# Exploring the Utility of Metadata Record Graphs and Network Analysis for Metadata Quality Evaluation and Augmentation

## Mark E. Phillips

UNT Libraries,
University of North Texas,
Denton, TX, USA
E-mail: mark.phillips@unt.edu

## Oksana L. Zavalina

College of Information
University of North Texas,
Denton, TX, USA
E-mail: oksana.zavalina@unt.edu

## Hannah Tarver

UNT Libraries,
University of North Texas,
Denton, TX, USA
E-mail: hannah.tarver@unt.edu

**Abstract:** Our study explores the possible uses and effectiveness of network analysis, including Metadata Record Graphs, for evaluating collections of metadata records at scale. We present the results of an experiment applying these methods to records in the University of North Texas (UNT) Digital Library and two sub-collections of different compositions: the UNT Scholarly Works collection, which functions as an institutional repository, and a collection of architectural slide images. The data includes count- and value-based statistics with network metrics for every Dublin Core element in each set. The study finds that network analysis provides useful information that supplements other metrics, for example by identifying records that are completely unconnected to other items through the subject, creator, or other field values. Additionally, network density may help managers identify collections or records that could benefit from enhancement. We also discuss the constraints of these metrics and suggest possible future applications.

**Keywords:** Metadata Record Graphs; metadata quality; metadata linking; network analysis; metadata evaluation; digital libraries; quality metrics; network graphing; Dublin Core.

**Reference** to this paper should be made as follows: Phillips, M. E., Zavalina, O. L., and Tarver, H. (xxxx). 'Exploring the Utility of Metadata Record Graphs and Network Analysis for Metadata Quality Evaluation and Augmentation', *International Journal of Metadata, Semantics and Ontologies*, Vol. x, No. x, pp.xxx–xxx.

**Biographical notes:** Mark Edward Phillips is the Associate Dean for Digital Libraries at the University of North Texas Libraries. His research interests include metadata quality, digital library infrastructure, digital preservation, and web archiving.

Dr. Oksana L. Zavalina is an Associate Professor in the Department of Information Science in the College of Information, University of North Texas. Her research interests are related to all aspects of information organisation (and specifically metadata) in traditional libraries, digital libraries, and aggregations.

Hannah Tarver is the Head of the Digital Projects Unit in the University of North Texas Libraries. Her research interests include metadata quality, controlled vocabularies, authority control, and digital library management.

## 1 Introduction

Metadata quality in electronic catalogues and digital library systems has been addressed using a variety of metrics in small collections and large aggregations. Many organisations have quality control measures in place to review metadata records or particular field usage, though this becomes increasingly difficult when the number of records becomes too large to verify information individually. Additionally, analysis of individual records may be more useful for finding certain kinds of problems (e.g., information that does not match an item, or formatting errors) but could miss others (e.g., consistency of usage for names or subjects within a collection or among multiple collections). The same is true for any particular quality assessment or metric.

As collections have grown, researchers and metadata managers have continually looked for additional ways to evaluate metadata and identify possible issues. Information science and computer science have successfully used network analysis to facilitate information access by leveraging the relations between information objects, such as research papers (e.g., Web of Science) or web pages (e.g., Google ranking algorithm). However, in the research efforts aimed at evaluating and improving overall metadata quality to facilitate information access, the interconnectedness of metadata records has been largely overlooked.

Digital libraries often represent these connections as a hypertext link that allows a user to travel from one resource to a list of resources that contain the same subject (or other) data value in metadata records representing them, oftentimes as a search result; users then can choose to navigate to another record that also has the specified term. The research project presented in this article tests the use of network analysis as a metric for metadata quality. In this model, a 'node' is represented by a single metadata record and the connections are built by shared data values in a particular metadata field, such as the records in Figure 1 that share a subject field data value of 'Oceans'.

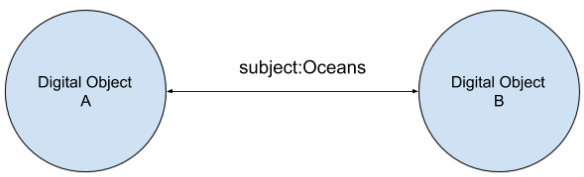


**Figure 1** Simple Metadata Record Graph connecting two resources.

We believe that network analysis and particularly the use of these graphs hold a strong potential to aid in metadata management for two reasons. First, such an approach is in line with the concept of linked data. Second, it is based on the idea that users find linking functionality helpful to discover 'more items like' one they have found. This paper describes a study aimed at determining the most efficient and effective ways to employ operationalisations of the proposed concept of the Metadata Record Graph to evaluate and augment collections of metadata records.

