Investigating the use of metadata record graphs to analyze subject headings in the Digital Public Library of America

Mark E. Phillips Hannah Tarver

March 4, 2022

Abstract

Purpose

This research furthers metadata quality research by providing complementary network-based metrics and insights to analyze metadata records and identify areas for improvement.

Design/methodology/approach

Metadata record graphs apply network analysis to metadata field values; this study evaluates the interconnectedness of subjects within each hub aggregated into the Digital Public Library of America. It also reviews the effects of NACO normalization – simulating revision of values for consistency and breaking up pre-coordinated subject headings – to simulate applying faceted application of subject terminology to Library of Congress Subject Headings.

Findings

Network statistics complement count- or value-based metrics by providing context related to the number of records a user might actually find starting from one item and moving to others via shared subject values. Additionally, connectivity increases through normalization of values to correct or adjust for formatting differences or by breaking pre-coordinated subject strings into separate topics.

Research limitations/implications

This analysis focuses on exact-string matches, which is the lowest common denominator for searching, although many search engines and digital library indexes may use less stringent matching methods. In terms of practical implications for evaluating or improving subjects in metadata, the normalization components demonstrate where resources may be most effectively allocated for these activities (depending on a collection).

Originality

Although the individual components of this research are not particularly novel, network analysis has not generally been applied to metadata analysis. This research furthers previous studies related to metadata quality analysis of aggregations and digital collections in general.

Keywords

Metadata, Digital libraries, Data analysis, Networks

Article classification

Research paper

Introduction

Cultural memory organizations such as galleries, libraries, archives, and museums are actively engaged in making digitized and born digital collections available through digital repositories, digital archives, digital libraries, and other websites. These online collections utilize descriptive metadata enabling users to discover, locate, and view the resources. Subject metadata, in particular, provides users with general entry points for all resource types that are often grouped into topical, form, chronological, and geographic terms. In addition to formal controlled subjects – for example, Library of Congress Subject Headings (LCSH) or the Getty Art & Architecture Thesaurus (AAT) – many organizations make use of uncontrolled keywords, tags, or categories to create access points for their collections.

The Digital Public Library of America (DPLA) aggregates descriptive metadata from a large number of partner hubs across the U.S. comprising over 3,000 cultural heritage institutions who contribute more than 32 million metadata records. Since this metadata is normalized to a single format (Digital Public Library of America, 2017), it allows for larger scale, cross-institution metadata analysis.

The goal of this research project is to investigate the use of the subject metadata field by the DPLA hubs. In addition to standard count- and data-value-based statistics, this study demonstrates the use of metadata record graphs, which is a technique of generating network statistics based on shared metadata field values to connect records. Metadata record graphs are a novel application of network analysis to assess metadata quality by demonstrating the level of connection among a set of metadata records within a collection.

Research questions

This study seeks to answer the following research questions.

RQ1. What is the existing level of connection between metadata records in DPLA hubs?

RQ2. How do statistics for subject metadata record graphs change when string normalization is applied to hubs in the DPLA?

RQ3. How does separating pre-coordinated subject terms affect metadata record graphs?

Literature review

Libraries have spent many decades organizing and evaluating data that documents resources, to ensure that users can be connected to the materials that they need or want. As resources have moved online, particularly into digital libraries or archives, cultural heritage organizations (and others) have continued evaluation work around metadata as the materials became increasingly findable by larger user groups. Some of the earliest web-based evaluation started in the mid-1990s; researchers Moen et al. (1998) reviewed the quality of metadata embedded in U.S. federal government websites. Subsequently, as organizations invested more in digital library initiatives during the early 2000s – for example, the National Science Foundation's National Science Digital Library – significant amounts of metadata migrated to online systems and large aggregations of metadata.

On the philosophical side of metadata quality, a number of researchers have proposed

enveloping frameworks to help conceptualize and discuss specific components of quality. Specifically, the framework proposed by Bruce and Hillmann (2004) provided a foundation for much of the work in this area, although it also echoed work from other projects in the field, such as the *Quality Assurance Framework* (Statistics Canada, 2002) developed for the Canadian Government in 1997 (updated 2002). These frameworks identified categories of quality such as accuracy, timeliness, accessibility, and provenance, which could also be aligned with the four generic user tasks identified by the International Federation of Library Association's (IFLA) Functional Requirements for Bibliographic Records (FRBR) (Riva, Le Boeuf, & Žumer, 2016): find, identify, select, and acquire.

The next step in much of the research extended these concepts at a more granular level and helped to identify possible metrics, such as the the Stvilia and Gasser framework (Stvilia, Gasser, Twidale, & Smith, 2007) which outlines additional points of reference and suggests computational methods of calculating data quality. Other work set out to specifically identify calculable metrics – for example, (Ochoa & Duval, 2009) and (Trippel, Broeder, Durco, & Ohren, 2014) – or to provide an overview and comparison of various frameworks and metrics (Tani, Candela, & Castelli, 2013). Over time, these methods have been reviewed, applied in repository environments, and revised by other researchers (e.g., Palavitsinis, Manouselis, & Sanchez-Alonso, 2014). Additionally, specific aspects of metadata quality have been evaluated using computational metrics, or operationalized within systems, such as the analysis of metadata quality over time in the UNT Libraries' Digital Collections (Tarver, Zavalina, Phillips, Alemneh, & Shakeri, 2014) using a locally-calculated "completeness" score to review record changes. On a larger scale, Király and Büchler also evaluated completeness as a metric to study the quality in the Europeana metadata aggregation (Király & Büchler, 2018).

In addition to the Király and Büchler study, a number of researchers have made use of aggregations for metadata quality research, to compare values from many sources and to have a larger, more varied dataset to draw conclusions. As open metadata sharing protocols

were widely adopted, such as the Open Archive Initiatives and the Protocol for Metadata Harvesting (OAI-PMH), a number of projects made use of these mechanisms to bring metadata together from multiple sources, such as the Digital Library Foundation's (DLF) Aquifer project (2007-2009), OAISter at the University of Michigan (2002-2009), and the Institute of Museum and Library Services (IMLS) Digital Collections and Content (DCC) at the University of Illinois Urbana-Champaign (2002-2009). More recently, the Digital Public Library of America (DPLA) in the U.S. (2013-present) has expanded these efforts, similar to the Europeana project in the European Union (2008-present). Research using these large aggregations has included a number of quantitative analysis studies (Greenberg, 2001; Ward, 2003; Eklund, Miksa, Moen, Snyder, & Polyakov, 2009; Tarver, Phillips, Zavalina, & Kizhakkethil, 2015; Zavalina, Zavalina, & Miksa, 2016) to assess particular quality aspects. Generally, these aspects report on field usage (e.g., Shreeves et al., 2005) or evaluation of values within a particular field, such as the Dublin Core *subject* (Harper, 2016; Tarver et al., 2015) or *date* fields (Zavalina, Alemneh, Kizhakkethil, Phillips, & Tarver, 2015).

In addition to external researchers, Europeana has created the Europeana Task Force on Metadata Quality (Dangerfield & Kalshoven, 2015) to investigate ways of improving metadata quality and analysis within the aggregation. Similar work has not been done by the Digital Public Library of America, though there have been related initiatives such as the Digital Library Federation (DLF) Assessment Interest Group (AIG) Metadata Working Group (http://dlfmetadataassessment.github.io/), which is working to aggregate best practices around metadata quality and assessment.

