An Exploratory Analysis of Subject Metadata in the Digital Public Library of America

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

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

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

Priya Kizhakkethil
University of North Texas, USA
priyakizhakkethil@my.unt.edu

Abstract

This paper presents results of an exploratory quantitative analysis of subject representation in the large dataset of over 8 million item-level metadata records in the Digital Public Library of America (DPLA) originating from a number of institutions that serve as content or service hubs of DPLA. The findings demonstrate both similarities and differences in subject representation across content and service hub providers. This benchmark study provides empirical data about the distribution of subjects at the hub level (e.g., minimum, maximum, and average number of subjects per record; number of records without subjects; and number of unique subjects) as well as distribution by hub type (content or service hubs), and subjects shared across similar hubs or across the entire aggregation.

Keywords: metadata aggregations; keywords; metadata values; subject analysis; subject terms

1. Introduction and Background

Cultural heritage institutions and funding agencies worldwide have invested intensively in digitization projects; however in many cases, access to those digitized collections often remained in separate pockets or silos. Large-scale digital libraries now bring together hundreds of individual digital collections and millions of items produced by these projects. The Digital Public Library of America (DPLA) is currently one of the most prominent such aggregations. Arising out of a vision from the early 1990s of a national digital library, shared by librarians, scholars, educators, and others, DPLA brings "different viewpoints, experience, and collections together in a single platform and portal, providing open and coherent access to our society's digitized cultural heritage" ("About", dp.la, 2015). Functioning on a distributed network model, DPLA consists of a group of national partners providing both content and services (Ma, 2014). DPLA was formed in 2010 and got underway in 2013 with support from a number of funding agencies which include the Alfred P. Sloan Foundation, the Arcadia Fund, the Institute of Museum and Library Services (IMLS), the John S. and James L. Knight Foundation, and the National Endowment for the Humanities (Mitchell, 2013).

Relying on a distributed network of partners to host and preserve digital information, DPLA focuses on the compilation of metadata to augment the discovery of these resources and to provide a useful platform where libraries and their patrons can make the best use of them. In addition, DPLA also provides APIs (Application Profile Interfaces) and maximally-open data to software developers, researchers, and others for building discovery tools along with providing access and communication (Ma, 2014). The DPLA community has also embraced the tenets of open data and adopted an advocacy stance in support of open access policies. On its launch in April 2013, a discovery platform provided access to an initial data set contributed by eighteen partners, or "hubs," comprising more than two million records in over 3,200 collections. Since

the launch, the size of the aggregate collection and the number of partner institutions have continued to grow (Mitchell, 2013).

The internal data model of DPLA is based on the Resource Description Framework (RDF) and employs JSON-LD (JavaScript Object Notation-based serialization for Linked Data) for dissemination of metadata via API output. Based on the Europeana data model, the emphasis is on supporting the creation of graph structures and the standard is essentially a data aggregation and sharing service. Since the primary goal is the compilation of harvested data, some of the data gathered from providers is stored along with data generated or extracted during the data collection process. The DPLA metadata model is based on RDF and the central descriptive metadata standard employed is the Dublin Core (DC) (Mitchell, 2013). The metadata aggregated and normalized by DPLA is in the public domain and has no copyright restriction; DPLA data can be downloaded as JSON files, allowing for sharing or data analysis.

Although metadata analysis can lead in many directions, one field of significance is a subject field, since subject representation has applications in information retrieval, as well as in disciplines such as automated language processing and knowledge engineering that reference knowledge structures. In Svenonius (2000) definition, the "subject language" depicts what a document is about. Similarly, Soergel (2009) defines subject metadata in digital libraries as information concerning what the information object is about and why it is relevant.

Assigning subject metadata is based on subject analysis, for which various models have been proposed (e.g., Beghtol, 1986; Hjørland, 1998; Langridge, 1989; Šauperl, 2002; Wilson, 1968). These models guide the metadata creators to examine a document not only for its content, but also for author's intentions, for viewpoints and possible bias, and to take into account when assigning subject terms the intended audience and intellectual level, as well as possible uses of information. According to Wilson (1968), since most works are multifaceted and cover more than one subject, the notion of "the" subject of a work is "indeterminate" (p. 318), i.e., in some cases it would be impossible in principle to decide between more than one different and equally precise descriptions to be the one and only subject of a work. Hjørland (1992) further developed this idea of multiplicity of a document's subjects by taking the approach that subjects of a document can be defined as the informative or epistemological potentials of that document. According to Hjørland (1997), these intellectual potentials of a document can differ depending on periods of time and societal development, as well as across different domains, which would ideally require periodically revising subject headings in bibliographic records.