## 2 Literature Review

Libraries have long been concerned with the quality and consistency of item representation in their catalogues and have been documenting analyses of metadata quality (e.g., Mason, 2007; Hider and Tan, 2008; Hill, 2008). With the move to digital media, this trend has continued to digital library holdings for digitised materials and born-digital objects under the auspices of various cultural heritage institutions. Metadata in these systems often acts as the sole or primary method of supporting the functions of finding, accessing, and managing the digital objects. As collections have grown to thousands, or even millions, of digital items, it has become a challenge for managers to evaluate the quality of metadata records across collections, or even to clearly define the appropriate characteristics. To address this problem, a number of researchers have worked on clarifying metadata quality as a concept. Among the first notable activities in this area was the work carried out as part of the United States Government Information Locator Service (GILS) and National Science Digital Library (NSDL). GILS and NSDL efforts spurred discussions and experiments around metadata quality and resulted in creation and extension of frameworks for quantifying and understanding metadata quality (e.g., Moen et al., 1998; Bruce and Hillmann, 2004; Stvilia and Gasser, 2007). These metadata quality frameworks were operationalised by suggesting metrics to measure their components (Ochoa and Duval, 2009; Király, 2015; Chen et al., 2011). Many of these metrics rely on the aggregate values of metadata records in digital collections. Examples of aggregated values, also referred to as count- and value-based metrics include the number of records in an aggregation that contain a given element, the number of unique data values of a single metadata field present across all metadata records, and the average number of instances of a metadata field in records across the aggregation.

Some projects have applied metadata quality metrics, particularly those related to completeness – the third most important metadata quality criterion (Park, 2009) – into production systems. These projects include, for example, those presented in Zavalina, Alemneh, Kizhakkethil, Phillips and Tarver, 2015; Király and Büchler, 2018, etc. Aggregations that bring together metadata from different sources inevitably face problems with metadata quality, and because of this, evaluation of metadata gains more and more importance (Hillmann, 2008). Thus, several studies, in their analysis of large collections of metadata, focused on counting instances of data values in metadata and

then performing standard descriptive statistics on these counts to better understand collections in an aggregated environment. These include examination of metadata records in the Digital Collections and Content (DCC) aggregation (Jackson, Han, Groetsch, Mustafoff, and Cole, 2008), and the Digital Public Library of America (DPLA) aggregation (Tarver, Phillips, Zavalina, and Kizhakkethil, 2015; Harper, 2016).

Metadata quality research in the digital library environment overlaps substantially with studies of traditional library catalogue records, which frequently have similar information documented in similar ways. Over the decades, a number of studies of the quality of traditional library metadata expressed in the MARC (Machine Readable Cataloguing) metadata standard have been undertaken. For example, the findings of a study by Mayernik (2010) that examined the distribution of MARC21 fields in bibliographic records pointed at the relatively low level of interconnectedness between these metadata records due to the fact that most MARC21 fields were used in a small number of metadata records, and a smaller number of fields appeared in nearly all the records. The large-scale MARC Content Designation Project carried out in the second half of the 2000s examined the extent of application for the hundreds of 'data elements' (fields and subfields) available in the MARC21 metadata standard by analysing all 56 million MARC21 bibliographic records in WorldCat, the largest database of MARC21 metadata at the time. Of particular interest to the issue of interconnectedness of metadata records, the MARC Content Designation Project team looked for a set of commonly- or frequently-occurring data elements in bibliographic records in groups of MARC21 metadata records arranged based on format or type of material described by a record (Moen et al., 2006). Researchers also compared utilisation of MARC21 fields and subfields in the WorldCat metadata records with the National, Core, and Minimal level record standards (Eklund et al., 2009). Similarly, several years later, another team of researchers (Smith-Yoshimura, Argus, Dickey, Naun, Ortiz, and Taylor, 2010) examined patterns of MARC21 field and subfield usage in the WorldCat database and its implications on metadata practices; they found that only a small subset of available fields occur in WorldCat records. The research team proposed six factors for practitioners to consider when making decisions regarding creation of MARC21 metadata records: strive for consistency in choice and application of a field; respond to local user needs; focus on authorised names, classification, and controlled vocabularies as the need for surrogate 'descriptive metadata' will decrease; use appropriate fields to reflect the resource; MARC data cannot continue to exist on its own, separate from the rest of the 'information universe'; and recognise that MARC data is used for far more than just user retrieval and identification so accuracy matters (pp. 13-14).