Despite this previous research, many aspects of metadata quality remain open questions due to the complexity of the issue. Most of the studies related to metadata quality metrics have utilized count- or data-value-based metrics, which are useful to provide a high-level overview of large metadata collections. Currently, few projects that have applied network analysis to metadata analysis; Ochoa and Duval (2009) introduced a network-based metric in their work called the QLink, though they were unable to apply it with real-world collection

data. Since that work, exploratory research using network analysis in the form of metadata record graphs has explored several metadata sets in the UNT Libraries' Digital Collections, including a sampling of six collections (Phillips, Zavalina, & Tarver, 2019, 2020) – to review possible applications of network analysis as a tool – and to specifically analyze subject usage in a large theses and dissertation collection (Phillips, Tarver, & Zavalina, 2019). This work has demonstrated the usefulness of network analysis as a relatively novel, but complementary method, of assessing metadata quality.

This project builds on previous work regarding the use of the subject field (Tarver et al., 2015; Harper, 2016) within DPLA metadata by applying metadata record graphs and network analysis to quantify metadata connectedness in this corpus of records. Unlike the more common count- and value-based statistics, this method provides insight regarding the ability of users to navigate between different resources and to discover new resources that meet their needs based on shared data values, particularly subject terms.

Methodology

The Digital Public Library of America (DPLA) is a metadata aggregator that brings together collections of metadata from organizations across the U.S. and presents them in a single search interface at its website (https://dp.la). In addition to a search system, the DPLA makes its aggregated data available through various application programming interfaces (APIs) as well as allowing for the bulk download of monthly data snapshots. This metadata is provided with a Creative Commons CC0 designation, placing the metadata in the public domain. This project makes use of a snapshot of the DPLA metadata obtained from the bulk download system.

The DPLA data was downloaded in December 2019 and consists of 36,698,952 records from 43 hubs (see Table I 1). The DPLA model for metadata aggregation is a hub-and-spoke model with two types of hubs: content hubs and service hubs. Content hubs provide

metadata records from a single data feed (such as a museum, governmental organization, or other institution maintaining digital holdings) to the DPLA. Hubs that provide additional services (generally content hosting or metadata aggregation) are referred to as service hubs in the DPLA. Service hubs may aggregate metadata from dozens of other institutions within a geographic area and then provide a single data feed to the DPLA. The combined content and service hubs contribute a wide range of records to the DPLA; the smallest contribution comes from the Library of Congress with 4,481 records, up to the National Archives and Records Administration that is contributing 14,466,357 (Fig 1 1). For each of these hubs, count- and data-value-based statistics for the subject field were calculated, including: the total number of metadata records, the number of unique subject instances, the percentage of records that contain at least one subject heading, the unique number of subject headings, and a standardized entropy value for all subjects in a given hub. Additional descriptive statistics related to the number of subject instances per record include the mean and mode fields per record as well as the frequency of the mode instances across the collection.

For this study, the DPLA subject metadata was processed to create a "metadata record graph". This graph, or network, is the result of several stages of processing from the originally downloaded metadata records (discussed in Phillips et al., 2020). The code used for creating the metadata record graphs is available under an open source license in a GitHub repository (Phillips, 2020b). The resulting network represents the metadata records as nodes, and the edges in the graph are connections between those records via shared subject metadata values. To create these network graphs from the DPLA metadata, the following general steps are carried out:

1. Unique identifiers for each metadata record, paired with the data values for its specific element (such as the subject), are output and sorted to alphabetize data values.

2. Record identifiers for a shared data value are grouped with that value. These identifiers represent nodes that are connected by a common data value.

Hub Code	*Hub Name*	*Hub Type*	*Number of Records*
artstor	ARTstor	Content	134,475
bhl	Biodiversity Heritage Library	Content	233,948
bpl	Digital Commonwealth	Service	762,531
cdl	California Digital Library	Service	1,293,065
ct	Connecticut Digital Archive	Service	89,339
david_rumsey	David Rumsey	Content	92,905
dc	District Digital	Service	57,753
digitalnc	North Carolina Digital Heritage Center	Service	476,531
esdn	Empire State Digital Network	Service	403,238
florida	Sunshine State Digital Network	Service	233,821
georgia	Digital Library of Georgia	Service	691,450
getty	J. Paul Getty Trust	Content	99,585
gpo	United States Government Publishing Office	Content	194,690
harvard	Harvard Library	Content	65,739
hathitrust	HathiTrust	Content	2,912,330
ia	Internet Archive	Content	613,172
il	Illinois Digital Heritage Hub	Service	320,575
indiana	Indiana Memory	Service	343,296
kentucky	Kentucky Digital Library	Service	141,677
lc	Library of Congress	Content	4,481
maine	Digital Maine	Service	63,492
maryland	Digital Maryland	Service	102,485
mi	Michigan Service Hub	Service	493,965
minnesota	Minnesota Digital Library	Service	611,868
missouri	Missouri Hub	Service	227,799
mt	Big Sky Country Digital Network	Service	89,737
mwdl	Mountain West Digital Library	Service	1,086,844
nara	National Archives and Records Administration	Content	14,466,347
nypl	The New York Public Library	Content	2,048,825
ohio	Ohio Digital Network	Service	107,539
oklahoma	OKHub	Service	124,944
p2p	Plains to Peaks Collective & Service	Service	309,866
pa	PA Digital	Service	400,100
scdl	South Carolina Digital Library	Service	217,413
sd	Digital Library of South Dakota	Service	58,096
smithsonian	Smithsonian Institution	Content	3,610,489
tennessee	Digital Library of Tennessee	Service	124,896
texas	The Portal to Texas History	Service	1,379,325
usc	University of Southern California. Libraries	Content	1,210,860
virginias	Digital Virginias	Service	58,486
vt	Vermont Green Mountain Digital Archive	Service	51,452
washington	University of Washington	Content	141,873
wisconsin	Recollection Wisconsin	Service	547,650

Table 1: DPLA hub list with record counts

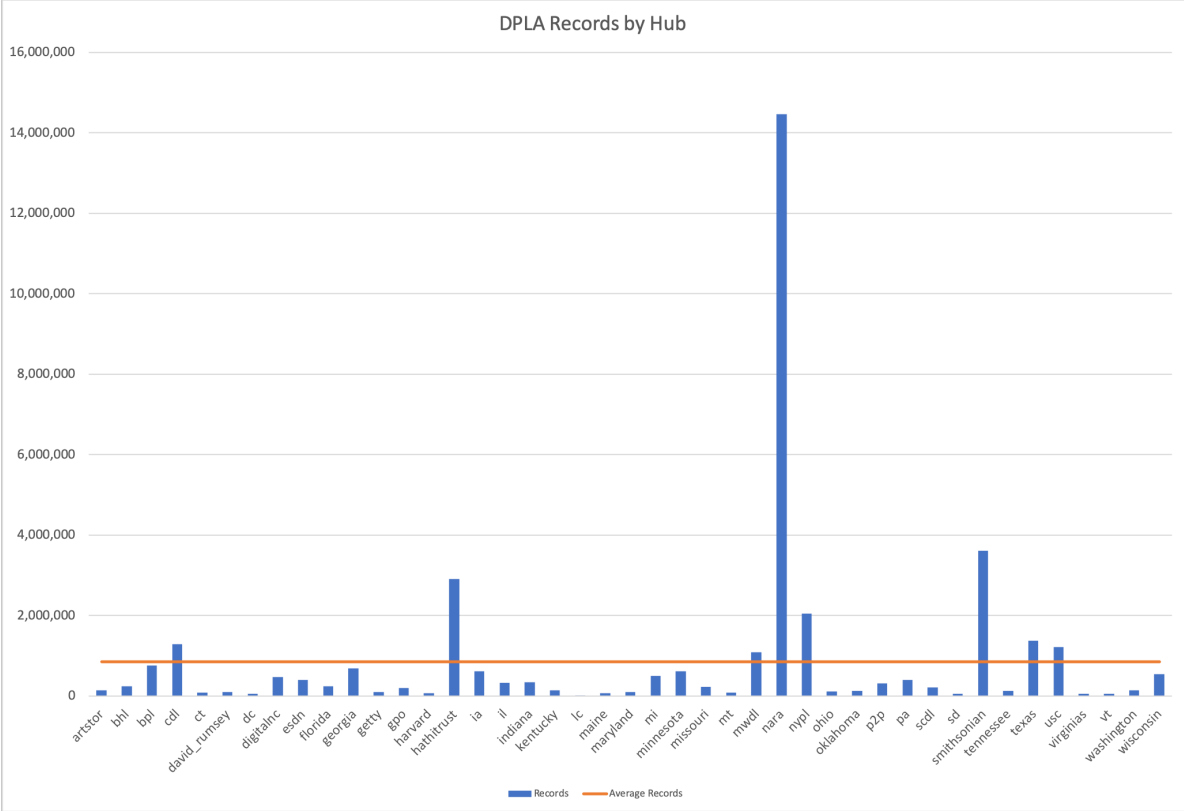

Figure 1: Records per DPLA hub

3. All combinations of these identifiers are generated, output, and sorted.

4. A final adjacency list is created with a metadata record identifier as the key, paired with identifiers for metadata records connected to that record by any shared data value.

