DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications 2019

Using Metadata Record Graphs to Understand Digital Library Metadata

Mark E. Phillips
UNT Libraries, USA
mark.phillips@unt.edu

Oksana L. Zavalina
UNT College of Information, USA
oksana.zavalina@unt.edu

Hannah Tarver
UNT Libraries, USA
hannah.tarver@unt.edu

Abstract

Digital collections in cultural heritage institutions are increasingly digitizing physical items, collecting born-digital items, and making these resources available online. Metadata plays a crucial role in the discovery and management of these collections, which makes it important to identify areas of metadata improvement. A number of frameworks and associated metrics support metadata evaluation but this paper focuses on a less-studied aspect of accessibility by using traditional network analysis to understand the connections between metadata records created through shared data values, in elements such as subject or creator. The goal of the research reported in this paper is to investigate potential uses of network analysis and to determine which metrics hold the most promise in effective assessment of metadata at the database or collection level. We introduce the Metadata Record Graph and analyze how it can be used to better understand various-sized collections of metadata.

Keywords: Metadata Record Graphs; networking metrics; metadata quality; digital libraries

1. Introduction

Cultural heritage institutions including archives, galleries, libraries, and museums are increasingly turning to digital technologies and infrastructures to manage their growing collections. These resources are collected from two major sources: born-digital resources (items created digitally and published in electronic format, often to the web) and digitization of analog material. Most institutions maintain their collections in digital libraries or digital asset management systems that help organize, provide access to, and preserve the digital resources. These systems use metadata, or data about the digital resources, to allow for the discovery, identification, and delivery of resources to users. Metadata also assists with inventory, tracking, and other internal management activities. Thus, the quality of metadata greatly affects usability of collections.

Metadata quality has received much attention over the past few decades. A number of metadata quality frameworks and broader information quality frameworks are used by the community to help guide investigations (e.g., Bruce & Hillman, 2004; Stvilia, Gasser, Twidale & Smith, 2007, etc.). Researchers have also suggested measurements or metrics for the different components of theses frameworks for evaluation (Stvilia, 2006; Ochoa & Duval, 2009; Király, Stiller, Charles, Bailer & Freire, 2018). An analysis of the frameworks, their parameters, and the metrics proposed by each is included in Tani, Candela, & Castelli (2013).

Many of the metrics operationalized in metadata quality frameworks use aggregate values to calculate descriptive statistics for different element instances in metadata records that make up targeted digital collections (Ward, 2003; Stvilia, Gasser, Twidale, Shreeves & Cole, 2004). Other approaches to develop a better understanding of the quality of metadata focused on specific metadata element(s) such as subject, date, or description (Harper, 2016; Tarver, Phillips, Zavalina, & Kizhakkethil, 2015; Zavalina, Phillips, Alemneh, Tarver, & Kizhakkethil, 2015; Tarver, Zavalina, & Phillips, 2017). These efforts worked directly with metadata records and their data values as the units of their analysis instead of higher-level aggregated counts.

Metadata quality analysis is not a new area of research; however, many of the metrics attempt to evaluate quality framework criteria that are straightforward to calculate, such as completeness or accuracy of value formatting and compliance with vocabularies. One area that has not been

https://doi.org/10.23106/dcmi.952141695

adequately represented in the literature regards metadata quality metrics based on the relationships between metadata records. In the past ten years there has been a growing interest in "linked data" and more specifically for cultural heritage institutions, "linked open data" or LOD. This movement has encouraged metadata managers to begin thinking about metadata records as collections of relationships in a network and not only as descriptive properties for local resources. By treating metadata records as nodes in a network that are connected by the data values as edges, tools and algorithms from the field of network analysis can be leveraged to learn more about metadata in specific collections. For example, imagine a very simple collection containing two digital resources (Figure 1). If both representative metadata records have a data value of "Shark" in the subject element, then the metadata records are linked because it would be possible in an online system to travel between one record and the other using the subject path of "Shark."