Subject metadata is crucial for providing access to information objects in both traditional library collections and digital collections and aggregations. To help achieve optimal recall and precision, it is recommended (e.g., ALCTS, 1999) to include Subject, Type, and Coverage elements in metadata records in digital libraries to accommodate different subject-related facets: topic, place, time period, language, etc. Gross & Taylor (2005) found that in the absence of subject headings in a catalog record, more than one third of the retrievals would be missed when a user performs a keyword search. In a study assessing the benefits of adding subject metadata to online records of the Northwestern University Library's Eighteenth Century Collections Online (ECCO), Garrett (2007) extends the arguments forwarded by Gross & Taylor (2005) on the benefit afforded by subject headings for providing access even when the full text of a work is accessible. In a replication of the 2005 study, Gross, Taylor & Joudrey (2015) found that even with the addition of tables of contents and summaries or abstracts in the catalog records (which reduced lost hits), the absence of subject headings leads to an average of 27% of the retrievals to be missed.

Evaluation of metadata in digital libraries has gained more importance to ensure metadata quality (Hillmann, 2008). Margaritopoulos et al. (2009; 2012) discuss subject metadata from the point of view of measuring metadata quality, and in particular, completeness of metadata records. They point out that multivalued metadata fields such as subject are normally considered complete

if populated with at least one value; however multiple instances should be considered to determine the richness of the field, which can make the evaluation more complicated.

The empirical assessment of metadata has not yet become a common practice. In particular, few of the available studies that analyzed item-level metadata in digital libraries, included subject-metadata-related components. Several quantitative studies of item-level metadata in digital libraries (Jackson, Han, Groetsch, Mustafoff, and Cole, 2008; Kurtz, 2010; Weagley, Gelches, & Park, 2010) did not focus specifically on subject metadata but looked at the percentage of records that included one or more instances of each metadata element, including the subject metadata elements. For example, Kurtz's (2010) study of metadata in three university repositories revealed that the Dublin Core Subject field was included in only 65% of records. Weagley, Gelches, and Park's (2010) study of metadata in six digital video repositories reported the same level (65%) of Subject field utilization. To the contrary, Jackson and colleagues (2008) found Subject field values in almost all (94%) of metadata records harvested through OAI-PMH. The Dublin Core Coverage metadata element was found to be included in 7% and 21% of metadata records in the Kurtz (2010) Weagley, Gelches, and Park (2010) studies respectively and in 51% of records in the Jackson et al. (2008) study. Another study (Ma, Lu, Lin, & Galloway, 2009), which combined quantitative and qualitative approaches in overall analysis of item-level metadata in the Internet Public Library (IPL), evaluated users' ratings of the subject representation in IPL metadata through controlled-vocabulary subject headings and free-text keywords; the completeness of keywords was perceived to be quite low.

The analysis of literature reveals that little research to date has been conducted with the goal of specifically evaluating subject metadata in digital libraries. Available studies of subject metadata in digital libraries focused on collection-level metadata which describes entire collections of information objects as opposed to item-level metadata which describes each individual information object. For example, Zavalina (2011) examined and compared the free-text collection-level subject metadata (i.e., data values in the Description metadata field) across multiple digital libraries. The follow-up study (Zavalina, 2012) compared the data values in free-text Description and four controlled-vocabulary subject metadata fields -- Subjects, Temporal Coverage, Geographic Coverage, and Object Types/ Genres -- in three digital libraries: American Memory, Opening History, and The European Library. These two studies used a detailed manual content analysis and focused more on the qualitative characteristics of subject metadata than on quantitative ones. Some quantitative indicators that were measured in Zavalina (2012) study include the data value length (measured as the number of characters) -- range, median, mean, variance and standard deviation -- of each of the 5 subject metadata fields in the records.

The study reported in this paper is one of the first attempts to systematically evaluate subject metadata, and the first one to use a very large aggregator such as the Digital Public Library of America as its target.

2. Methods

The research questions that guided this exploratory study are: How are the subjects of information objects represented in metadata records across collections in the Digital Public Library of America (DPLA)? What are the differences and similarities in subject metadata originating from content hubs and service hubs?