Metadata record graphs for the subject field were generated for each of the hubs, providing network statistics including: connected nodes, unconnected nodes, density, average degree, degree mode, and percent of degree mode. Finally, aggregate statistics, such as the Gini coefficient of the degree distribution and the mean value and standard deviation of the Qlink value introduced by Ochoa and Duval (2009), are included.

Previous investigations into the use of metadata record graphs included normalized data values to compare connections among subjects that differed only in punctuation, white space, or capitalization (Phillips, 2020a). The effect of this normalization has not been fully explored for large and disparate collections so the second component of this project involved

processing each hub's data values to measure how the network statistics change with the application of NACO normalization rules (Task Group on Normalization of the PCC Standing Committee on Automation, 2007).

Additionally, for this study, network statistics were generated for a selection of hubs that contain records with a high percentage of pre-coordinated subject headings from the Library of Congress Subject Headings (LCSH) after terms were deconstructed into their individual facet terms. This process seeks to measure the effect of applying FAST (faceted application of subject terminology) to bibliographic records that use LCSH, which divides a pre-coordinated subject value into its constituent parts. For example, **France -- History -- Wars of the Huguenots, 1562-1598 -- Juvenile literature** would be divided into four separate headings with the categories of Topical, Geographic, Period, and Form (Chan, Childress, Dean, O'Neill, & Vizine-Goetz, 2001). Metadata record graphs generated for the modified values provide a way to measure the difference in network connectivity as a result of this conversion.

Findings

Baseline information about subjects in DPLA metadata records can be established using count-based data related to each of the hubs and across the aggregation (see Table II 2). There is a wide range of subject field usage across hubs, including one hub (Library of Congress) that has at least one subject instance in every record, and several others that have only a handful of records without subjects. On the other end of the spectrum, several hubs, such as the J. Paul Getty Trust, have subject instances in fewer than a third of their total records; records contributed by the National Archives and Records Administration (NARA) have zero subject instances in 98.6 percent of the hub's records.

This study furthers similar research from 2015, when there were only 8,012,390 total metadata records in DPLA from 23 hubs (Tarver et al., 2015). Although the number of

Hub	*Total Records*	*Records with subject instances*	*% of records with subject instance*	*Unique data values in field*	*Mean field instances per record*	*Mode field instances per record*	*Frequency of mode instances per record*	*Entropy*
artstor	134,475	111,045	82.6%	28,135	3	1	59%	0.750
bhl	233,948	219,305	93.7%	28,003	4	2	24%	0.589
bpl	762,531	630,108	82.6%	141,730	3	1	31%	0.711
cdl	1,293,065	1,010,380	78.1%	328,105	4	2	19%	0.751
ct	89,339	80,740	90.4%	14,204	3	1	39%	0.713
david_rumsey	92,905	34,949	37.6%	146	1	1	69%	0.765
dc	57,753	40,967	70.9%	19,810	3	3	28%	0.748
digitalnc	476,531	368,036	77.2%	159,697	4	2	20%	0.685
esdn	403,238	338,165	83.9%	122,247	3	3	28%	0.723
florida	233,821	219,098	93.7%	52,302	3	1	36%	0.645
georgia	691,450	690,023	99.8%	165,996	3	2	42%	0.686
getty	99,585	22,798	22.9%	4,727	2	1	71%	0.642
gpo	194,690	193,373	99.3%	236,496	4	3	25%	0.907
harvard	65,739	39,397	59.9%	16,057	3	1	37%	0.740
hathitrust	2,912,330	2,123,460	72.9%	1,082,375	2	1	45%	0.870
ia	613,172	516,751	84.3%	117,841	3	1	37%	0.705
il	320,575	265,228	82.7%	89,341	4	1	21%	0.743
indiana	343,296	322,298	93.9%	67,554	4	4	24%	0.709
kentucky	141,677	4,130	2.9%	1,943	3	2	46%	0.671
lc	4,481	4,481	100.0%	5,859	3	2	56%	0.902
maine	63,492	63,368	99.8%	8,564	13	18	56%	0.429
maryland	102,485	102,411	99.9%	7,990	5	4	33%	0.529
mi	493,965	385,336	78.0%	49,417	4	6	30%	0.561
minnesota	611,868	473,575	77.4%	129,025	3	3	28%	0.701
missouri	227,799	157,154	69.0%	35,193	3	1	25%	0.732
mt	89,737	88,212	98.3%	42,307	5	1	33%	0.725
mwdl	1,086,844	1,041,310	95.8%	229,555	4	1	35%	0.710
nara	14,466,347	195,952	1.4%	3,228	2	1	73%	0.551
nypl	2,048,825	1,388,300	67.8%	63,087	4	2	28%	0.646
ohio	107,539	104,404	97.1%	39,979	3	1	34%	0.785
oklahoma	124,944	113,473	90.8%	48,409	4	4	29%	0.623
p2p	309,866	283,595	91.5%	144,772	4	3	21%	0.789
pa	400,100	356,325	89.1%	196,784	4	1	19%	0.780
scdl	217,413	192,171	88.4%	45,847	3	2	25%	0.721
sd	58,096	57,977	99.8%	10,082	5	4	18%	0.651
smithsonian	3,610,489	3,467,070	96.0%	610,404	7	6	27%	0.597
tennessee	124,896	107,244	85.9%	44,501	4	2	24%	0.760
texas	1,379,325	1,379,310	100.0%	323,673	20	22	24%	0.468
usc	1,210,860	359,737	29.7%	70,283	3	2	50%	0.603
virginias	58,486	54,870	93.8%	12,858	3	2	35%	0.701
vt	51,452	45,931	89.3%	13,291	4	3	25%	0.724
washington	141,873	107,422	75.7%	77,866	1	1	61%	0.860
wisconsin	547,650	533,107	97.3%	186,313	4	3	20%	0.711
dpla (all)	36,698,952	18,292,981	49.8%	4,310,440	5	1	20%	0.637

Table 2: Count-based and data-value-based statistics for the subject field for each DPLA hub before normalization

records and hubs have increased as expected, in the previous data, roughly 22.8 percent (1,827,276 records) had zero subject instances; in the current data, that percentage has increased to 50.1 percent (18,405,971 records) that have zero subject instances. An overview of the subject coverage for the hubs is presented in Figure 2 2. Since this research is interested in determining how well users can find items with similar topical content, this shift means that half of the records in DPLA cannot be found at all based on subject values. This lack of subject data is also expressed by network statistics regarding unconnected nodes (i.e., records that do not connect to any other records via subject terms).