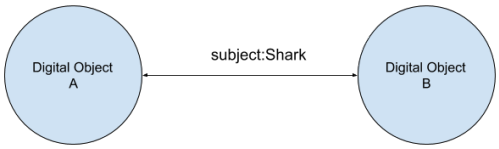

FIG 1. Simple representation of a Metadata Record Graph

For over 20 years, linked metadata records have been presented to users in online displays of library catalogs and digital library systems as a hypertext link that allows a user to travel from a metadata record to one or more metadata records for other resource(s) containing that same data value (Babu & O'Brien, 2000; Task Force on Guidelines for OPAC Displays, 2008). Though a common interaction, there has been insufficient research to understand how relationships between connected data values can help manage metadata collections. Further work in this area is expected to lead to new metadata quality metrics that can be used to assess metadata quality related to accessibility or broader findability in a wide range of digital collections. This research seeks to understand networks created by tens of thousands of metadata records connected by shared data values.

The following research questions guide this investigation:

RQ1: How can network analysis be used to help assess metadata quality in digital library collections?

RQ2: Which metrics from network theory can be used as metadata quality indicators for networks of metadata records?

RQ3: What metadata elements are most appropriate to create connections between metadata records?

2. Methods

Six selected collections of metadata records were downloaded from a digital repository and processed to create what we are referring to as "Metadata Record Graphs." These network graphs interpret metadata records as nodes; the edges represent connections between those records based on commonalities such as a shared subject metadata element data value, a shared creator metadata element data value, etc. The resulting Metadata Record Graph is an undirected graph with bidirectional edges (i.e., it is possible to move from a metadata record to another metadata record through an edge in either direction).

We generated a Metadata Record Graph for every metadata element in each collection by taking the following steps:

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

https://doi.org/10.23106/dcmi.952141695

2. Record identifiers for a shared data value were grouped with that value. These identifiers represent nodes that are connected by a common data value.
3. All combinations of these identifiers were generated, output, and sorted.
4. A final adjacency list was created with a metadata record identifier as the key, paired with identifiers for metadata records connected to that record by any shared data value.

These steps result in a final Metadata Record Graph, which is used to calculate various network statistics for this research.

The first characteristic we assessed was the density of the graph or network, which is a calculation of the actual connections (or edges) in a graph divided by the potential connections (possible edges). Density provides an indication of how well a collection of metadata records is connected; it is represented as a number between 0 and 1. A collection of metadata records with a density of 0 would not share any data values; i.e., every record in the network contains only unique data values. To the contrary, the density of a network consisting of metadata records that all share a common data value (e.g., the language of "English"), would be 1. Another common network analysis metric is the degree of a network's node, or the number of edges that intersect with a node. Once the degree for each of the nodes in the Metadata Record Graph is known, one can calculate the average degree and the degree distribution of the graph itself that provide an estimate of how connected the entire network is. The average degree metric is the average of degrees of nodes in the network (i.e., the average number of items that would be retrieved if data values are represented as clickable links); this value is rounded to a whole number of nodes. The degree distribution metric is the probability of a given degree occurring in the network.

In addition to the metrics discussed above, we calculated the Qlink metric proposed in the metadata quality literature (Ochoa & Duval, 2009) for each Metadata Record Graph node:

$$Qlink = \frac{links(instance_k)}{\max_{i=1}^{N}(links(instance_i))}$$

In this equation, $links(instance_k)$ represents the number of connections to or from the metadata record and N is the number of resources in the collection. This Qlink metric was analyzed similarly to the degree metric (with standard descriptive statistics) and visualized as a Qlink distribution. To support comparison of the proposed network-based metrics of metadata connectedness among different collections, we included in our analysis the metrics traditionally used in metadata evaluation:

- general count and data-value-based statistics for each metadata element in a collection of metadata records, and
- standardized entropy (Stvilia, Gasser, Twidale, Shreeves & Cole, 2004), which is calculated as a number between 0 and 1 representing the amount of unique or duplicated information present in a metadata element.