Content hubs are digital repositories that maintain a one-to-one relationship with DPLA, providing metadata records for items owned or produced by that organization, such as ARTstor, California Digital University, The U.S. Government Publishing Office, and Harvard Library. Service hubs are state, regional, or other collaborative entities that bring together digital objects from multiple cultural heritage institutions and provide metadata records from all hosted or aggregated materials to DPLA through a single data feed. Some of the service hubs of DPLA are the Connecticut Digital Archive, Digital Library of Georgia, and The Portal to Texas History ("hubs", dp.la, 2015).

Unlike the previous studies of subject metadata in digital libraries that analyzed a generalizable sample of metadata records, the researchers of this study took a "big data" approach that analyzed the whole dataset and therefore avoided sampling errors. To address the research questions, the researchers used DPLA's Bulk Download[1] to download the complete DPLA metadata dataset. This dataset was parsed into individual item records that contained both the original metadata from submitted by various DPLA hubs as well as a normalized version of the metadata in accordance with the DPLA Metadata Application Profile[2]. In total the DPLA dataset (Phillips, 2015) contained 8,012,390 metadata records which were used in this analysis.

Each metadata record was parsed and the DPLA-normalized metadata was extracted for processing. The raw data for each field and the number of instances of the element in each record were added to a Solr index that the researchers used for their analysis in this paper; since the researchers chose to focus on subject terms for the purposes of this study, the data was limited to the dc:subject field values. Below is an example of the extracted and calculated data added to the Solr index for each field in the DPLA Metadata Application Profile for each record (Fig. 1). The example is represented in the JavaScript Object Notation (JSON) format that the researchers used for submitting data to the Solr index; this example shows that the record had two subject values, "Sun" and "Men."

```
{
 "subject_ss": [
           "Sun",
           "Men"
         ],
 "subject_count_i": 2
  }
```

FIG. 1. Example JSON created from a metadata record.

The researchers decided that for each record they would calculate the number of instances of each element in the record, and if there were no instances of that element in a given record then the count for that element would default to 0 for analysis.

The researchers used the Solr search framework to form queries for data analysis. Two components were particularly useful: the StatsComponent, which provides high level statistics for a specified field or set of fields in the index, and the Facet feature, which groups values, provides a count of instances of elements, and presents the number of records with a given value for a defined element. When the built-in features of Solr were not sufficient to answer the questions posed by the researchers, they wrote a series of Python scripts that would interact with Solr directly and apply additional logic and calculation to the data.

3. Findings

After general review of the data, the first finding of this analysis was that the average number of subjects per record in DPLA is 2.99, with a standard deviation of 3.90. In the dataset, 1,827,276 records had zero subjects, representing 22.8 percent of total records (see Table 1). For each hub, Table 1 lists the hub type, minimum and maximum number of subjects in the hub's records, the number of items/metadata records, the total number of subject entries, the average number of subjects per record (mean), and standard deviation (stddev).

[1] http://dp.la/info/developers/download/.
[2] http://dp.la/info/developers/map/.

TABLE 1: Statistics for subject fields for each hub in the DPLA dataset.

Hub Name	Hub Type	Min	Max	Records	Subjects	Mean	Stddev
ARTstor	Content	0	71	56,342	194,948	3.46	3.47
Biodiversity Heritage Library	Content	0	118	138,288	454,624	3.29	3.41
David Rumsey	Content	0	4	48,132	22,976	0.48	0.69
Digital Commonwealth	Service	0	199	124,804	295,778	2.37	2.92
Digital Library of Georgia	Service	0	161	259,640	1,151,369	4.43	3.68
Harvard Library	Content	0	17	10,568	26,641	2.52	1.41
HathiTrust	Content	0	92	1,915,159	2,614,199	1.37	1.33
Internet Archive	Content	0	68	208,953	385,732	1.85	1.97
J. Paul Getty Trust	Content	0	36	92,681	32,999	0.36	1.21
Kentucky Digital Library	Service	0	13	127,755	26,009	0.20	0.78
Minnesota Digital Library	Service	1	78	40,533	202,484	5.00	2.66
Missouri Hub	Service	0	139	41,557	97,115	2.34	3.02
Mountain West Digital Library	Service	0	129	867,538	2,641,065	3.04	3.34
National Archives and Records Administration	Content	0	103	700,952	231,513	0.33	1.23
North Carolina Digital Heritage Center	Service	0	1,476	260,709	869,203	3.33	4.59
Smithsonian Institution	Content	0	548	897,196	5,763,459	6.42	4.65
South Carolina Digital Library	Service	0	40	76,001	231,270	3.04	2.35
The New York Public Library	Content	0	31	1,169,576	1,996,483	1.71	1.65
The Portal to Texas History	Service	0	1,035	477,639	5,257,702	11.01	4.97
United States Government Publishing Office	Content	0	30	148,715	457,097	3.07	1.75
University of Illinois at Urbana-Champaign	Content	0	22	18,103	67,955	3.75	2.87
University of Southern California Libraries	Content	0	119	301,325	863,535	2.87	2.67
University of Virginia Library	Content	0	15	30,188	95,328	3.16	2.33