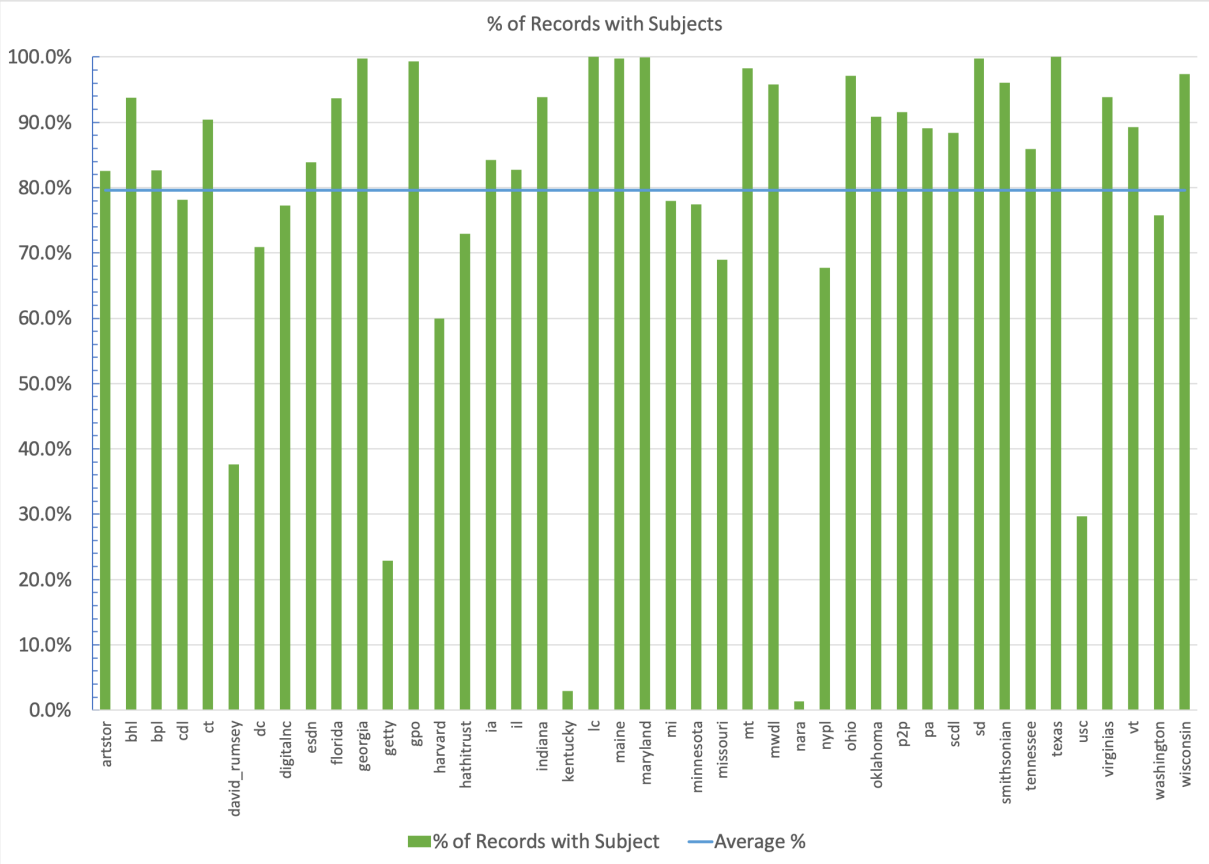

Figure 2: % of hub records with subject

Additionally, metadata record graphs for the hub values provide initial network statistics (see Table III 3) about the interconnectedness of records (i.e., the number of records connected by subject values). These network statistics are based on exact-string matching, so connected notes represent only subject values with identical capitalization, punctuation, and

so on. Overall, about 30 percent of hubs had a node connection rate of at least 90 percent (i.e., 10% or fewer of the hub's records have completely unique or no subject values). In terms of density (a measure of how tightly connected the nodes are), one hub (NARA) had a density of "0" and two others were barely measurable at .0001 (HathiTrust and Kentucky Digital Library). A majority (34 hubs) have a density of .06 or less, but five hubs have extremely higher densities: OKHub and the Smithsonian Institution (.3-.4), The Portal to Texas History (.5+), Digital Maine and Digital Maryland (more than .6). Even though roughly half of the hubs (25) have a connected-node rate of 80 percent or more, the overall density is extremely low, meaning that most records connect to few other records via shared subject values.

Hub	*Total nodes*	*Connected nodes*	*Unconnected nodes*	*% Connected nodes*	*Total edges*	*Density*	*Average degree*	*Degree mode*	*Frequency of degree mode*	*Degree distribution Gini coefficient*	*Qlink mean*	*Qlink std*
artstor	134,475	102,261	32,214	76.0%	130,698,938	0.0145	1,944	0	24%	0.708	0.156	0.225
bhl	233,948	218,313	15,635	93.3%	3,443,366,720	0.1258	29,437	0	7%	0.502	0.264	0.241
bpl	762,531	603,938	158,593	79.2%	2,669,348,894	0.0092	7,001	0	21%	0.735	0.069	0.121
cdl	1,293,065	996,955	296,110	77.1%	5,879,892,002	0.0070	9,095	0	23%	0.711	0.101	0.152
ct	89,339	79,745	9,594	89.3%	120,547,105	0.0302	2,699	0	11%	0.648	0.196	0.257
david_rumsey	92,905	34,947	57,958	37.6%	39,814,589	0.0092	857	0	62%	0.759	0.080	0.134
dc	57,753	39,697	18,056	68.7%	19,385,831	0.0116	671	0	31%	0.751	0.163	0.280
digitalnc	476,531	363,307	113,224	76.2%	2,046,046,032	0.0180	8,587	0	24%	0.726	0.091	0.147
esdn	403,238	329,947	73,291	81.8%	2,514,275,000	0.0309	12,470	0	18%	0.782	0.180	0.341
florida	233,821	217,311	16,510	92.9%	1,375,618,792	0.0503	11,766	42,411	18%	0.652	0.277	0.365
georgia	691,450	686,403	5,047	99.3%	3,789,553,083	0.0159	10,961	50,194	6%	0.641	0.200	0.264
getty	99,585	22,527	77,058	22.6%	19,168,709	0.0039	385	0	77%	0.915	0.070	0.229
gpo	194,690	171,761	22,929	88.2%	44,943,655	0.0024	462	0	12%	0.878	0.043	0.144
harvard	65,739	38,819	26,920	59.1%	54,569,105	0.0253	1,660	0	41%	0.747	0.144	0.233
hathitrust	2,912,330	1,916,304	996,026	65.8%	588,650,461	0.0001	404	0	34%	0.870	0.010	0.038
ia	613,172	500,785	112,387	81.7%	2,680,982,875	0.0143	8,745	0	18%	0.850	0.104	0.250
il	320,575	260,663	59,912	81.3%	425,199,093	0.0083	2,653	0	19%	0.610	0.098	0.114
indiana	343,296	320,478	22,818	93.4%	1,512,969,656	0.0257	8,814	0	7%	0.564	0.256	0.274
kentucky	141,677	3,557	138,120	2.5%	536,508	0.0001	8	0	97%	0.986	0.013	0.099
lc	4,481	3,338	1,143	74.5%	95,686	0.0095	43	0	26%	0.733	0.066	0.109
maine	63,492	63,313	179	99.7%	1,268,022,447	0.6291	39,943	49,927	56%	0.205	0.729	0.357
maryland	102,485	102,328	157	99.8%	3,451,301,495	0.6572	67,352	82,619	16%	0.191	0.778	0.369
mi	493,965	370,490	123,475	75.0%	13,467,568,491	0.1104	54,528	157,160	32%	0.646	0.347	0.452
minnesota	611,868	469,423	142,445	76.7%	2,533,857,490	0.0135	8,282	0	23%	0.706	0.132	0.193
missouri	227,799	152,412	75,387	66.9%	265,864,922	0.0102	2,334	0	33%	0.719	0.088	0.138
mt	89,737	83,611	6,126	93.2%	155,663,041	0.0387	3,469	0	7%	0.533	0.174	0.169
mwdl	1,086,844	1,009,215	77,629	92.9%	18,678,392,054	0.0316	34,372	171,318	16%	0.748	0.201	0.352
nara	14,466,347	195,848	14,270,499	1.4%	2,668,946,979	0.0000	369	0	99%	0.993	0.005	0.058
nypl	2,048,825	1,385,228	663,597	67.6%	88,820,023,009	0.0423	86,703	0	32%	0.683	0.159	0.221
ohio	107,539	102,241	5,298	95.1%	97,731,516	0.0169	1,818	11,333	10%	0.745	0.125	0.233
oklahoma	124,944	112,827	12,117	90.3%	2,513,280,559	0.3220	40,231	69,477	41%	0.414	0.552	0.451
p2p	309,866	279,631	30,235	90.2%	443,658,008	0.0092	2,864	0	10%	0.728	0.097	0.172
pa	400,100	346,617	53,483	86.6%	523,194,162	0.0065	2,615	0	13%	0.666	0.083	0.115
scdl	217,413	188,343	29,070	86.6%	260,483,099	0.0110	2,396	0	13%	0.698	0.147	0.213
sd	58,096	57,853	243	99.6%	232,587,299	0.1378	8,007	6,533	7%	0.488	0.327	0.286
smithsonian	3,610,489	3,460,675	149,814	95.9%	2,473,014,106,985	0.3794	1,369,905	2,168,473	8%	0.358	0.557	0.400
tennessee	124,896	106,706	18,190	85.4%	104,865,833	0.0134	1,679	0	15%	0.578	0.116	0.127
texas	1,379,325	1,379,152	173	100.0%	537,137,079,278	0.5647	778,840	1,021,959	10%	0.228	0.568	0.261
usc	1,210,860	357,596	853,264	29.5%	12,660,034,419	0.0173	20,911	0	70%	0.857	0.105	0.253
virginias	58,486	53,802	4,684	92.0%	59,200,209	0.0346	2,024	0	8%	0.558	0.260	0.272
vt	51,452	45,615	5,837	88.7%	44,012,246	0.0333	1,711	0	11%	0.559	0.189	0.196
washington	141,873	69,065	72,808	48.7%	32,722,466	0.0033	461	0	51%	0.886	0.080	0.218
wisconsin	547,650	511,990	35,660	93.5%	3,049,465,429	0.0203	11,137	0	7%	0.613	0.110	0.130
dpla (all)	36,698,952	17,877,514	18,821,438	48.7%	3,907,450,670,257	0.0058	212,946	0	51%	0.888	0.076	0.210

