An Exploratory Study of the Description Field in the Digital Public Library of America

Hannah Tarver
University of North Texas Libraries, USA
hannah.tarver@unt.edu

Oksana Zavalina
University of North Texas, USA
oksana.zavalina@unt.edu

Mark Phillips
University of North Texas Libraries, USA
mark.phillips@unt.edu

Abstract

This paper presents results of an exploratory quantitative analysis regarding the application of a free-text Description metadata element and data values associated with this element. It uses a dataset containing over 11.6 million item-level metadata records from the Digital Public Library of America (DPLA), originating from a number of institutions that serve as DPLA's content or service hubs. This benchmark study provides empirical quantitative data about the Description fields and their data values at the hub level (e.g., minimum, maximum, and average number of description fields per record; number of records without free-text description fields; length of data values; etc.) and provides general analysis and discussion in relation to the findings.

Keywords: metadata aggregations, metadata values, free-text fields, item descriptions.

1. Introduction and Background

Two kinds of metadata coexist in records created according to various metadata standards: controlled-vocabulary metadata which draws values from formally-maintained list of terms, and free-text metadata which relies on natural language. Free-text metadata -- for example, the Description metadata element in the Dublin Core (DC) metadata scheme; various notes (e.g., 5XX fields) in MARC records; Abstract, Note, and Table of Contents elements in the Metadata Object Description Schema (MODS); Scope and Content elements of the Encoded Archival Description (EAD) metadata scheme; etc. -- have been considered an important part of metadata records as a rich source of information on the nature of information object(s) described by each record.

Best practice recommendations have been developed regarding data values for the Description element and its semantic equivalents in metadata records describing information objects -- Cataloging Cultural Objects (CCO) (Baca et al., 2006), Categories for the Description of Works of Art (CDWA) (Baca et al., 2009), OSU Knowledge Bank Metadata Application Profile for Digital Video (Ohio State University Libraries, 2006) etc. -- as well as in metadata records describing physical collections of manuscripts (National Union Catalog of Manuscript Collections, 2010) and collections of archival materials (OLAC Cataloging Policy Committee, 2002; Encoded Archival Description, 2002, 2015).

Cataloging Cultural Objects (Baca et al., 2006) and Categories for the Description of Works of Art (Baca et al., 2009) suggest recording information about subject, significance, and function in an item-level free-text Description element. OSU Knowledge Bank Metadata Application Profile for Digital Video (Ohio State University Libraries, 2006) recommends inclusion of provenance and history of the work, as well as the nature of the language of the resource. Dublin Core Usage Guide (Hillmann, 2005) provides guidelines on how to use item-level metadata elements; however, it does not detail what information should be included in Description, besides a broad recommendation, "Description may include but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content" (section 4.3).

Several documents discuss specific guidelines in relation to collection-level metadata rather than item-level records. National Union Catalog of Manuscript Collections (2010) suggests that collection-level metadata creators for manuscript collections provide in the Description element: information about types of materials included in the collection; topics with which the materials in the collection deal; geographical areas with which the materials in the collection deal; associated dates, events, and historical periods dealt with by the materials in the collection; names, dates, and biographical identification of persons and names of corporate bodies significant (by quality and/or quantity of material) to the collection; and specific phases of career/activity of the major person or corporate body responsible. Summary Notes for Catalog Records (OLAC, 2002) recommends inclusion of information about specific types and forms of materials present; significant people, topics, places, and events covered; span of dates covered by the collection; history of the work; unique characteristics of the collection; reason and function of the collection; audience; and user interaction. The previous version of Encoded Archival Description (EAD, 2002) recommended inclusion of such characteristics as form and arrangement of materials; significant subjects represented; places represented; events represented; significant organizations and individuals represented; and collection strengths. The current version of Encoded Archival Description (EAD3, 2015) adds a recommendation to provide information functions and activities that generated the materials being described, and gaps in the materials to help the user evaluate the potential relevance of the materials being described.

The guidelines on constructing data values for free-text metadata elements, such as Description, are intended to facilitate users' access to information objects and collections through these rich metadata fields, but the suggestions are not necessarily followed in creation of metadata records. Empirical research focusing on analysis of data values in free-text Description metadata allows researchers not only to determine the level of adherence to guidelines but also, importantly, to categorize information typically found in these data values. For example, two studies of collection-level metadata in large-scale repositories in the United States and Europe (Zavalina et al, 2008; Zavalina, 2012) resulted in the list of 19 properties of a digital collection that are represented in Description fields in collection-level metadata records: topical coverage; geographic coverage; temporal coverage; collection title; size; collection development information; provenance; importance of collection; uniqueness; comprehensiveness; intended audience; navigation and functionality; participating, hosting or contributing institutions; copyright information; frequency of additions to collection; funding sources; types/genres of items; creators of items; and language of items. Several of these categories of information (e.g., creators of items in collection, etc.) did not appear in the existing guidelines for the free-text Description field but were nevertheless included by metadata creators who considered these important for more efficient information access and discovery.