This data showed some interesting results including that only the Minnesota Digital Library had at least one subject for all 40,533 of its records. There were two hubs, North Carolina Digital Heritage Center and The Portal to Texas History, which had individual records containing more than 1,000 subject headings (1,476 and 1,035 respectively). The average subjects-per-record ranged from 0.2 at the Kentucky Digital Library to 11.0 at The Portal to Texas History.

The next step was to break down the data based on hub types (service versus content hubs) for comparison (see Table 2). The researchers found that the average number of subjects for content hubs was 2.3 subjects per record, while the service hubs averaged 4.7 subjects per record. This means that service hubs tend to have twice as many subjects and keywords in their records as content hubs.

TABLE 2: Statistics for the subject field based on category (content hub or service hub).

Hub Type	Min	Max	Records	Subjects	Mean	Stddev
Content Hub	0	548	5,736,178	13,207,489	2.3	3.08
Service Hub	0	1,476	2,276,176	10,771,995	4.7	5.06

Further analysis of the metadata records originating from content hubs and service hubs showed that content hubs had a total of 1,590,456 records (28%) without any subjects compared to service hubs which had only 236,811 (10%) records without subjects.

The researchers also calculated additional metrics at the hub level for the DPLA records: the number of records without subjects, percentage of records without subjects, the mode of number of subjects-per-record, unique subjects, subjects unique to a single hub, and finally the entropy of the subject field for the specified hub (see Table 3). Entropy in this context represents a measure

of the average information content or similarity of values for a particular field, i.e., collections that have fewer unique values (more similar terms) will have a lower entropy score.

TABLE 3: Additional statistics for subject fields for each hub in the DPLA dataset.

Hub Name	Records	Records Without Subjects	% Without Subjects	Average Subjects per Record	Subject Count Mode	Unique Subjects	Subjects Unique to Hub	Entropy*
ARTstor	56,342	6,586	11.7	3.5	3	9,560	4,941	0.73
Biodiversity Heritage Library	138,288	10,326	7.5	3.3	2	22,004	9,136	0.65
David Rumsey	48,132	30,167	62.7	0.5	0	123	30	0.76
Digital Commonwealth	124,804	6,040	4.8	2.4	1	41,704	31,094	0.77
Digital Library of Georgia	259,640	3,216	1.2	4.4	2	132,160	114,689	0.67
Harvard Library	10,568	167	1.6	2.5	2	9,257	7,204	0.76
HathiTrust	1,915,159	525,874	27.5	1.4	1	685,733	570,292	0.88
Internet Archive	208,953	44,872	21.5	1.8	1	56,911	28,978	0.8
J. Paul Getty Trust	92,681	73,978	79.8	0.4	0	2,777	1,852	0.6
Kentucky Digital Library	127,755	117,790	92.2	0.2	0	1,972	1,337	0.62
Minnesota Digital Library	40,533	0	0	5	4	24,472	17,545	0.74
Missouri Hub	41,557	11,451	27.6	2.3	0	6,893	4,338	0.69
Mountain West Digital Library	867,538	49,473	5.7	3	1	227,755	192,501	0.68
National Archives and Records Administration	700,952	619,212	88.3	0.3	0	7,086	3,589	0.63
North Carolina Digital Heritage Center	260,709	41,323	15.9	3.3	2	99,258	84,203	0.66
Smithsonian Institution	897,196	29,452	3.3	6.4	7	348,302	325,878	0.62
South Carolina Digital Library	76,001	7,460	9.8	3	2	23,842	18,110	0.72
The New York Public Library	1,169,576	208,472	17.8	1.7	1	69,210	52,002	0.62
The Portal to Texas History	477,639	58	0	11	10	104,566	87,076	0.49
United States Government Publishing Office	148,715	1,794	1.2	3.1	2	174,067	105,389	0.92
University of Illinois at Urbana-Champaign	18,103	4,221	23.3	3.8	0	6,183	3,076	0.63
University of Southern California Libraries	301,325	35,106	11.7	2.9	2	65,958	51,822	0.59
University of Virginia Library	30,188	229	0.8	3.2	1	3,736	2,425	0.6