In addition to drawing on library catalogue records evaluation practice and research, digital library metadata quality research borrows terminology and methods from other related fields that analyse data values, including network analysis. Network analysis – and its more generalised field of graph theory – has played an important role in a number of areas of information science over the past sixty years. One area of information science that heavily relies on network analysis and graph theory is that of citation analysis. Citation analysis began in the 1960s with the work of Derek J. de Solla Price (1965) and Eugene Garfield (1955) who described the concept put into production as the Scientific Citation Index (SCI). The networks of citations that connected either publications or authors were used to both discover new publications in a field and to identify, score, and rank the impact of the research literature. These citation indexes and databases continue to play a major role in the research component of scientific discovery with products like Web of Science, Scopus, and Google Scholar. The generic graph model of information has been used heavily in the past thirty years within information science and computer science to describe networked information and systems like the Internet. Perhaps the most well-known application of graph analysis applied to networks of documents is the PageRank algorithm proposed by Lawrence Page and Sergey Brin (Page et al., 1998) which became the basis of Google search and modern searching on the web.

An area of practice and research related to graph theory and network analysis is linked data and its use in the library and information context. Linked data as a term was coined by Tim Berners-Lee (2006) in a design note about the Semantic Web project that he had envisioned as the logical extension of the early web. In recent 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 think about their metadata collections as relationships in a network and not only as static descriptions of their local resources.

The idea of importance of establishing connections between resources is not new. In the late 19th to early 20th century, Charles Ammi Cutter (1904) defined the goals of a library catalogue, including the ability for users to see what holdings the library has for a given author or subject, or in a given type or genre of literature. In addition to the titles of information objects, names of agents that created or contributed to the creation of information objects, and 'aboutness' of information objects (their topical, geographical, temporal, or other subjects) have traditionally been the focus of organising access to information. In brick-and-mortar libraries of the not so distant past – up until the 1980s – two catalogues often existed side-by-side. These included the alphabetic catalogue in which metadata records (in the form of catalogue cards) were arranged in alphabetical order by title and by creator names (multiple records were needed in this catalogue to represent the co-authored works and works with multiple titles), and the subject catalogue in

which records were organised systematically, by field of knowledge, according to classification system (again, multiple records were needed for multi-subject works – one for each of its major subjects). It was thus common for the same information object to have three or more metadata records representing it in the library card catalogue to allow for multiple access points.

The major improvement to this functionality today is that online catalogues and databases provide new 'affordances' (e.g., Bates, 1989) not available in card catalogues: they are no longer limited by the size of the catalogue's card to fit metadata, by representing only physical objects or a specific physical location that a card catalogue was tied to. It is possible now to provide listings of the items in a collection organised by data values in any field of a metadata record, and in fact many digital library systems chose to do so beyond the creator, subject, and title fields' data values. Moreover, it is now possible to link metadata records based not only on the full data value (e.g., a subject string such as 'Information Science – Study and Teaching') but also on its components if necessary (e.g., linking metadata records that have only 'Study and Teaching' in the subject strings in their subject data values).

The web allows users to make connections between web pages or resources using links. These links form the foundation for navigation between information objects that users access every day and are a common feature in databases of cultural heritage institutions as well. Library catalogues and digital repositories constructed over the past three decades have enabled users to take advantage of the connections that exist between records in the form of shared data values (that are hyperlinked) to help discover other similar resources. This use of a link to connect records that share a data value is the basis for much of the work carried out in metadata and cataloguing practice, to not only describe resource but also to make connections between similar resources. In a physical library we can often co-locate resources on the shelf allowing for additional discoveries to be made; in the digital environment this is more challenging and makes the connections between our metadata records more important.

The principle of metadata's linking function and its relation to metadata quality has been incorporated in a Theory of Metadata Enrichment and Filtering that was recently proposed by Alemu (2016, 2017). The four principles of this theory – enrichment, linking, openness, and filtering – provide a way of understanding metadata creation, enhancement, and use in a modern information landscape. This theory provides a framework to understand how the concept of linking within a collection of metadata records can improve the value of the metadata as a whole.

## 3 Study Purpose and Research Questions

The study presented in this paper intends to address questions about how shared data values in metadata records – that may serve as links to similar resources for users – may be evaluated as an aspect of metadata quality across large collections of records. In addition to making practical contribution to the field of metadata and metadata quality for cultural heritage collections, this research seeks to further investigate, measure, and provide support to Alemu's Theory of Metadata Enrichment, particularly the principle of linking. This study will help provide a foundation for understanding the extent to which metadata collections have implicitly linked their metadata through shared metadata values, as well as establishing metrics that can be used to measure this concept by creating Metadata Record Graphs and measuring their characteristics.

Our study seeks to answer the following research questions:

- To what extent can network analysis be used as a method to quantify metadata quality in digital library collections?
- What are the meaningful metrics from network theory that can be used as metadata quality indicators for networks of metadata records?