For the purposes of this research, the network analysis assumed exact string matches and did not attempt to account for differences in semantic meanings or other information not evident in the string text. This is meant to simulate hyperlinks or string searches used by most digital libraries.

3. Data

This study analyzed a subset of records from the UNT Libraries' Digital Collections, which can be accessed via The Portal to Texas History (https://texashistory.unt.edu/) and the UNT Digital Library (https://digital.library.unt.edu/). Collectively, the Digital Collections encompass roughly 2.7 million items in a single administrative system, described with the UNTL metadata format (https://digital2.library.unt.edu/untl.xsd). Metadata elements include the standard 15 Dublin Core elements as well as 7 others (collection, institution, degree, citation, primary source, note, and meta)

https://doi.org/10.23106/dcmi.952141695

defined locally (https://library.unt.edu/digital-projects-unit/metadata/). This research only analyzed the data values in the Dublin Core elements.

Metadata in the Digital Collections is available using the Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH) at https://texashistory.unt.edu/oai/ (Portal) and https://digital.library.unt.edu/oai/ (Digital Library). OAI-PMH allows for programmatic harvesting of metadata from repositories and has been an important component of the scholarly repository landscape since its release in 2002. In March 2019, we harvested records in their native UNTL format from 6 collections that represent a broad range of material types. Table 1 shows the names and codes of collections selected for our analysis, the number of metadata records harvested (publicly-visible items in each collection), and a brief description of the scope.

TABLE 1: Overview of Digital Collections Analyzed in the Study

Collection Name	Collection Code	Metadata Records	Collection Description
College of Music Recordings	COMR	6,398	Audio/video recordings of recitals from UNT
Technical Report Archive and Image Library	TRAIL	25,132	U.S. government-sponsored technical reports
Texas Patents	TXPT	14,354	U.S. patents submitted by Texas inventors
Texas State Publications	TXPUB	11,219	Materials published by Texas state agencies
UNT Theses and Dissertations	UNTETD	19,292	Born-digital and digitized UNT theses/dissertations
UNT Photography Collection	UNTPC	16,659	Images documenting UNT history

4. Results

Tables below report statistics from Metadata Record Graphs created for each metadata element in metadata records of two collections — TRAIL and UNTETD — chosen because they represent very different kinds of collections. Table 2 provides traditionally-calculated metrics for the TRAIL collection based on counts of metadata elements, as well as standardized entropy and Metadata Record Graph density metrics.

TABLE 2: Data-Value-Based Overview of the TRAIL Collection Metadata (n=25,132)

Element Name	Records with Element Instances	% of Records with Element Instances	Unique Data Values in Element	Mean Element Instances Per Record	Mode Element Instances Per Record	Frequency of Mode Instances Per Record	Entropy	Graph Density for Metadata Element in Collection
title	25,132	100%	41,977	3	2	49%	0.763	0.1084
creator	24,526	98%	17,990	2	1	45%	0.935	0.0010
contributor	23,193	92%	2,451	1	1	70%	0.539	0.1092
publisher	10,940	44%	220	1	1	100%	0.533	0.0239
date	25,008	100%	5,087	1	1	100%	0.903	0.0009
language	25,132	100%	3	1	1	100%	0.001	0.9998
description	25,132	100%	32,558	2	2	100%	0.908	0.0000
subject	25,132	100%	21,147	3	2	53%	0.839	0.0194
coverage	7,388	29%	3,021	2	1	59%	0.754	0.0074
source	977	4%	468	1	1	100%	0.796	0.0000
relation	536	2%	516	1	1	87%	0.961	0.0000
rights	13,793	55%	6	3	3	100%	0.631	0.3012
resourceType	25,132	100%	15	1	1	100%	0.078	0.9232
format	25,132	100%	3	1	1	100%	0.121	0.9432
identifier	24,952	99%	80,335	4	4	57%	0.980	0.0022

Table 3 shows the network statistics calculated from the Metadata Record Graphs for each metadata element in the TRAIL collection: the number of connected and unconnected nodes, total number of edges, density, average degree, and the mean value and standard deviation for the Qlink metric.