For item-level metadata, several studies looked at frequency of application of Dublin Core metadata elements, including the free-text Description element, in metadata aggregations. For example, in Ward's (2003) study of over 900,000 Dublin Core metadata records harvested from 82 OAIster data providers, it was observed that Description element was included in slightly over a half (50.9 %) of all records and that 72% of data providers included this element in their records. Jackson and colleagues (2008), in their study of metadata harvested into IMLS DCC aggregation , did not report the observed percentage of metadata records that include Description field, but reported its systematic inclusion in the records from 31 (89%) out of 35 harvested digital collections. The findings of these studies show the level of application of Description elements to range substantially.

Other studies measured application of metadata elements, including Description element in DC-based metadata and/or its counterparts from other metadata schemes in individual digital collections. For example, Kurtz's (2010) analysis of metadata applications in three digital repositories hosted by university libraries and using Dublin Core demonstrated fluctuations in the level of Description element usage, from 40% to 75% of metadata records. A study of metadata application in digital video collections (Weagley, Gelches, & Park, 2010) revealed a much higher

level of application of Description metadata element (99% of metadata records, the highest of all elements and at the same level with the Title element) than other studies that measured application of Dublin Core metadata. This might be due to the specific nature of these digital collections. Similar observations were made for three digital image collections in a study (Park, 2006) which found the Description element to be included in all 100% of Dublin Core metadata records across the collections.

1.1. Digital Public Library of America

The Digital Public Library of America (DPLA) is a prominent aggregation of metadata, currently comprising over 13 million metadata records from libraries, archives and museums in the United States to provide free public access. DPLA functions on a distributed network model and consists of a group of national partners or "hubs" providing both content and services (Ma, 2014). Content hubs constitute large libraries, museums, archives and other digital repositories which maintain a one-to-one relationship with DPLA. Service hubs are state, regional, or other collaborations which host, aggregate, or otherwise bring together digital objects from cultural heritage institutions and provide metadata to the DPLA through a single data feed such as OAI-PMH.

The internal data model of DPLA is based on the Resource Description Framework (RDF) and the central descriptive metadata standard employed is the Dublin Core (Mitchell, 2013). In DPLA, some of the metadata gathered from providers is stored along with metadata generated or extracted during the aggregation process. The metadata aggregated and normalized by DPLA is in the public domain and has no copyright restriction; DPLA metadata can be harvested via the OAI-ORE standard for sharing or data analysis. JSON-LD (JavaScript Object Notation-based serialization for Linked Data), an RDF-inspired serialization, is disseminated via API output.

In the most recent version of DPLA metadata documentation, there is an inconsistency regarding the status of the Description property: in the Introduction to version 4 of the DPLA metadata model (Digital Public Library of America, 2015a, p.9), the Description property of the sourceResource class is named a "recommended" metadata element -- i.e., an element that should be included in a metadata record if the information is available -- but in the complete DPLA Metadata Application Profile document (Digital Public Library of America, 2015b, p.20), this property is not included in the listing of required or recommended properties. In DPLA's metadata application profile, which is based on an RDF serialization of the Dublin Core descriptive metadata standard, the DPLA Description element maps to dcterms:description (Digital Public Library of America, 2015a). Native metadata -- metadata used internally by institutions that serve as DPLA hubs -- is often more detailed and relies on richer metadata schemes than Dublin Core, such as MODS or MARCXML. Multiple metadata elements from these metadata schemes (e.g., MODS abstract, tableOfContents, and note; various 5XX MARC fields; etc.) map to a single metadata element (Description) in Dublin Core. Therefore, as a result of normalizing and aggregating native metadata into DPLA, it is likely that metadata records contain multiple Description fields with varying kinds of data values.

The review of the literature demonstrates the lack of recent empirical, quantitative studies of free-text description metadata. The study reported in this paper is one of the first attempts to systematically evaluate this kind of metadata, and the first one to use a very large aggregator such as Digital Public Library of America as its target.

2. Methods

The research questions that guided this exploratory study fall into two areas: (1) What is the overall usage of the Description field by hubs in the DPLA dataset? And (2) How can high-level attributes such as length of data values provide insight into metadata practices regarding the free-text Description metadata field among DPLA hubs?

To address these research questions, we applied the quantitative content analysis research method. Unlike many previous studies of metadata in large-scale digital libraries that analyzed a generalizable sample of metadata records, the authors of this study took a "big data" approach that analyzes the whole dataset and therefore avoids sampling errors. The authors used DPLA's Bulk Download to harvest the metadata dataset (http://dp.la/info/developers/download/). This dataset was parsed into individual records that contain both the original metadata submitted by various DPLA hubs and a normalized version based on the DPLA Metadata Application Profile (http://dp.la/info/developers/map/).