Table 3: Network statistics for unnormalized subject field from each DPLA hub

NACO normalized values

The next step was to compare statistics after normalizing subject values based on NACO standardization (see Table IV 4). According to NACO rules, every string is normalized in the following ways: all letters are made lowercase, leading and trailing spaces are stripped, diacritics are removed and symbols (except # & +) are switched to blanks, super- and sub-script numbers are changed to digits, and some characters are replaced with ASCII equivalents. This normalization simulates the effect of editing subject values (to account for minor differences in spacing, punctuation, or capitalization) to determine if the effort of changing values would improve the overlap among subjects and increase the chances that items with similar topics can be found together.

Overall, the number of unconnected nodes changed by 13,226 (or an average of 307.6 per hub); that is, after normalizing subjects, more than 13,000 previously unconnected metadata records have at least one connection to another metadata record with the same term. This rate of change varies wildly among the individual hubs. Two hubs – NARA and David Rumsey – have no change in the number of connected nodes as a result of normalization. Another nine had ten or fewer new node connections. In one unusual case, Vermont Green Mountain Digital Archive had 33 fewer connected nodes after normalization, due to a number of subject values that have punctuation but no alphanumeric content (e.g., --); initially, those values connected to one another, but punctuation was stripped from all values during normalization rendering the subject instances "empty". On the other end of the spectrum, two hubs have an increase in connected nodes by more than 1,000 records: the Internet Archive (1,110) and HathiTrust (6,214). Among the other hubs, roughly one-third have a change in unconnected nodes of 11-100 records, 101-200 records, and 201-800 records, respectively.

Aside from unconnected nodes, there were also noticeable changes in connectivity. Hubs that have very many or very few records without subjects tended not to change significantly, and density only changed measurably in about half of the hubs. However, among the records

that did change, eight hubs increased by more than 1 percent in terms of total edges; District Digital increased by nearly 10 percent. Given the initial number of edges, even small percentages represent millions of new edges; for example, 1,854,139 for District Digital and 126,016,824 new edges for the Internet Archive (a 4.7% change). The largest increase in density is reflected in the values for Digital Library of South Dakota, a relatively small hub with 58,096 total records. After NACO normalization, the number of edges increased by 2,568,977 (1.1%), but the density increased by .0015, meaning that the existing values became even more closely connected, although the number of newly connected nodes only increased by four records.