* Entropy calculated using the formula from Stvilia, Gasser, Twidale, Shreeves, & Cole (2004)

The data in Table 3 is helpful to identify hubs that have more coverage in the subject fields of their records. There is a range from the previously-mentioned Minnesota Digital Library that has zero records without subjects, or The Portal to Texas History that has 58 records (.01%) without subjects, to the National Archives and Records Administration with 88.3% and Kentucky Digital Library with 92.2% of their records lacking subject headings. The calculation of the number of subjects that are unique to a Hub showed that the Smithsonian Institution has 94% of its subjects unique to just the Smithsonian, while several other hubs share roughly half of their subjects with at least one other institution: ArtStor (52%), Biodiversity Heritage Library (42%), Internet Archive (51%), NARA (51%), University of Illinois at Urbana-Champaign (50%). The researchers theorize that the generally high number of unique subjects may be caused by the standard library practice of generating subject headings using the Library of Congress Subject Headings (LCSH); because of geographic and temporal qualification of the subjects, this creates a higher number of unique strings. Further analysis in this area could be performed to normalize LCSH into its constituent pieces and re-run the analysis to determine what effect this has on the dataset.

The researchers compiled the same information by hub type (see Table 4) to analyze the overlap of subject terms between hubs of different types.

TABLE 4: Makeup of unique subjects per hub type in the DPLA.

Hub Type	Records	Unique Subjects	Subjects Unique to Hub Type	% of Subjects Unique to Hub Type
Content Hub	5,736,178	1,311,830	1,253,769	96
Service Hub	2,276,176	618,081	560,049	91

A large percentage of subjects -- 96% for content hubs and 91% for service hubs -- are unique to that hub type. In fact, only 3% of the total unique subjects in the dataset are shared between content hubs and service hubs (see Fig. 2).