https://doi.org/10.23106/dcmi.952141695

TABLE 3: Network and Qlink Statistics for the TRAIL Collection Metadata (n=25,132)

Element Name	Connected Nodes	Unconnected Nodes	Total Edges	Density	Average Degree	Qlink Mean	Qlink Std
title	25,104	28	34,234,877	0.108	2,724	0.49	0.31
creator	20,842	4,290	305,040	0.001	24	0.05	0.15
contributor	22,904	2,228	34,484,643	0.109	2,744	0.29	0.24
publisher	10,853	14,279	7,552,428	0.024	601	0.16	0.29
date	22,865	2,267	299,764	0.001	24	0.08	0.13
language	25,132	0	315,720,759	1.000	25,125	1.00	0.01
description	19,021	6,111	4,749	0.002	57	0.23	0.27
subject	24,066	1,066	6,140,004	0.019	489	0.11	0.19
coverage	6,836	18,296	2,325,844	0.007	185	0.06	0.19
source	535	24,597	13,137	0.000	1	0.01	0.09
relation	116	25,016	455	0.000	0	0.00	0.03
rights	13,793	11,339	95,116,528	0.301	7,569	0.55	0.50
resourceType	25,129	3	291,530,610	0.923	23,200	0.96	0.19
format	25,132	0	297,861,427	0.943	23,704	0.97	0.16
identifier	5,645	19,487	683,004	0.002	54	0.05	0.20

The same two sets of metrics obtained for the UNTETD collection are presented in Table 4 (traditional calculations) and Table 5 (network statistics).

TABLE 4: Data-Value-Based Overview of the UNTETD Collection Metadata (n=19,292)

Element Name	Records with Element Instances	% of Records with Element Instances	Unique Data Values in Element	Mean Element Instances Per Record	Mode Element Instances Per Record	Frequency of Mode Instances Per Record	Entropy	Graph Density for Metadata Element in Collection
title	19,292	100%	19,290	1	1	100%	1.000	0.0000
creator	19,292	100%	18,500	1	1	100%	0.998	0.0000
contributor	17,872	93%	6,111	3	3	36%	0.877	0.0071
publisher	19,291	100%	8	1	1	100%	0.528	0.3893
date	19,285	100%	730	1	1	87%	0.862	0.0053
language	19,292	100%	5	1	1	100%	0.016	0.9971
description	19,176	99%	25,385	2	2	62%	0.978	0.0001
subject	19,284	100%	62,615	5	5	18%	0.953	0.0023
coverage	4,264	22%	1,059	1	1	78%	0.683	0.0054
source	0	0%	0	0	0	100%	0.000	0.0000
relation	259	1%	391	2	1	70%	1.000	0.0000
rights	19,292	100%	17,253	4	4	95%	0.404	1.0000
resourceType	19,292	100%	1	1	1	100%	0.000	1.0000
format	19,292	100%	1	1	1	100%	0.000	1.0000
identifier	17,233	89%	44,223	3	3	39%	0.999	0.0000

TABLE 5: Network and Qlink Statistics for the UNTETD Collection Metadata (n=19,292)

Element Name	Connected Nodes	Unconnected Nodes	Total Edges	Density	Average Degree	Qlink Mean	Qlink Std
title	74	19,218	55	0.000	0	0.00	0.03
creator	1,577	17,715	803	0.000	0	0.03	0.09
contributor	17,851	1,441	1,323,847	0.007	137	0.16	0.15
publisher	19,287	5	72,439,560	0.389	7,510	0.75	0.30
date	19,221	71	993,288	0.005	103	0.34	0.14
language	19,292	0	185,541,874	0.997	19,235	1.00	0.04
description	7,418	11,874	25,384	0.000	3	0.10	0.19
subject	17,593	1,699	423,445	0.002	44	0.08	0.11
coverage	3,990	15,302	1,002,075	0.005	104	0.06	0.18
source	0	19,292	0	0.000	0	0.00	0.00
relation	0	19,292	0	0.000	0	0.00	0.00
rights	19,292	0	186,080,599	1.000	19,291	1.00	0.00
resourceType	19,292	0	186,080,986	1.000	19,291	1.00	0.00
format	19,292	0	186,080,986	1.000	19,291	1.00	0.00
identifier	269	19,023	5,140	0.000	1	0.01	0.07