For this analysis, each record was parsed from the DPLA dataset and processed to extract the Description field information, along with the DPLA identifier for the record and the originating provider/hub. The resulting dataset comprises 11,654,800 records. Because the Description field is not required and is repeatable, some records contain no Description values while other records contain multiple instances of the Description field. The original 11,654,800 records in the DPLA dataset contained a total of 17,884,946 description values. Each record was further processed to generate metrics about individual Description field instances. Examples of these metrics include: length of description (number of characters); number of words; average word length; and proportion of description that consists of letters, punctuation, or integers. In total there were 20 descriptive metrics generated for each of the description values in the dataset.

3. Findings

All of the Description field values were loaded into the Apache Solr Full-Text indexer where various components of that system including the facet and the statistics components were used to explore the dataset.

For each analysis, the findings were broken down by hub. A relatively small number (11,422) of records did not include hub source information; for the purposes of maintaining completeness of the dataset, these are categorized as records originating from "undefined provider."

3.1. Usage

The first general analysis included a count of instances of Description values per record (Table 1). Since this field is repeatable and serves as point to which many free-text fields map from the hubs, some records have more than one instance of the Description field.

As shown in Table 1, there is a wide range of usage in the Description field across hubs. In some cases, a large majority of records have no Description field values. These include collections from the National Archives and Records Administration (NARA, 98.83%), Kentucky Digital Library (98.66%), and items with undefined provider (99.89%). On the other end of the spectrum, The Portal to Texas History includes Description fields in 99.98% of its metadata records and several others -- the United States Government Publishing Office (GPO), J. Paul Getty Trust, David Rumsey, and University of Illinois at Urbana-Champaign -- also have at least one Description field value in more than 99% of their records.

The number of Description instances per record also represents a drastic range (see Fig. 1). Eight hubs -- Biodiversity Heritage Library, Empire State Digital Network, Kentucky Digital Library, Minnesota Digital Library, NARA, Tennessee Digital Library, University of Virginia Library, and University of Washington -- have no more than one Description value in any record (see Appendix A for additional statistics). However, some item records contain an extremely large number of values. The Smithsonian Institution has at least one record containing 179 separate Description entries; the Digital Library of Georgia and Indiana Memory each have at least one record with 98 separate entries. While these numbers seem to be outliers on the whole, five other hubs have records containing at least 25 separate Description values: HathiTrust (77), GPO (65), Internet Archive (35), J. Paul Getty Trust (25), and University of Illinois at Urbana-Champaign (25).

Additionally, our analysis considered the total number of Description field instances in metadata records per hub, as well as the percentage of those Description field data values that are unique (Table 1). The three hubs that have less than 1% uniqueness are the same hubs that have few Description field instances in their records: Kentucky Digital Library, NARA, and undefined provider. This suggests that the few records that *do* contain Description field values from these hubs have significant content overlap.

TABLE 1: Distribution of Description field instances in metadata records by hub.

Hub	Records	Records with 0 Description Instances		Records with 1+ Description Instances		Total Instances	Unique Description Values	
artstor	107,665	40,851	37.94%	66,814	62.06%	128,922	34,490	26.75%
bhl	123,472	64,928	52.59%	58,544	47.41%	123,472	46,235	37.45%
cdl	312,573	80,450	25.74%	232,123	74.26%	563,967	300,983	53.37%
david_rumsey	65,244	168	0.26%	65,076	99.74%	166,314	32,093	19.30%
digital-commonwealth	222,102	8,932	4.02%	213,170	95.98%	455,369	110,200	24.20%
digitalnc	281,087	70,583	25.11%	210,504	74.89%	241,224	162,178	67.23%
esdn	197,396	48,660	24.65%	148,736	75.35%	197,396	91,001	46.10%
xgeorgia	373,083	9,344	2.50%	363,739	97.50%	821,067	271,437	33.06%
getty	95,908	229	0.24%	95,679	99.76%	264,268	32,419	12.27%
gpo	158,228	207	0.13%	158,021	99.87%	690,883	208,307	30.15%
harvard	14,112	3,106	22.01%	11,006	77.99%	23,645	14,487	61.27%
hathitrust	2,474,530	1,068,159	43.17%	1,406,371	56.83%	4,077,994	1,449,785	35.55%
indiana	62,695	18,819	30.02%	43,876	69.98%	74,009	35,907	48.52%
internet_archive	212,902	40,877	19.20%	172,025	80.80%	521,102	128,870	24.73%
kdl	144,202	142,268	98.66%	1,934	1.34%	144,202	693	0.48%
mdl	483,086	44,989	9.31%	438,097	90.69%	483,086	195,321	40.43%
missouri-hub	144,424	17,808	12.33%	126,616	87.67%	169,332	89,907	53.10%
mwdl	932,808	57,899	6.21%	874,909	93.79%	1,195,954	741,141	61.97%
nara	700,948	692,759	98.83%	8,189	1.17%	700,948	4,667	0.67%
nypl	1,170,436	775,361	66.25%	395,075	33.75%	1,170,438	61,423	5.25%
scdl	159,092	33,036	20.77%	126,056	79.23%	159,598	53,974	33.82%
smithsonian	1,250,705	68,871	5.51%	1,181,834	94.49%	2,805,327	343,372	12.24%
the_portal_to_texas_history	649,276	125	0.02%	649,151	99.98%	1,271,500	234,696	18.46%
tn	151,334	2,463	1.63%	148,871	98.37%	151,334	129,605	85.64%
uiuc	18,231	127	0.70%	18,104	99.30%	63,403	25,123	39.62%
undefined_provider	11,422	11,410	99.89%	12	0.11%	11,436	16	0.14%
usc	1,065,641	852,076	79.96%	213,565	20.04%	1,076,016	182,084	16.92%
virginia	30,174	21,081	69.86%	9,093	30.14%	30,174	1,118	3.71%
washington	42,024	8,838	21.03%	33,186	78.97%	42,024	20,710	49.28%