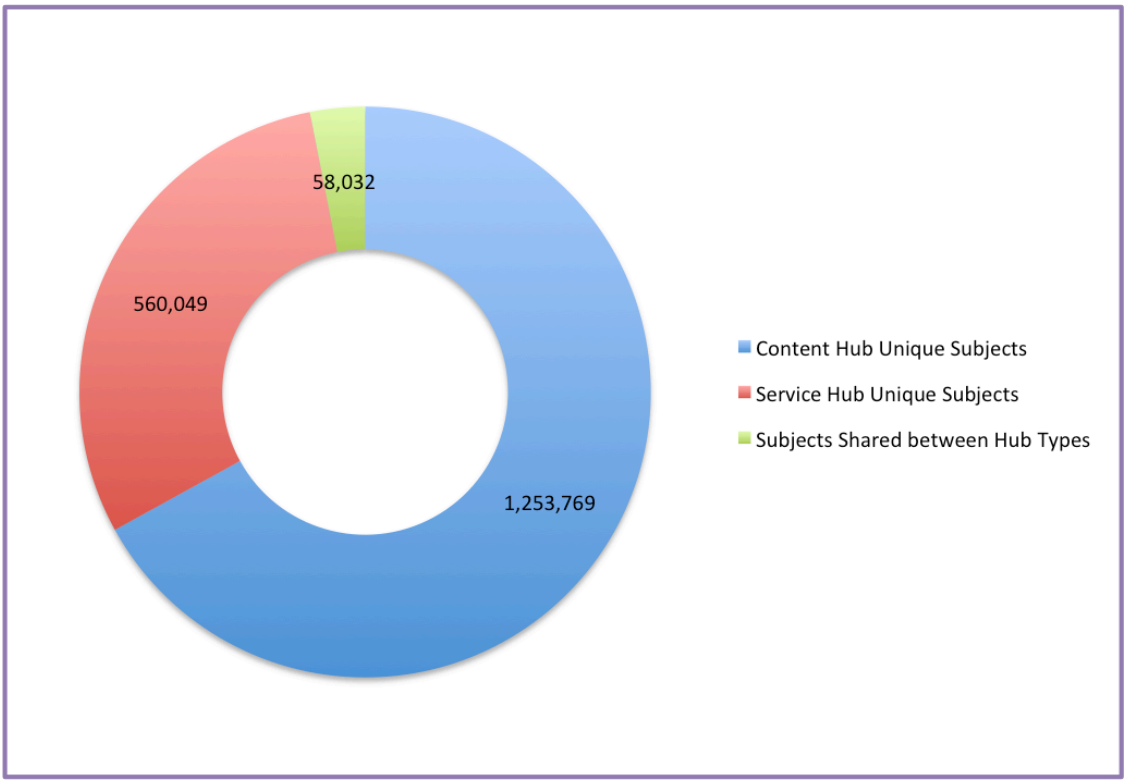

FIG. 2. Comparison of unique and shared subjects between hub types in DPLA.

The next step was to look at shared subjects, which is significant since subject terms have a relatively unique ability to connect users with disparate resource types, and across multiple partner collections, that have common topical content. However, this assumes a level of consistency in subject assignment, so the analysis determined how many subjects were shared across individual hubs and the subjects common to the highest number of hubs (see Table 5).

TABLE 5: Subjects shared, by number of hubs.

Unique Subject Count	# of Hubs with Subject		Unique Subject Count	# of Hubs with Subject
1,717,478	1		302	12
114,047	2		245	13
21,126	3		199	14
8,013	4		152	15
3,905	5		117	16
2,187	6		63	17
1,330	7		62	18
970	8		32	19
689	9		20	20
494	10		7	21
405	11		7	22

Table 5 demonstrates that a large majority of subjects (roughly 92%) are unique to a single hub. Subjects that are shared between two hubs represent 6% of total subjects and only 1% are shared among three hubs. The total number of remaining subjects, shared among four or more hubs, amounts to only 1.5% of total subjects.

The seven subject headings that are shared by twenty-two hubs are: African Americans, Animals, Architecture, Children, Education, Horses, and Transportation.

4. Discussion and Conclusions

Subject terms have a unique place in metadata for several reasons. First, every item has one or more "topics," or content that can be described in topical ways, so it is reasonable to expect that complete metadata records should include subject terms, or that records without terms could be updated given time and resources. This differs from many other fields in a metadata record, for which entries may remain blank simply because the information (e.g., creator, location, etc.) is not known about the item. Secondly, although many metadata fields may be complete with a single data value (e.g., creation date or item language), subject fields often occur as multiple entries in each record, and in most cases, additional subject terms are directly related to additional access points for users (i.e., providing more subject terms increases the findability and usefulness of a metadata record). Finally, to some degree, subject representation requires a certain level of active consideration – that is, a metadata creator has looked at the item, thought about the content, and then generated or assigned subject terms. This suggests that data values associated with subject fields in metadata records can often be tied to curation activities within individual hubs, as opposed to data values in other fields of metadata records which *may* be copied directly from the source item or from accompanying collection-level information (e.g., book titles, or creator names).

This analysis provides a framework for general discussion regarding subjects in digital collections, and in large aggregates. One noticeable finding is the high variability of the number of instances of subject fields across records, ranging from no subjects to more than one thousand. Reasons for these variances would have to be explored locally at individual hubs – for example, records that do not have any subject terms may be due to workflow issues, a lack of tools to discover incomplete records or resources to fix known deficits, or even local practices that do not require or encourage subject representation. Several hubs also had records containing a large number of subject terms (i.e., more than 100, more than 500, or more than 1,000). Based on the experiences of the researchers handling records in The Portal to Texas History, some of the numbers may be slightly inflated due to the normalization process that DPLA uses when importing records. For example, the Portal has a locally-established hierarchical subject vocabulary, the UNT Libraries Browse Subjects (UNTL-BS), that is parsed into separate

keywords when records are harvested and added to DPLA; for example the hierarchical subject string "Business, Economics and Finance - Transportation - Automobiles" becomes keywords "Business, Economics and Finance," "Transportation," and "Automobiles." This means that a record with only one or two controlled terms from the UNTL-BS list may have six or eight keyword terms in the DPLA-normalized record. While this may not account for the extremely large numbers, it could impact some hubs more than others. Additionally, most of the records in the Portal containing higher numbers of subject terms tend to contain many personal names. In fact, the outlier in this dataset is a metadata record representing a ledger of inquest records for which the partner particularly requested that all of the names be included in the metadata (since the ledger is handwritten).