Hub	*Total nodes*	*Connected nodes*	*Unconnected nodes*	*% Connected nodes*	*Total edges*	*Density*	*Average degree*	*Degree mode*	*Frequency of degree mode*	*Degree distribution Gini coefficient*	*Qlink mean*	*Qlink std*
artstor	134,475	102,360	32,115	76.1%	130,880,293	0.0145	1,947	0	24%	0.707	0.156	0.225
bhl	233,948	218,323	15,625	93.3%	3,443,472,372	0.1258	29,438	0	7%	0.502	0.264	0.241
bpl	762,531	604,148	158,383	79.2%	2,683,209,236	0.0092	7,038	0	21%	0.734	0.069	0.121
cdl	1,293,065	997,340	295,725	77.1%	5,920,641,587	0.0071	9,158	0	23%	0.710	0.102	0.152
ct	89,339	79,778	9,561	89.3%	121,818,320	0.0305	2,727	0	11%	0.644	0.197	0.254
david_rumsey	92,905	34,947	57,958	37.6%	40,053,934	0.0093	862	0	62%	0.758	0.081	0.134
dc	57,753	39,714	18,039	68.8%	21,239,970	0.0127	736	0	31%	0.752	0.163	0.282
digitalnc	476,531	363,397	113,134	76.3%	2,052,886,609	0.0181	8,616	0	24%	0.725	0.091	0.147
esdn	403,238	330,715	72,523	82.0%	2,525,933,651	0.0311	12,528	0	18%	0.780	0.181	0.341
florida	233,821	217,439	16,382	93.0%	1,379,954,897	0.0505	11,804	42,493	18%	0.652	0.243	0.319
georgia	691,450	686,486	4,964	99.3%	3,793,112,763	0.0159	10,971	50,194	6%	0.640	0.200	0.264
getty	99,585	22,534	77,051	22.6%	19,201,305	0.0039	386	0	77%	0.915	0.070	0.229
gpo	194,690	171,971	22,719	88.3%	45,028,506	0.0024	463	0	12%	0.877	0.043	0.144
harvard	65,739	38,824	26,915	59.1%	54,574,923	0.0253	1,660	0	41%	0.747	0.144	0.233
hathitrust	2,912,330	1,922,518	989,812	66.0%	591,793,370	0.0001	406	0	34%	0.869	0.010	0.038
ia	613,172	501,895	111,277	81.9%	2,806,999,699	0.0149	9,156	0	18%	0.848	0.099	0.236
il	320,575	260,834	59,741	81.4%	434,053,985	0.0084	2,708	0	19%	0.608	0.100	0.116
indiana	343,296	320,585	22,711	93.4%	1,538,719,215	0.0261	8,964	0	7%	0.555	0.260	0.273
kentucky	141,677	3,579	138,098	2.5%	537,130	0.0001	8	0	97%	0.986	0.013	0.099
lc	4,481	3,339	1,142	74.5%	95,722	0.0095	43	0	25%	0.733	0.066	0.109
maine	63,492	63,318	174	99.7%	1,268,039,066	0.6291	39,943	49,927	56%	0.205	0.729	0.357
maryland	102,485	102,329	156	99.8%	3,451,568,691	0.6572	67,358	82,619	15%	0.191	0.778	0.369
mi	493,965	370,688	123,277	75.0%	13,469,290,438	0.1104	54,535	157,160	32%	0.646	0.347	0.452
minnesota	611,868	469,536	142,332	76.7%	2,548,812,064	0.0136	8,331	0	23%	0.705	0.133	0.194
missouri	227,799	152,568	75,231	67.0%	268,059,612	0.0103	2,353	0	33%	0.717	0.089	0.138
mt	89,737	83,719	6,018	93.3%	156,005,688	0.0387	3,477	0	7%	0.532	0.174	0.169
mwdl	1,086,844	1,009,936	76,908	92.9%	18,694,595,334	0.0317	34,402	171,318	16%	0.747	0.201	0.352
nara	14,466,347	195,848	14,270,499	1.4%	2,668,946,979	0.0000	369	0	99%	0.993	0.005	0.058
nypl	2,048,825	1,385,242	663,583	67.6%	88,822,612,304	0.0423	86,706	0	32%	0.683	0.159	0.221
ohio	107,539	102,275	5,264	95.1%	98,488,266	0.0170	1,832	11,333	10%	0.742	0.126	0.233
oklahoma	124,944	112,857	12,087	90.3%	2,513,343,511	0.3220	40,232	69,477	40%	0.414	0.552	0.451
p2p	309,866	279,857	30,009	90.3%	445,732,952	0.0093	2,877	0	10%	0.725	0.097	0.171
pa	400,100	347,299	52,801	86.8%	526,506,662	0.0066	2,632	0	13%	0.664	0.084	0.115
scdl	217,413	188,461	28,952	86.7%	272,221,618	0.0115	2,504	0	13%	0.696	0.153	0.220
sd	58,096	57,857	239	99.6%	235,156,276	0.1393	8,095	6,533	7%	0.486	0.325	0.283
smithsonian	3,610,489	3,460,778	149,711	95.9%	2,473,025,346,977	0.3794	1,369,912	2,168,473	7%	0.577	0.557	0.400
tennessee	124,896	106,718	18,178	85.4%	109,841,704	0.0141	1,759	0	15%	0.228	0.117	0.127
texas	1,379,325	1,379,158	167	100.0%	537,143,846,793	0.5647	778,850	1,021,960	10%	0.857	0.568	0.261
usc	1,210,860	357,654	853,206	29.5%	12,660,461,192	0.0173	20,912	0	70%	0.558	0.105	0.253
virginias	58,486	53,810	4,676	92.0%	59,213,057	0.0346	2,025	0	8%	0.557	0.260	0.272
vt	51,452	45,582	5,870	88.6%	44,099,035	0.0333	1,714	0	11%	0.885	0.190	0.196
washington	141,873	69,527	72,346	49.0%	33,032,345	0.0033	466	0	51%	0.611	0.080	0.217
wisconsin	547,650	512,520	35,130	93.6%	3,064,919,452	0.0204	11,193	0	6%	0.358	0.111	0.130
dpla (all)	36,698,952	17,900,319	18,798,633	48.8%	3,922,352,974,662	0.0058	213,758	0	51%	0.887	0.075	0.207

Table 4: Network statistics for NACO normalized subject field from each DPLA hub

One observation is that it is possible to have a metadata record that contains *subject* values but not be connected to other records. Unconnected nodes reflect both records that

have no subject values and records containing subject values not shared by any other record within the hub. Records containing one or more subjects that occur only once within the hub's collection is calculable by subtracting the number of connected nodes from the number of records containing subjects (see Table V 5). This provides more information about how the records are connected by highlighting the number of records that have extremely specific subject values versus records that do not have any subject values.

The percentage of records with unconnected subject values is relatively low, even for hubs that have a higher than average number of unconnected nodes, due to the large percentage of records with no subjects. There are two main exceptions: Library of Congress, in which 25.49 percent of the hub's subject values are unique to a single record, and the University of Washington, in which more than 26 percent (37,895) of the records have unconnected subject values. However, a noticeable data point in Table 5 is that a number of collections contain a significantly higher percentage of records with unconnected subject values than the average of 2.9 percent across the 43 hubs (M = 2.9%, SD = 5.6%). For example, University of Washington, Library of Congress, Government Publishing Office, and ARTstor have values of over six percent.

FAST normalized values

Aside from general subject values, this research explored whether other changes might affect collocation of topics. In particular, this analysis has focused on exact string matches, but some hubs use primarily pre-coordinated subject strings – for example, Library of Congress Subject Headings (LCSH) – which may contain similar values based on topics or geographic locations but in multiple combinations that do not match exactly. One way that some organizations address this issue internally is by implementing FAST terms, which break pre-coordinated strings into individual subject values.

For the purposes of this research, the subject values were divided at every double-dash

Hub	*Total Records*	*Records with subjects*	*Records without subjects*	*Network connected nodes*	*Network unconnected nodes*	*Records with unconnected subject values*	*% of records with unique Hub subjects*
artstor	134,475	111,045	23,430	102,360	32,115	8,685	6.46%
bhl	233,948	219,305	14,643	218,323	15,625	982	0.42%
bpl	762,531	630,108	132,423	604,148	158,383	25,960	3.40%
cdl	1,293,065	1,010,380	282,685	997,340	295,725	13,040	1.01%
ct	89,339	80,740	8,599	79,778	9,561	962	1.08%
david_rumsey	92,905	34,949	57,956	34,947	57,958	2	0.00%
dc	57,753	40,967	16,786	39,714	18,039	1,253	2.17%
digitalnc	476,531	368,036	108,495	363,397	113,134	4,639	0.97%
esdn	403,238	338,165	65,073	330,715	72,523	7,450	1.85%
florida	233,821	219,098	14,723	217,439	16,382	1,659	0.71%
georgia	691,450	690,023	1,427	686,486	4,964	3,537	0.51%
getty	99,585	22,798	76,787	22,534	77,051	264	0.27%
gpo	194,690	193,373	1,317	171,971	22,719	21,402	10.99%
harvard	65,739	39,397	26,342	38,824	26,915	573	0.87%
hathitrust	2,912,330	2,123,460	788,870	1,922,518	989,812	200,942	6.90%
ia	613,172	516,751	96,421	501,895	111,277	14,856	2.42%
il	320,575	265,228	55,347	260,834	59,741	4,394	1.37%
indiana	343,296	322,298	20,998	320,585	22,711	1,713	0.50%
kentucky	141,677	4,130	137,547	3,579	138,098	551	0.39%
lc	4,481	4,481	0	3,339	1,142	1,142	25.49%
maine	63,492	63,368	124	63,318	174	50	0.08%
maryland	102,485	102,411	74	102,329	156	82	0.08%
mi	493,965	385,336	108,629	370,688	123,277	14,648	2.97%
minnesota	611,868	473,575	138,293	469,536	142,332	4,039	0.66%
missouri	227,799	157,154	70,645	152,568	75,231	4,586	2.01%
mt	89,737	88,212	1,525	83,719	6,018	4,493	5.01%
mwdl	1,086,844	1,041,310	45,534	1,009,936	76,908	31,374	2.89%
nara	14,466,347	195,952	14,270,395	195,848	14,270,499	104	0.00%
nypl	2,048,825	1,388,300	660,525	1,385,242	663,583	3,058	0.15%
ohio	107,539	104,404	3,135	102,275	5,264	2,129	1.98%
oklahoma	124,944	113,473	11,471	112,857	12,087	616	0.49%
p2p	309,866	283,595	26,271	279,857	30,009	3,738	1.21%
pa	400,100	356,325	43,775	347,299	52,801	9,026	2.26%
scdl	217,413	192,171	25,242	188,461	28,952	3,710	1.71%
sd	58,096	57,977	119	57,857	239	120	0.21%
smithsonian	3,610,489	3,467,070	143,419	3,460,778	149,711	6,292	0.17%
tennessee	124,896	107,244	17,652	106,718	18,178	526	0.42%
texas	1,379,325	1,379,310	15	1,379,158	167	152	0.01%
usc	1,210,860	359,737	851,123	357,654	853,206	2,083	0.17%
virginias	58,486	54,870	3,616	53,810	4,676	1,060	1.81%
vt	51,452	45,931	5,521	45,582	5,870	349	0.68%
washington	141,873	107,422	34,451	69,527	72,346	37,895	26.71%
wisconsin	547,650	533,107	14,543	512,520	35,130	20,587	3.76%
dpla (all)	36,698,952	18,292,981	18,405,971	17,900,319	18,798,633	392,662	1.07%