Among larger collections, however, the amount of duplication in Description values does not follow similar patterns. The four hubs containing more than 1 million items -- HathiTrust, New York Public Library, the Smithsonian Institution, and University of Southern California Libraries -- have uniqueness values ranging from a mere 5.25% to nearly 36%. In addition to the four largest contributors, two other hubs have more than 1 million descriptions, though fewer items: The Portal to Texas History (1,271,500 descriptions with only 18.5% uniqueness) and Mountain West Digital Library (1,195,945 descriptions with roughly 62% uniqueness). Tennessee Digital Library has the highest level of uniqueness (86%) with only 151,334 items. Overall, there do not appear to be any generalizable correlations among collection size, number of descriptions, and uniqueness.

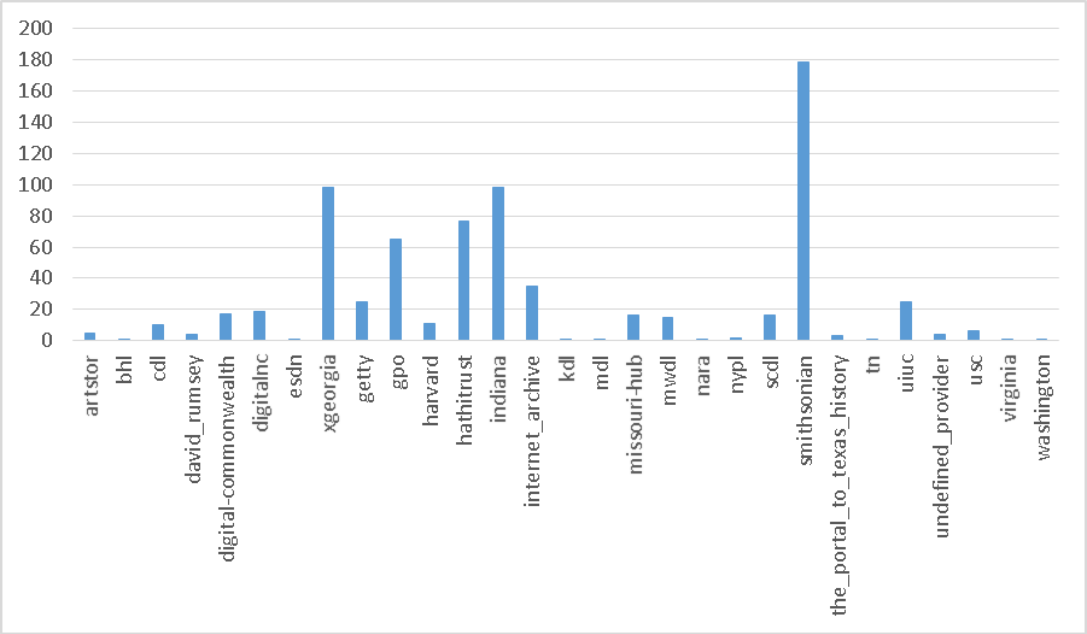

FIG. 1. Largest number of Description instances in any record, by DPLA hub.

3.2. Description Length

After looking at usage of the Description field by hubs, we wanted to gain a better sense of the field values and compare them across the dataset.