Another discussion point arising from this analysis is the differences in average number of subject terms between hub types. DPLA content hubs provide more than double the number of records that service hubs contribute, however the average number of subjects per record for content hubs is half that of the service hubs. This may be related to the fact that service hubs aggregate or host materials from multiple institutions, and therefore the initial metadata creation or maintenance may be distributed among content holders. Overall, the numbers show a large amount of variance even among hubs of the same type, so it is hard to say with certainty if the differences are more representative of an actual divide by hub type, or of radical differences among individual hubs.

While determining the accuracy and "quality" of subject metadata in these records would be essentially impossible on a large scale, this analysis does provide data related to completeness, i.e., whether or not all records have subject(s), assuming that every record should include at least one subject term. It also highlights those metadata records that do not fit the model of an average record within a particular digital library and may be indicative of problem records or lower quality metadata. On a local level, subject analysis similar to the analysis presented in this paper could help individual hubs to discover gaps or possible areas of metadata enhancement within their own collections. Some examples include identifying records that have no subject entries or for which the number of unique values is higher or lower than expected for the known content.

Aside from records in individual hubs, the findings also highlight the lack of overlap across all of the collections in DPLA since the majority (92%) of subject terms in metadata records are unique to a particular hub. While some of this uniqueness in subject terms might be explained by uniqueness of items contributed to DPLA by individual hubs and varying subject matter of these items, this factor would only contribute a single-digit percentage of uniqueness of subject terms in DPLA. It is likely that most of the 92% uniqueness is due merely to the lack of a common controlled vocabulary. Since DPLA aims to bring items together for access, using fewer unique subject terms across DPLA would appear to be of importance to facilitate finding and collocating materials across the aggregate. However, implementing any plan to improve consistency in subject representation across such a large number of records and content providers would be difficult, time consuming, and could require extensive resources as well as buy-in from the many hubs. Perhaps one option based on the kind of analysis in this paper would be to provide better access to currently-used or most-used subject terms in DPLA metadata for persons who maintain records at individual hubs. While it would not be an immediate fix, it could create an opportunity to start promoting intentional subject overlap.

4.1. Further Study

As this study has shown, the availability of data from DPLA creates an opportunity for various kinds of metadata analysis across aggregated collections or at local institutions. Additional analyses of subject representation in DPLA could look at field values across the collection after basic normalizations. For example, known Library of Congress Subject Heading (LCSH) terms could be broken into constituent pieces in the same way that OCLC parses values into Faceted Application of Subject Terminology (FAST) terms (e.g., "Children--Texas" into "Children" and "Texas"). This could show whether a larger overlap in subject matter exists than is apparent from

analysis of original subject strings. Qualitative studies could also provide context regarding the data in this study, such as the reasons that some records have no subject terms, the differences in the number of subject terms across hubs or hub types, and additional information about the lack of overlap in subject terms within DPLA.

In addition to the dc:subject metadata field, several other fields particularly lend themselves to cross-collection analysis at an aggregate level. For example, coverage field(s) function similarly to subject in the way that they represent content of materials. Analysis of dates, time periods, and geographic elements in coverage values could show where topics converge, or where information could be easily added to provide more item-level access.

On an even broader scale, comparisons of DPLA with other large international digital libraries or aggregates (such as Europeana, Canadiana, etc.) would provide a more extensive dataset for studies in metadata completeness or metadata field usage. The data in this study provides a baseline that could be used as a point of comparison regarding subject term representation in individual metadata records or overlap across large collections and aggregates.

References

ALCTS/CCS/SAC/Subcommittee on Metadata and Subject Analysis. (1999). Subject Data in the Metadata Record: A Report from the ALCTS/CCS/SAC/Subcommittee on Metadata and Subject Analysis Working Draft. Retrieved from http://archive.ifla.org/VII/s12/mom/appendx3.htm

Beghtol, Claire. (1986). Bibliographic classification theory and text linguistics: aboutness analysis, intertextuality and the cognitive act of classifying documents. *Journal of Documentation, 42*(2), 84-113.

Digital Public Library of America. (n.d.). Retrieved March 20, 2015, from http://dp.la/.

FAST (Faceted Application of Subject Terminology). (2013, August). Retrieved April 1, 2015 from http://www.oclc.org/research/themes/data-science/fast.html.

Garrett, Jeffrey. (2007). Subject headings in full-text environments: The ECCO experiment. *College & Research Libraries, 68*(1), 69-81.