https://doi.org/10.23106/dcmi.952141695

A full set of count-based and network-based statistics results for each of the Dublin Core metadata elements in each collection is presented in an open-access companion dataset (Phillips, Tarver & Zavalina, 2019). For brevity, we have elected to include in this paper comparative statistics for the Metadata Record Graphs created by the subject element from each of the six collections, as the subject is one of the few metadata elements that could be consistently enhanced or modified to affect the network values. Table 6 presents a combination of both network statistics and traditionally-calculated statistics for this metadata element.

TABLE 6: Subject Metadata Element: Statistics for Six Collections

Collection	Records	Unique Data Values	Entropy	Network-Graph Statistics			
				Unconnected Nodes	Density	Average Degree	Qlink Mean
COMR	6,398	3,150	0.791	146	0.060	382	0.22
TRAIL	25,132	21,147	0.839	1,066	0.019	489	0.11
TXPT	14,354	12,268	0.588	1	0.982	14,092	0.99
TXPUB	11,219	15,325	0.772	25	0.134	1,508	0.32
UNTETD	19,292	62,615	0.953	1,699	0.002	44	0.08
UNTPC	16,659	6,170	0.582	1	0.924	15,393	0.93

Figure 2 presents Qlink distributions for the Metadata Record Graphs of the subject element from each of the six collections on both standard and log scales. It should be noted that the Qlink distribution and a more traditional degree distribution both provide the same visual plot but with different scales. The Qlink acts as a scale normalization for easier comparisons.