Our analysis shows that the length of Description field values in all DPLA metadata records averaged 83.3 words. The range of Description lengths was very broad, with a standard deviation of 373.71 and a maximum length of 130,592 words (approximately 45 pages of text).

Table 2 shows the distribution of Description value lengths by hub. Our analysis identified five hubs with the highest average lengths, ranging from 201 to 447 words: David Rumsey, J. Paul Getty Trust, Minnesota Digital Library, Missouri Hub, and Tennessee Digital Library. On the other side of the spectrum, three out of five hubs with the shortest average length of data values (i.e., under 10 words) are the same three hubs with the lowest number of records containing Description fields and the lowest level of uniqueness: Kentucky Digital Library (2.71 words), NARA (2.03 words) and undefined provider (0.21 words). The other two hubs with the shortest lengths of data values are Biodiversity Heritage Library (6.29 words) and University of Virginia Library (9.98 words).

It is also notable that the spread of lengths is vast for some hubs, e.g., Missouri Hub with an average of 210 characters, but a standard deviation of 2325. Mountain West Digital Library and David Rumsey both have extremely large standard deviations also, with 905.5 (average 154.6 characters) and 861.92 (average 447.36 characters) respectively. The smallest standard deviation (aside from "undefined provider") is Biodiversity Heritage Library (8.48), though the average length is only 6.28 characters.

Figure 2 shows lengths of Description values on a log-log scale. A noticeable spike at 10 characters sets off the group of extremely short descriptions. Although 4.1 million records have no Description values (i.e., a length of 0), they do not display on the log scale; the set from 1-10 characters is more than 2 million descriptions (roughly 2%). On the far left axis, nearly 800,000 values are only a single character long. From that point, the graph shows a clear inverse relationship between the number of characters and the number of records in which they appear (i.e., records with larger numbers of characters tend to occur less frequently). However, there are several obvious spikes, particularly around 800-1,000 characters and 1,500-1,800 characters. These longer values likely represent full sentences and paragraphs rather than the single or few words in the shorter values.

TABLE 2: Description field length statistics by hub.

Hub	Minimum Length	Maximum Length	Instances	Sum of Lengths	Mean/ Average	Standard Deviation
artstor	0	6,868	128,922	9,413,898	73.02	178.31
bhl	0	100	123,472	775,600	6.28	8.48
cdl	0	6,714	563,967	65,221,428	115.65	211.47
david_rumsey	0	5,269	166,314	74,401,401	447.36	861.92
digital-commonwealth	0	23,455	455,369	40,724,507	89.43	214.09
digitalnc	0	9,785	241,224	45,759,118	189.66	262.89
esdn	0	9,136	197,396	23,620,299	119.66	170.67
xgeorgia	0	12,546	821,067	135,691,768	155.05	210.85
getty	0	2,699	264,268	80,243,547	303.64	273.36
gpo	0	1,969	690,883	33,007,265	47.81	58.20
harvard	0	2,277	23,645	2,424,583	102.54	194.02
hathitrust	0	7,276	4,077,994	174,039,559	42.66	88.03
indiana	0	4,477	74,009	6,893,350	93.93	189.30
internet_archive	0	7,685	521,102	41,713,913	79.68	174.94
kdl	0	974	144,202	390,829	2.71	24.95
mdl	0	40,598	483,086	105,858,580	219.13	345.47
missouri-hub	0	130,592	169,332	35,593,253	210.14	2325.08
mwdl	0	126,427	1,195,954	174,126,243	145.60	905.51
nara	0	2,000	700,948	1,425,165	2.03	28.13
nypl	0	2,633	1,170,438	48,750,103	41.65	161.88
scdl	0	3,362	159,598	18,422,935	115.37	164.74
smithsonian	0	6,076	2,805,327	139,062,761	49.52	137.37
the_portal_to_texas_history	0	5,066	1,271,500	132,235,329	104.00	95.95
tn	0	46,312	151,334	30,513,013	201.63	248.79
uiuc	0	4,942	63,403	3,782,743	59.65	172.44
undefined_provider	0	469	11,436	2,373	0.21	6.09
usc	0	29,861	1,076,016	60,538,490	56.26	193.20
virginia	0	268	30,174	301,042	9.98	17.91
washington	0	1,000	42,024	5,258,527	125.13	177.40