Table 5: Records with unconnected subject values based on NACO normalized data

(--) subdivision marker, although in true FAST terms some values would change from their LCSH counterparts. Not all cultural heritage organizations assign LCSH terms, so an initial assessment determined which hubs had the highest percentage of subjects containing double-

dashes. The four highest were: HathiTrust (73.64%), University of Washington (77.16%), the Government Publishing Office (85.88%), and the Library of Congress (97.24%).

Although there was no significant change in the network statistics for these hubs after NACO normalization, there were more obvious changes after separating pre-coordinated values (see Table 6). For example, unconnected nodes decreased in every case; records from the Library of Congress decreased from 1,400 to 0 unconnected nodes (a difference of 25%), but the GPO hub decreased from nearly 23,000 records (initially) to just under 1,600 unconnected nodes. Additionally, the number of edges increased exponentially in every case – surpassing 71 trillion in the HathiTrust hub – and there was a similar increase in the total density. This shift may be most obvious in the difference for "average degree", which is the average number of other records a user would find by clicking on any subject in the hub. For three of the hubs, the initial average degree was around 400-460 records; for the Library of Congress, the average was only 43. After dividing pre-coordinated terms, the average degree increased significantly in every case, though there is a range among the first three hubs from an average of more than 14,000 (University of Washington) to more than 61,000 (Government Printing Office). Even the Library of Congress increased to an average of more than 4,000 records. In each of these cases, that means that pre-coordinted terms contained large numbers of overlapping topics that could not be found together with string matching.

Overall, this data suggests that moving from pre-coordinated strings to separate topical, geographic, chronological, or other terms would significantly increase the connectivity among records with similar content. It seems reasonable that this would also likely increase connections between records that contain LCSH terms and others that use less complex subject terms, or keywords.

Hub	*Total nodes*	*Connected nodes*	*Unconnected nodes*	*% Connected nodes*	*Total edges*	*Density*	*Average degree*	*Degree mode*	*Frequency of degree mode*	*Degree distribution Gini coefficient*	*Qlink mean*	*Qlink std*
gpo-unnormalized	194,690	171,761	22,929	88.2%	44,943,655	0.0024	462	0	12%	0.878	0.043	0.144
gpo-naco	194,690	171,971	22,719	88.3%	45,028,506	0.0024	463	0	12%	0.877	0.043	0.144
gpo-fast	194,690	193,103	1,587	99.2%	6,003,432,013	0.317	61,672	0	1%	0.431	0.517	0.434
hathitrust-unnormalized	2,912,330	1,916,304	996,026	65.8%	588,650,461	0.0001	404	0	34%	0.870	0.010	0.038
hathitrust-naco	2,912,330	1,922,518	989,812	66.0%	591,793,370	0.0001	406	0	34%	0.869	0.010	0.038
hathitrust-fast	2,912,330	2,068,273	844,057	71.0%	71,596,986,200	0.017	49,168	0	29%	0.800	0.084	0.164
lc-unnormalized	4,481	3,338	1,143	74.5%	95,686	0.0095	43	0	26%	0.733	0.066	0.109
lc-naco	4,481	3,339	1,142	74.5%	95,722	0.0095	43	0	25%	0.733	0.066	0.109
lc-fast	4,481	4,481	0	100.0%	9,256,454	0.922	4,131	4,057	30%	0.054	0.923	0.119
washington-unnormalied	141,873	69,065	72,808	48.7%	32,722,466	0.0033	461	0	51%	0.886	0.080	0.218
washington-naco	141,873	69,527	72,346	49.0%	33,032,345	0.0033	466	0	51%	0.611	0.080	0.217
washington-fast	141,873	102,388	39,485	72.2%	1,040,630,085	0.1034	14,670	0	28%	0.659	0.276	0.374

Table 6: Network statistics for FAST normalized subject field from each select hubs

Significance

Subject metadata is particularly useful for user access to materials because topics can be applied regardless of material types and do not necessarily rely on contextual information that may not be known (e.g., creator, creation date, etc.). For these same reasons, it is one of the few descriptive metadata fields that could be changed or enhanced to improve metadata quality without requiring external information or research; this means that quality analysis could be used to actively improve records. Finally, subject metadata is crucial in situations where there is not full text to fall back on such as collections of photographs, audiovisual collections, or in applications where just metadata is aggregated, such as the DPLA.

This study applies the concept of metadata record graphs to metadata records in DPLA, to demonstrate how this approach might supplement other quality metrics for analyzing subject metadata at scale. Although these network metrics are generally useful, as discussed in the Findings, several of the metrics calculated in the three stages of this study did not lend themselves to easily actionable interpretation. For example, the Qlink distributions (Ochoa & Duval, 2009) and Gini coefficient of distributions were challenging to interpret. This study found they were not readily helpful for metadata analysis work. This is consistent with the findings in (Phillips, 2020a), which further discussed the various network metrics and their possible applications in the management of metadata and metadata analysis.

This study relies on the premise that subjects should connect based on exact string matching, although many search engines automatically tend to correct for some differences

in capitalization, spacing, and even punctuation. However, especially within an environment where metadata values are aggregated or indexed outside their original context, assuming exact string matching provides the greatest amount of consistency and best measure of connectivity. Additionally, the NACO normalizations and LCSH transformations simulate possible changes that could be applied to existing subject values and show how they would affect overall interconnectedness of records, potentially leading users to additional related materials. Changes in network metrics through processing and string value normalization can provide guidance to metadata practitioners by alerting them to situations where metadata values contain unexpected variation.

At present, the DPLA is the largest collection of metadata for cultural heritage organizations in the U.S. and this project represents a big data approach that seeks to investigate the subject values of all records in the aggregation instead of a subset chosen by sampling. This approach also allows for comparisons among different metadata practices, whether by individual institutions (content hubs) or intermediary aggregators (service hubs). Although metadata quality in general, and subjects specifically, are important to institutions, often resources for correcting or enhancing metadata are limited. This research suggests that network analysis may be an additional tool for identifying records that would benefit the most from review or editing, or to guide decisions about provisioning resources for metadata work.

Conclusion

There are several explicit take-aways from the data in this study. First, applying network statistics to subject values provides information to complement other count- or value-based metrics. This is most obvious in the way that modifications to values may change unique value counts, while there is a significantly different change in the edge count and density. The first reflects (roughly) the number of topics, while the second provides context about what a user might actually find starting from one record and moving to others. Second,

normalization of values to correct or adjust for differences in capitalization, punctuation, and so on does have an effect on density, although the specific amount of change depends on the collection. Third, shorter, separate topics rather than joined, pre-coordinated subject strings provide more overlap among similar subjects. This is particularly true since long, pre-coordinated strings may be more difficult for search engines to collocate when they are not exact string matches.