https://doi.org/10.23106/dcmi.952141695

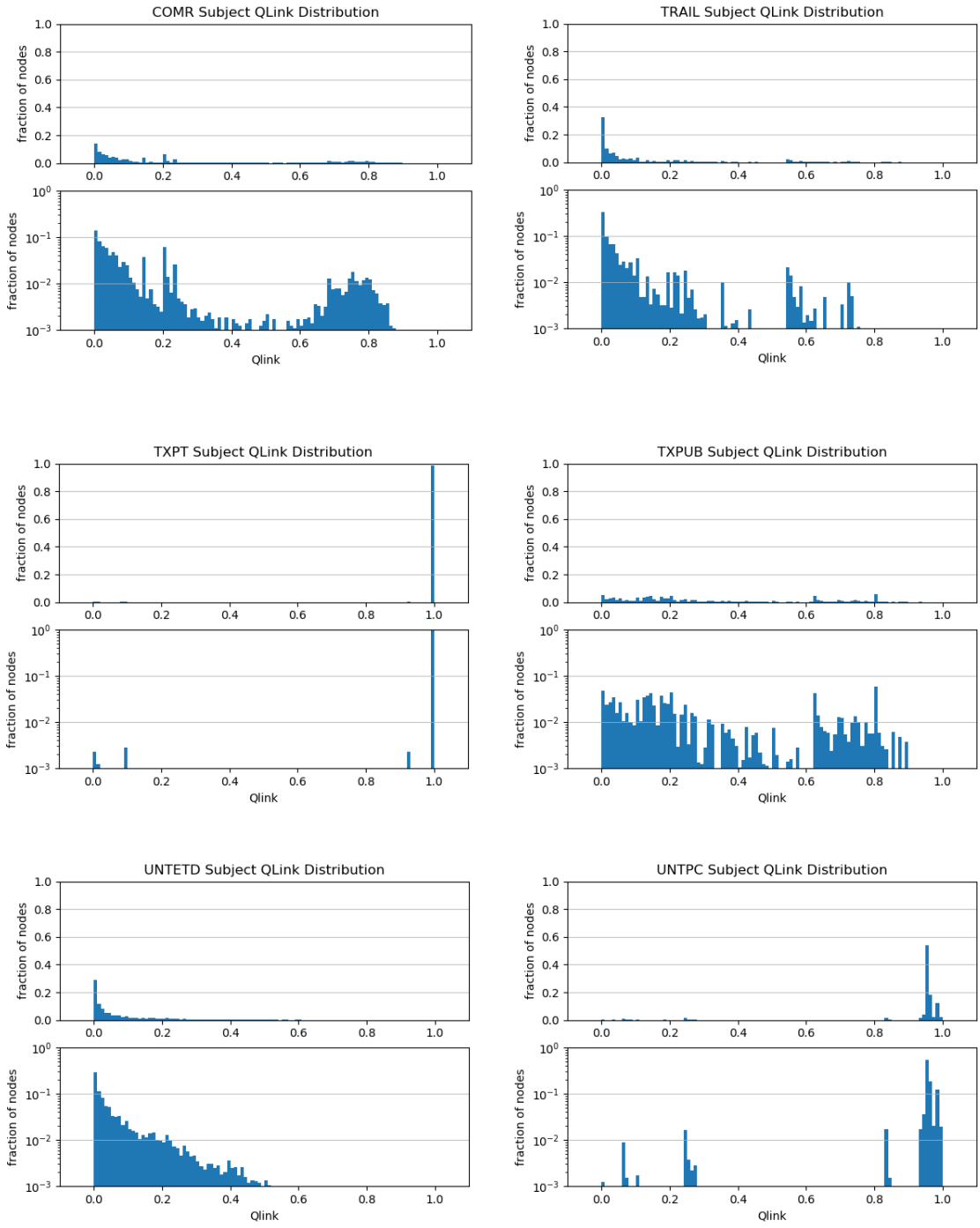

FIG. 2. Subject Qlink Distribution for each collection.

5. Discussion

Traditional metrics provide a general overview of the makeup of metadata records in a collection, such as the percentage of records with data values for a particular metadata element, which outlines overall usage in the collection. For example, both the TRAIL and UNTETD collections have 100% usage of the title element across all records. However, the TRAIL collection has 41,977 unique title values (a 1:1.7 ratio of unique titles per record) compared to the 19,290 unique title date values in the UNTETD collection (essentially a 1:1 ratio). The entropy calculation provides a better

https://doi.org/10.23106/dcmi.952141695

understanding of how much information is included within the metadata element in a collection. Entropy values closer to 1 indicate more unique information whereas entropy values closer to 0 indicate less information in that metadata element. To illustrate this phenomenon, one may imagine adding a new data value to a particular metadata element in the collection. If the entropy — calculated before this addition — is high for that metadata element in a collection, there is a higher probability that the new data value will be unique. In the situation when the entropy is lower, one can assume that the new data value will have already occurred in the collection.

The network statistics obtained in this study demonstrated that there is an inverse relationship between the density of the Metadata Record Graph and the entropy of that same metadata element in a collection. This trend is most noticeable for elements such as title and creator in the UNTETD collection where there is a high entropy — 1 and 0.998 respectively — and a low network density — 0 in both cases. To the contrary, metadata elements such as resourceType, format, and language in both collections exhibit low entropy and very high network density. This observation is explainable because if there are only a few values in a collection that are used very often, such as language values in a primarily English-language collection, one would expect that the resulting Metadata Record Graph for that element would be highly connected. A user would be able to view one metadata record containing a language code of English and find all of the other metadata records in the collection that also have the same language code.