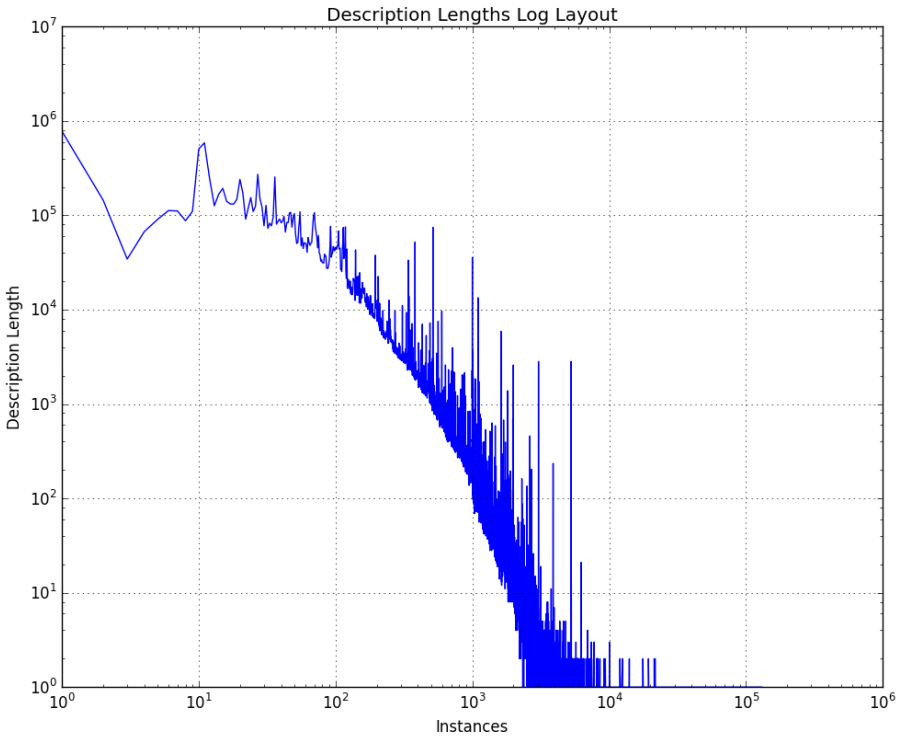

FIG. 2. Lengths of all Description values in the DPLA dataset.

4. Discussion

Although the amount of information gleaned from this form of analysis can be somewhat limiting, there are some definite points that we can make based on our observations, and in the course of considering our research questions with the data in mind.

First, we wanted to look at overall usage of the Description field among DPLA hubs. The number of records without Description values suggests that different hubs (and perhaps different partner institutions, in the case of service hubs) may not all consider the Description field to be equally important, or may not enforce the usage of a Description field. However, considering DPLA's practices, which map many descriptive or "note" type fields that do not map elsewhere into the Description field, it appears that many hubs do not allow or actively record descriptive information of any kind at an item level. This is somewhat at odds with the literature, which suggests the use of free-text fields as a way of most adequately imparting important information to users about the items. Comparing perceived importance with actual usage could be an interesting source of potential future research.

Our other initial question sought to discover what we might learn about usage of the Description field (among records containing Description values) by looking at their various attributes. The most obvious pattern is that there is essentially no pattern -- the number of element instances and the length of field data values vary wildly across DPLA contributing hubs and, in some cases, within hub collections. However, we do wish to offer some additional explanations about why this emerged in the data and what it could mean going forward.

4.1. Description Lengths

Aside from outliers, it is hard to draw definitive conclusions about the range of lengths. For example, longer Description field lengths could indicate more rigorous description standards in these hubs (e.g., specific guidelines on the level of detail that should be included in Description field values). The lengths could also be dependent on specific description practices for the types of information objects that are prevalent in collections of these hubs (e.g., image-based materials may have longer descriptions than collections of primarily printed text with OCR files). Exploring these aspects could be a useful opportunity for future research.

In addition to the large data analysis, we surveyed a random sample of roughly 200 Description field data values per hub for a total of 2800 data values. Although this sample is not very large and we did not have enough time to draw definitive conclusions within the scope of this paper, the Description values do provide some insight into the statistics and allow us to make more educated evaluations of the previous analysis (see Table 3).

TABLE 3: Selected Description field values from DPLA records.

Description Value	Information Type
1 glass negative: b&w; 8 x 10 in.; sulfiding.	Physical object description
This material has been provided by The Royal College of Surgeons of England. The original may be consulted at The Royal College of Surgeons of England	Rights or usage statement
This image shows a section of Thorn Cemetery including gravestones.	Object content description
Microform.	Object type or format
Title supplied by cataloger.	Note or metadata source
This series contains transcripts of proceedings, depositions, and oral examinations prepared exclusively for or in the District Court. The depositions and oral examinations were taken out of court and are primarily interviews with School Board representatives and employees concerning the development, implementation, and review of desegregation plans.	Collection-level content description
P950.	Identifier or call number

The sample includes data values containing a variety of information, such as rights and use statements, physical descriptions, and collection-level descriptions. These kinds of descriptions

may account, in part, for the relatively high number of duplicate values observed in the data. For example, if the same use statement or collection information is propagated across a large group of item records, it would reduce the number of unique data values in a collection of metadata records. Similarly, records may have the same Description field values describing physical attributes that are identical for many items (e.g., 13 p. or 4 x 5 in.), or describing the content attributes of a large number of serial items.