All of these findings suggest that nearly any digital collection may find benefits in apportioning resources to adjust or add subject values in metadata records as a way of improving topical-based retrieval.

References

Bruce, T. R., & Hillmann, D. (2004). The continuum of metadata quality: defining, expressing, exploiting. In D. Hillman & E. L. Westbroooks (Eds.), *Metadata in practice.* Chicago: ALA Editions.

Chan, L. M., Childress, E., Dean, R., O'Neill, E. T., & Vizine-Goetz, D. (2001). A faceted approach to subject data in the dublin core metadata record. *Journal of Internet Cataloging*, *4*(1-2), 35-47. doi: 10.1300/J141v04n01_05

Dangerfield, M., & Kalshoven, L. (2015). *Report and recommendations from the task force on metadata quality.* Europeana. Retrieved from `https://pro.europeana.eu/post/metadata-quality-task-force-report`

Digital Public Library of America. (2017, 12 7). *Metadata application profile, version 5.0* (Tech. Rep.). Retrieved from `http://dp.la/info/map`

Eklund, A. P., Miksa, S. D., Moen, W. E., Snyder, G., & Polyakov, S. (2009). Comparison of MARC content designation utilization in OCLC worldcat records with national, core, and minimal level record standards. *Journal of Library Metadata*, *9*(1-2), 36–64. Retrieved from `https://doi.org/10.1080/19386380903095073` doi: 10.1080/19386380903095073

Greenberg, J. (2001). A quantitative categorical analysis of metadata elements in image-applicable metadata schemas. *Journal of the American Society for Information Science and Technology*, *52*(11), 917–924. doi: 10.1002/asi.1170

Harper, C. (2016). Metadata analytics, visualization, and optimization: Experiments in statistical analysis of the Digital Public Library of America (DPLA). *The Code4Lib Journal*, *33*. Retrieved from `https://journal.code4lib.org/articles/11752`

Király, P., & Büchler, M. (2018). Measuring completeness as metadata quality metric in Europeana. In *IEEE international conference on big data, big data 2018, seattle, wa, usa, december 10-13, 2018* (pp. 2711–2720). Retrieved from `https://doi.org/10.1109/BigData.2018.8622487` doi: 10.1109/BigData.2018.8622487

Moen, W. E., Stewart, E. L., & McClure, C. R. (1998, 4). Assessing metadata qual-

ity: findings and methodological considerations from an evaluation of the US Government Information Locator Service (GILS). In *Proceedings IEEE International Forum on Research and Technology Advances in Digital Libraries* (pp. 246–255). doi: 10.1109/ADL.1998.670425

Ochoa, X., & Duval, E. (2009). Automatic evaluation of metadata quality in digital repositories. *International Journal on Digital Libraries*, *10*(2), 67–91. doi: 10.1007/s00799-009-0054-4

Palavitsinis, N., Manouselis, N., & Sanchez-Alonso, S. (2014). Metadata quality in digital repositories: Empirical results from the cross-domain transfer of a quality assurance process. *Journal of the Association for Information Science and Technology*, *65*(6), 1202–1216. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.23045 doi: 10.1002/asi.23045

Phillips, M. E. (2020a). *Exploring the use of metadata record graphs for metadata assessment* (Doctoral dissertation, University of North Texas, Denton, Texas). Retrieved from https://digital.library.unt.edu/ark:/67531/metadc1707350/

Phillips, M. E. (2020b). *metadata-record-graphs.* https://github.com/vphill/metadata-record-graphs. GitHub.

Phillips, M. E., Tarver, H., & Zavalina, O. L. (2019). Using metadata record graphs to understand controlled vocabulary and keyword usage for subject representation in the UNT theses and dissertations collection. *Cadernos BAD*(1), 61–76.

Phillips, M. E., Zavalina, O. L., & Tarver, H. (2019). Using metadata record graphs to understand digital library metadata. *International Conference on Dublin Core and Metadata Applications*, 49–58. Retrieved from https://dcpapers.dublincore.org/pubs/article/view/4237

Phillips, M. E., Zavalina, O. L., & Tarver, H. (2020). Exploring the utility of metadata record graphs and network analysis for metadata quality evaluation and augmentation. *International Journal of Metadata, Semantics and Ontologies*, *14*(2). doi: 10.1504/IJMSO.2020.108326

Riva, P., Le Boeuf, P., & Žumer, M. (2016). *FRBR-library reference model* (Tech. Rep.). Retrieved from https://www.ifla.org/files/assets/cataloguing/frbr-lrm/frbr-lrm_20160225.pdf

Shreeves, S. L., Knutson, E. M., Stvilia, B., Palmer, C. L., Twidale, M. B., & Cole, T. W. (2005). Is quality metadata shareable metadata? the implications of local metadata practices for federated collections. In H. A. Thompson (Ed.), *Proceedings of the twelfth*

national conference of the association of college and research libraries (p. 223-237). Chicago, IL: Association of College and Research Libraries.

Statistics Canada. (2002). *Statistics canada's quality statistics canada's quality assurance framework.* Minister of Industry. Retrieved from https://unstats.un.org/unsd/dnss/docs-nqaf/Canada-12-586-x2002001-eng.pdf

Stvilia, B., Gasser, L., Twidale, M. B., & Smith, L. C. (2007). A framework for information quality assessment. *Journal of the American society for information science and technology*, *58*(12), 1720–1733. doi: 10.1002/asi.20652

Tani, A., Candela, L., & Castelli, D. (2013). Dealing with metadata quality: The legacy of digital library efforts. *Information Processing & Management*, *49*(6), 1194-1205. doi: 10.1016/j.ipm.2013.05.003

Tarver, H., Phillips, M. E., Zavalina, O., & Kizhakkethil, P. (2015). An exploratory analysis of subject metadata in the Digital Public Library of America. *International Conference on Dublin Core and Metadata Applications*, 30–40.

Tarver, H., Zavalina, O., Phillips, M. E., Alemneh, D., & Shakeri, S. (2014). How descriptive metadata changes in the UNT Libraries' Collections: A case study. In *International Conference on Dublin Core and Metadata Applications* (pp. 43–52). Retrieved from http://dcpapers.dublincore.org/pubs/article/view/3701

Task Group on Normalization of the PCC Standing Committee on Automation. (2007, 11). *Authority file comparison rules* (Tech. Rep.). Retrieved from https://www.loc.gov/aba/pcc/naco/documents/SCA_PccNormalization_Final_revised.pdf

Trippel, T., Broeder, D., Durco, M., & Ohren, O. (2014, 5). Towards automatic quality assessment of component metadata. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 3851–3856). Reykjavik, Iceland: European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2014/pdf/1011_Paper.pdf

Ward, J. (2003, 5). A quantitative analysis of unqualified Dublin Core Metadata Element Set usage within data providers registered with the open archives initiative. In *Joint Conference on Digital Libraries, 2003. Proceedings.* (pp. 315–317). doi: 10.1109/JCDL.2003.1204883

Zavalina, O. L., Alemneh, D. G., Kizhakkethil, P., Phillips, M. E., & Tarver, H. (2015). Extended date/time format (EDTF) in the Digital Public Library of America's metadata: Exploratory analysis. *Proceedings of the Association for Information Science and Technology*, *52*(1), 1-5. doi: 10.1002/pra2.2015.145052010066

Zavalina, O. L., Zavalina, V., & Miksa, S. D. (2016). Quality over time: A longitudinal quantitative analysis of metadata change in RDA-based MARC bibliographic records representing video resources. *Proceedings of the Association for Information Science and Technology*, *53*(1), 1–5. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/pra2.2016.14505301125 doi: 10.1002/pra2.2016.14505301125