Perhaps the easiest-to-understand network metric for a metadata element is the number of unconnected nodes in a Metadata Record Graph. For example, the TRAIL collection has 4,290 unconnected nodes in the creator element network, which means 17% of the metadata records are isolated from the rest of the records in the collection. In the UNTETD collection there are 17,715 isolated metadata records (92%) based on unconnected nodes in the creator element. Such a drastic difference is to be expected due to the nature of these collections. In a collection of theses and dissertations defended at UNT, the same author name would only appear in more than one metadata record if a student received multiple degrees from the university. On the other hand, in a large collection of technical reports (such as TRAIL) the overlap between metadata records based on author name(s) occurs more frequently as creators often author multiple reports, and it would be common to navigate from one report to other reports authored or co-authored by the same person.

The average degree of a metadata element's Metadata Record Graph provides a similar insight into the network of metadata records compared to network density. The average degree will be high if the density of the network is also high. This helps develop a sense of the number of connections per node that occur, i.e., the average number of other metadata records a user could navigate to directly from that metadata record by clicking on any data value in that given metadata element.

Based on our evaluation, the Qlink metric seems to be the most useful for metadata evaluation as plots that provide a sense of the shape of the distribution of Qlink across the network. Our analysis has revealed that the distributions for most of the networks based on the subject metadata element matches other network distributions — like the Web and many social networks — where there are a large number of nodes with few connections or a low degree, and a smaller number of nodes with a high degree, or many connections (e.g., COMR, TRAIL, TXPUB, and UNTETD collections). The distributions for the TXPT and UNTPC collections are quite different visually, with a spike near or at the value of 1.0. Based on our knowledge of the collections, we believe this is due to the effects of common subject data values. For example, the TXPT collection consists solely of patents from Texas and each metadata record in this collection would share a broad subject heading (e.g., "Patents -- Texas."), so the records would all be connected based on those general topics, even if there is much less overlap on more item-specific subject terms.

Finally, a comparison of the Metadata Record Graphs for the subject metadata element reveals substantial differences in the network density: the TXPT and UNTPC collections exhibit very high density while the COMR, TRAIL, and UNTETD collections have a very low density and the

https://doi.org/10.23106/dcmi.952141695

TXPUB collection falls somewhere in the middle. Other metrics such as unconnected nodes, average degree and the Qlink mean seem to follow the same trend.

Overall, our analysis leads to the conclusion that Metadata Record Graphs are a useful tool for assessing metadata, as its network metrics represent the connectivity of the metadata records, which is not apparent with other common forms of analysis. Some of these metrics are easily interpreted, such as density and number of unconnected nodes in the network. Other statistics — such as average degree and Qlink with its associated average — are not as straightforward and require more familiarity with network analysis to enable adequate interpretation. Plotting Qlink distributions helps to substantially simplify evaluation of the Qlink metric for values across the Metadata Record Graph.

Results from this investigation indicate that the Metadata Record Graph based on the subject metadata element is likely the most useful to consider as an indicator of connectedness for metadata records, since other element data values tend to fall within specific parameters (i.e., mostly the same or mostly different) for a single collection. Subject is also the primary metadata element that metadata creators can modify to adjust the network properties, compared to other information associated with an item that is less subjective (e.g., publishers or dates). For example, if a collection is overly connected with highly generic subject terms then more specific terms can be used; conversely, if a network has a low density and is not very connected, metadata creators have the ability to add more generic subject terms. Both of these types of adjustments to network properties are expected to aid in bringing users to a larger number of relevant resources.

6. Future Work

Future research into Metadata Record Graphs is needed. One potentially fruitful direction would be to compare network metrics over a larger set of collections than just the six analyzed in this study to see if other useful information or patterns emerge. Another area of work would be to comparatively evaluate usability of Metadata Record Graphs for subject data values based on different kinds of standardized subject terms such as Library of Congress Subject Headings with other controlled vocabularies (e.g., Art and Architecture Thesaurus, ERIC terms, etc.) and more free-form subject terms (i.e., keywords). Finally, as this study assumed exact matching between data values in networks of metadata records to connect them, it would be useful to investigate how different normalizations of strings can be used to further connect metadata records in a network. For example, basic normalizations — making all values lowercase, stripping punctuation and extra spacing, converting non-ASCII characters, or similar combinations of value manipulations — could simulate the potential level of interconnectedness if metadata managers were to standardize formatting, assert name authority, or do other clean-up of data values. In text-heavy elements, such as title or description, networks created by other methods, such as clustering or term vector models, could simulate the likelihood that users would find like items with general keyword searches (rather than full string searches).