Gross, Tina, and Arlene G. Taylor. (2005). What have we got to lose? The effect of controlled vocabulary on keyword searching results. *College & Research Libraries, 66*(3), 212-230.

Gross, Tina, Arlene G. Taylor, and Daniel N. Joudrey. (2015). Still a lot to lose: The role of controlled vocabulary in keyword searching. *Cataloging & Classification Quarterly, 53*(1), 1-39.

Hillmann, Diane Ileana. (2008). Metadata quality: from evaluation to augmentation. *Cataloging & Classification Quarterly, 46*(1), 65-80.

Hjorland, Birger. (1992). The concept of 'subject' in information science. *Journal of Documentation, 48*(2), 172-200. doi: 10.1108/eb026895

Hjørland, Birger. (1997). The concept of subject or subject matter and basic epistemological positions. In Information Seeking and Subject Representation: An Activity-Theoretical Approach to Information Science. Westport CT: Greenwood Press, 55-103.

Hjørland, Birger. (1998) Theory and metatheory of information science: a new interpretation. *Journal of Documentation 54*, 606-621.

Jackson, Amy S., Myung-Ja Han, Kurt Groetsch, Megan Mustafoff, and Timothy W. Cole. (2008). Dublin Core metadata harvested through OAI-PMH. *Journal of Library Metadata, 8*(1), 5-21.

Kurtz, Mary. (2010). Dublin Core, DSpace, and a brief analysis of three university repositories. *Information Technology & Libraries, 29*(1), 40-46.

Langridge, Derek Wilton. (1989). *Subject analysis: principles and procedures*. London: Bowker-Saur.

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

Ma, Shanshan, Caimei Lu, Xia Lin, and Mike Galloway. (2009). Evaluating the metadata quality of the IPL. Proceedings of the Annual Meeting of American Society for Information Science and Technology, 49. http://www.asis.org/Conferences/AM09/open-proceedings/papers/49.xml

Margaritopoulos, Thomas, Merkourios Margaritopoulos, Ioannis Mavridis, and Athanasios Manitsaris. (2009). A fine-grained metric system for the completeness of metadata. In Fabio Sartori, Miguel-Angel Sicilia, & Nikos Manouselis (Eds.), Metadata and semantic research (pp. 83-94). Berlin: Springer.

Margaritopoulos, Merkourios, Thomas Margaritopoulos, Ioannis Mavridis, and Athanasios Manitsaris. (2012). Quantifying and measuring metadata completeness. *Journal of the American Society for Information Science and Technology, 36*(4), 724–737. doi:10.1002/asi.21706.

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

Phillips, Mark Edward. (February 2015). Digital Public Library of America: Bulk Metadata Download, February 2015 [dataset]. http://digital.library.unt.edu/ark:/67531/metadc502991

Šauperl, Alenka. (2002). Subject determination during the cataloging process: observation. Lanham, MD: Scarecrow Press.

Soergel, Dagobert. (2009). Digital libraries and knowledge organization. In S. R. Kruk & B. McDaniel (Eds.), Semantic Digital Libraries, (pp. 9-39). Berlin: Springer.

Svenonius, Elaine. (2000). The intellectual foundations of information organization. Cambridge, MA: MIT Press.

Stvilia, Besiki, Les Gasser, Michael B. Twidale, Sarah L. Shreeves, and Tim W. Cole. (2004). Metadata quality for federated collections. In S. Chengulur-Smith, L. Raschid, J. Long, & C. Seko (Eds.), Proceedings of the International Conference on Information Quality — ICIQ 2004 (pp. 111–125). Cambridge, MA: MIT.

Weagley, Julie, Ellen Gelches, and Jung-Ran Park. (2010). Interoperability and metadata quality in digital video repositories: A study of Dublin Core. *Journal of Library Metadata, 10*(1), 37-57.

Wilson, Patrick. (1968). *Two kinds of power: An essay on bibliographic control.* Berkeley: University of California Press.

Zavalina, Oksana L. (2011). Free-text collection-level subject metadata in large-scale digital libraries: A comparative content analysis. In T. Baker, D. I. Hillmann & A. Isaac (Eds.), Proceedings of the International Conference on Dublin Core and Metadata Applications, (pp. 147-157). The Hague: Dublin Core Metadata Initiative.

Zavalina, Oksana 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. doi: 10.1504/IJMSO.2012.050182