This also provides one explanation for the large number of extremely short Description values. Some hubs use Description fields to hold data values that contain only an item count, page count, or a short term from a controlled vocabulary. In addition, some Description values identify names of places, people, or events without contextual information, which likely accounts for some shorter data values.

4.2. Mapping

Since DPLA is an aggregation, much of the information available in these records is in a shortened format depending on how it is harvested or the level of normalization to fit the DPLA profile. As a relatively generic, free-text field (which also has no strict guidance or recommendations), Description serves as a mapping point for many different native metadata fields. This also makes it difficult to determine if the variety of information types observed in the dataset analyzed in this study is due to differing perceptions of "Description" among hubs and contributors, if there is simply no better place to map the information in DPLA, if the contributed records are too inconsistent to map more accurately, or some combination of all of these factors. However, it does seem that some information found in DPLA metadata records' Description fields could/should be mapped to a more appropriate field (e.g., rights statements).

This is another area that could benefit from much deeper research in terms of how different institutions define or perceive item-level (and collection-level) metadata, both in native systems and as part of an aggregation. Additional research may also consider classifying values currently mapped to Description and the possibility of automatically identifying some information to map values more accurately or to mark them for review for quality control.

4.3. Context and Quality

While not conclusive, several of the statistics identified within this research can help identify metadata records within the DPLA dataset that are in need of remediation. Specifically, records that have Description field values of more than 20,000 characters should be reviewed as to their appropriateness to local descriptive metadata input rules. In many cases, the values at the high end of the length spectrum likely contain the full text of the materials described by the records, and suggests possible problems with the quality of metadata records.

At the same time, records with extremely short values suggest the need for additional review in order for users to understand the information in its aggregated form. Institutions could consider a change in the way that the data values are entered, if one of the primary goals for those institutions is to make information shareable/aggregatable (thought it may not be). Aside from local changes, perhaps there is some potential for preserving or representing more of the contextual information that has been lost within the aggregation.

Even in a native system, extremely short descriptions that are part of a free-text field may suggest a lack of relevant information about the item. For example, a three-word description, such as "A view east" could be accurate in relating to an item without providing sufficient context to help users understand an item's relevance; this statement could refer to a photograph (of nearly anything), a poem title, a map, etc. Similarly, although identifying a name or location is generally considered important, without any context, a proper name remains extremely vague -- e.g., is the name of a person describing an individual pictured in a photo or artwork, a donor, one person in a group photo, or the subject of an obituary or text? From this perspective, contextual information

within a Description or free-text field could be considered highly important to the quality of the field value and the metadata record's usefulness.

5. Conclusions

The empirical data collected and analyzed in this study allows us to make a conclusion that simple statistical analyses can provide a better understanding field usage within a large metadata set. In this case, by investigating the Description fields from the Digital Public Library of America, we were able to consider a wide range of conceptual and technical models for metadata creation by a large number of institutions across the country. This diversity allows for a better understanding of practices than similar analysis within a single institution. However, our findings also show that the Description field and the nature of aggregated free-text fields are areas that would greatly benefit from additional research that was outside our scope and time constraints.

5.1. Further Research

This research was not able to take advantage of the majority of the Description attributes indexed in the methods described above. Performing similar analysis on these additional attributes would result in a better understanding of how the Description field is being used at a wide range of institutions, beyond the usage and length metrics.

Some areas of specific interest for further research include the use of language by each of the providers. This was calculated by identifying, for each of the Description values, the percentage of words that come from various lists of frequently-used English words (e.g., comparing data values to the 1,000 and 5,000 most frequently used English words, and against a standard English dictionary). Additionally, further investigation in this area could provide insight into the reading levels and intended audiences of the metadata being created at each of the provider/hubs. Along these same lines, research into how descriptive information helps users find items and the perception of usefulness by user communities could help to refine guidelines around Description field usage and importance.

On a broader level, the analysis in this report represents a "distant reading" of metadata values in a large dataset. In order to further understand the use of the Description field in the DPLA metadata aggregation, a "close reading" of the Description field values would be beneficial to practitioners and technologists working with metadata aggregations.

References

Baca, M. and P. Harpring (Eds.). (2009) Categories for the Description of Works of Art (CDWA), Getty Research Institute, Santa Monica.

Baca, M., et al. (2006) Cataloging Cultural Objects: A Guide to Describing Cultural Works and their Images, American Library Association, Chicago.