Metadata Record Graphs provide an opportunity for new metrics to help metadata creators and managers assess the metadata in their digital collections and identify areas where changing or normalizing data values would increase network density and, as a result, improve users' ability to find related materials. The metrics from the field of network analysis offer different insights into the collections of metadata records that are not easily achieved with more traditional count-based metadata statistics. While providing a new opportunity to evaluate metadata, Metadata Record Graphs may require additional study and documentation to develop a solid understanding for interpretation, given the wide range of available metrics that can be calculated for networks as well as for their individual nodes.

https://doi.org/10.23106/dcmi.952141695

7. References

Babu, Ramesh and Ann O'Brien. (2000). Web OPAC interfaces: an overview. The Electronic Library, 18(5) 316–330.

Bruce, Thomas R. and Diane Hillmann. (2004). The continuum of metadata quality: defining, expressing, exploiting. In D. Hillman & E. L. Westbrooks (Eds.), Chicago: ALA Editions.

Harper, Corey. (2016). Metadata analytics, visualization, and optimization: Experiments in statistical analysis of the Digital Public Library of America (DPLA). The Code4Lib Journal, 33.

Király, Péter, Juliane Stiller, Valentine Charles, Werner Bailer, and Nuno Freire. (2018). Evaluating data quality in Europeana: Metrics for multilinguality. In Metadata and Semantic Research - 12th International Conference, MTSR 2018, Limassol, Cyprus, October 23-26, 2018, Revised Selected Papers, 199–211.

Ochoa, Xavier and Erik Duval. (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

Phillips, Mark Edward, Hannah Tarver, and Oksana L. Zavalina (2019). Metadata Record Graphs for Six Collections from the UNT Libraries' Digital Collections. https://digital.library.unt.edu/ark:/67531/metadc1532397/

Stvilia, Belsiki. (2006). Measuring Information Quality. PhD thesis, University of Illinois at Urbana-Champaign.

Stvilia, Besiki, Les Gasser, Michael B. Twidale, and Linda C. Smith. (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

Stvilia, Besiki, Les Gasser, Michael B. Twidale, Sarah L. Shreeves, and Tim W. Cole. (2004). Metadata quality for federated collections. Proceedings of the Ninth International Conference on Information Quality (ICIQ-04), 111–125.

Tarver, Hannah, Mark E. Phillips, Oksana L. Zavalina, and Priya Kizhakkethil. (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, Hannah, Oksana L. Zavalina, and Mark E. Phillips. (2017). An exploratory study of the description field in the Digital Public Library of America. In International Conference on Dublin Core and Metadata Applications. 34–44.

Tani, Alice, Leonardo Candela, and Donatella Castelli. (2013). Dealing with metadata quality: The legacy of digital library efforts. Information Processing & Management, 49(6), 1194–1205.

Task Force on Guidelines for OPAC Displays (Ed.) & Standing Committee of the IFLA Section on Cataloguing. (2008). IFLA Guidelines for Online Public Access Catalogue (OPAC) Displays. Final Report May 2005. Berlin, Boston: De Gruyter Saur.

Ward, Jewel. (2003). A quantitative analysis of unqualified Dublin Core Metadata Element Set usage within data providers registered with the Open Archives Initiative. Proceedings of the Joint Conference on Digital Libraries 2003, 315–317.

Zavalina, Oksana L., Mark E. Phillips, Daniel Gelaw Alemneh, Hannah Tarver and Priya Kizhakkethil. (2015). Extended Date/Time Format (EDTF) in the Digital Public Library of America's metadata: Exploratory analysis. In Proceedings of the Association for Information Science and Technology, 52(1) 1-5.

https://doi.org/10.23106/dcmi.952141695