata is deployed, original values provided by contributors will be transformed by the IDHH before they are harvested by the DPLA and thus will not be available through the DPLA. Therefore, in addition to existing tools that allow for the analysis of small samples of contributors' metadata, a mode of analyzing large and, if possible, complete samples of Type and other metadata values directly from contributors will need to be developed. Discussing new methods in depth is beyond the scope of this paper but may involve a change of harvesting platform from REPOX to one that provides more powerful data analysis, such as the Michigan Service Hub's solution now widely adopted by other DPLA partners: Combine[22].

6.2. Continuing to Work with Providers

Type metadata will be a moving target: current providers' metadata may change and new collections with potentially new type metadata values are being added regularly. New collections' metadata will continue to be thoroughly assessed. Moreover, the IDHH been in communication with providers planning to improve their Type metadata, as well as other metadata fields throughout the first half of 2019.

6.3. Continuing to Work with the DPLA

Regular contact with the DPLA will be essential. In addition to the ongoing discussions on the DPLA's role in normalizing provider data, it will be important to discuss any remaining or new issues after Type metadata statistics are re-analyzed. Conversation with DPLA will be especially crucial now that Ingestion 3 has been deployed and the IDHH is about to be migrated to the new system.

[22] https://combine.readthedocs.io/en/master/