Digital Public Library of America (2015a, March 5). An introduction to the DPLA metadata model. Retrieved from http://dp.la/info/wp-content/uploads/2015/03/Intro_to_DPLA_metadata_model.pdf

Digital Public Library of America (2015b, March 5). Metadata application profile: Version 4.0. Retrieved from http://dp.la/info/wp-content/uploads/2015/03/MAPv4.pdf

Encoded Archival Description. (2002). Retrieved from http://www.loc.gov/ead/ .

Encoded Archival Description: EAD3. (2015). Retrieved from http://www2.archivists.org/sites/all/files/TagLibrary-VersionEAD3.pdf.

Jackson, A.S., M. Han, K. Groetsch., M. Mustafoff and T. W. Cole. (2008). Dublin Core metadata harvested through OAI-PMH. *Journal of Library Metadata, 8*(1), 5-21.

Hillmann, D. (2005). Using Dublin Core. Retrieved from http://dublincore.org/documents/usageguide/

Kurtz, M. (2010). Dublin Core, DSpace, and a brief analysis of three university repositories. *Information Technology & Libraries, 29*(1), 40-46. Retrieved from http://ejournals.bc.edu/ojs/index.php/ital/article/view/3157/2771

Ma, Hong. (2014). Techservices on the Web: DPLA: Digital Public Library of America. *Technical Services Quarterly, 31*(1), 83-84. doi: 10.1080/07317131.2014.845013

Mitchell, Erik T. (2013). Three case studies in linked open data. *Library Technology Reports*, *49*(5), 26-43.

National Union Catalog of Manuscript Collections. (2011). Online Data Sheet for Participating Institutions. Retrieved from http://www.loc.gov/coll/nucmc/lcforms.html.

OLAC Cataloging Policy Committee, Summary/Abstracts Task Force. (2002) Summary Notes for Catalog Records. Retrieved from http://www.olacinc.org/drupal/?q=node/21.

OSU Knowledge Bank Metadata Application Profile for Digital Video. (2011). Retrieved from https://library.osu.edu/documents/knowledge-bank/KnowledgeBankMetadataApplicationProfile2011.pdf

Park, J. (2006). Semantic interoperability and metadata quality: An analysis of metadata item records of digital image collections. *Knowledge Organization*, *33* (1), 20-34.

Ward, J. (2003). A quantitative analysis if unqualified Dublin Core metadata element set usage within data providers registered with the Open Archives Initiative. Proceedings of the 2003 Joint Conference on Digital Libraries, pp. 315-317.

Weagley, J., E. Gelches, & J. Park. (2010). Interoperability and metadata quality in digital video repositories: a study of Dublin Core. *Journal of Library Metadata*, *10*(1), 37-57. DOI: 10.1080/19386380903546984.

Zavalina, O.L. (2012). Exploring the richness of collection-level subject metadata in three large-scale digital libraries. *International Journal of Metadata, Semantics, and Ontologies*, *7*(3), 209-221.

Zavalina, O.L., C.L. Palmer, A. S. Jackson, and M.-J. Han. (2008). Evaluating descriptive richness in collection-level metadata. *Journal of Library Metadata*, *8*(4), 263-292.

Appendix A

TABLE 4: Distribution and statistics for Description field instances in metadata records by hub.

Hub	Records	Minimum	Median	Maximum	Average	Standard Deviation
artstor	107,665	0	1	5	.82	.84
bhl	123,472	0	0	1	.47	.50
cdl	312,573	0	1	10	1.55	1.46
david_rumsey	65,244	0	3	4	2.55	.80
digital-commonwealth	222,102	0	2	17	2.01	1.15
digitalnc	281,087	0	1	19	.86	.67
esdn	197,396	0	1	1	.75	.43
xgeorgia	373,083	0	2	98	2.32	1.56
getty	95,908	0	2	25	2.75	2.59
gpo	158,228	0	4	65	4.37	2.53
harvard	14,112	0	1	11	1.46	1.24
hathitrust	2,474,530	0	1	77	1.22	1.57
indiana	62,695	0	1	98	.91	1.21
internet_archive	212,902	0	2	35	2.27	2.29
kdl	144,202	0	0	1	.01	.12
mdl	483,086	0	1	1	.91	.29
missouri-hub	144,424	0	1	16	1.05	.70
mwdl	932,808	0	1	15	1.22	.86
nara	700,948	0	0	1	.01	.11
nypl	1,170,436	0	0	2	.34	.47
scdl	159,092	0	1	16	.80	.41
smithsonian	1,250,705	0	2	179	2.19	1.94
the_portal_to_texas_history	649,276	0	2	3	1.96	.20
tn	151,334	0	1	1	.98	.13
uiuc	18,231	0	3	25	3.47	2.13
undefined_provider	11,422	0	0	4	.00	.08
usc	1,065,641	0	0	6	.21	.43
virginia	30,174	0	0	1	.30	.46
washington	42,024	0	1	1	.79	.41