These broad questions guide this study and provide a framework for us to understand more specific questions related to graphs of metadata records including: What is the average density of networks of metadata records? How does the density of networks of metadata change as we add more records to the collection?

## 4 Methods

To address these research questions, we utilise a set of experiment-based case studies that analyse publicly-facing descriptive metadata collections hosted by the University of North Texas (UNT) Libraries. The UNT Digital Library (UNTDL) is a repository for resources created by faculty, staff, and students at the university as part of their research output, and for resources collected by the Libraries. These materials comprise both 'digitised analogue' material and born-digital resources from a wide range of academic disciplines, subject areas, and content creators. Each resource is described by a metadata record in a local format (UNTL) which contains 21 descriptive metadata elements: 15 Dublin Core elements (title, creator, contributor, publisher, date, language, description, subject, coverage, source, rights, relation, resource type, format, and identifier) as well as 6 additional elements (collection, partner, degree, citation, primary source, and note) defined locally. Descriptive metadata is created by a wide range of editors including student assistants, full- and part-time staff, and librarians at the Libraries.

Metadata in the UNT Digital Library is available to the public using the Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH). OAI-PMH allows for the programmatic harvesting of metadata from repositories and has been a core piece of the scholarly repository landscape for over fifteen years since its first release in 2002. This study uses records from the Digital Library harvested from the system in the original, UNTL format in May 2019.

In this study, we analyse the entire collection of metadata records in the UNT Digital Library as a whole. Additionally, we conduct more specific analysis for two collections: the UNT Scholarly Works (UNTSW) collection and the Professor Ray Gough Slide Collection (PRGSC). The UNT Scholarly Works collection serves as the institutional repository for the university and contains scholarly research and creative works deposited by members of its faculty and staff, as well as students. The Gough collection comprises solely slide images (positive photographs), documenting architecture observed by one of the university's professors during his travels around the world, including historic and modern buildings as well as archaeological sites. In both cases, metadata has been created by departmental staff or trained student assistants.

### *4.1 Data Collection and Processing*

For this research, a 'Metadata Record Graph' for each element in the established sets was created by processing metadata records downloaded via OAI-PMH. In the network graph, the nodes represent metadata records and the edges represent connections between those records, such as a common subject field data value, or a common contributor field data value. For each of the data sources (collections), the following general steps were used to create network graphs:

1. Metadata records are harvested from the UNT Digital Library system.
2. Unique identifiers for each metadata record are paired with the data values for a specific metadata field (such as subject), output, and sorted so that data values are arranged together alphabetically.
3. Identifiers for a shared data value are grouped with that value. These represent nodes that are connected by a common data value.
4. Connections are created between each identifier in a group, output, and sorted.
5. A final adjacency list is created with an identifier for a metadata record as the key with metadata records that are connected to that identifier.

The Metadata Record Graphs generated from collections of metadata illustrate how users could navigate from record to record in a digital library system in the same way that the comparative prominence of web pages is calculated based on algorithms such as the PageRank introduced by Page and Brin (1998). The resulting Metadata Record Graph is an undirected graph, or a graph where the edges are bidirectional. This means that it is possible to move from record to record though an edge (a shared data value) in either direction.

### *4.2 Data Analysis*

This study employs count- and value-based statistics along with network analysis theory and constructs (e.g., network graphs). Traditional statistics for count and value-based analysis include the number of records that contain an instance of the metadata field, the percentage of records that contain an instance of the field, the number of unique data values for a field, the mean and mode number of instances of a field per record, the frequency of the mode instances per record, and finally an entropy calculation for the data values of a field. We make use of some of network theory's commonly-used measures to evaluate the interconnectedness of metadata records in a digital repository. Once Metadata Record Graphs were created, we evaluated them by calculating several different network statistics: density, degree, average degree, and the Gini coefficient.

#### *Entropy*

In relation to traditional count- and value-based statistics, we are including a measure of normalised entropy (Stvilia, Gasser, Twidale, Shreeves and Cole, 2004) for each set of records. Entropy is represented as a number between 0 and 1 and expresses the amount of unique information in the data values. Values closest to 0 (high entropy) denote elements with more unique information – such as an identifier element that might be different for every record – while an entropy value closer to 1 (low entropy) would be expected for elements that have a few unique values used in a high percentage of records – such as a language or resource type element that might be taken from a relatively short strict controlled list.

#### *Density*

The first network metric is the density of the graph: a calculation of the actual connections (or edges) in a graph divided by the potential connections (possible edges). The density of a Metadata Record Graph provides a metric that can be used to understand how tightly connected the records in the collection are, or how strongly connected the nodes in the graph are, based on a value between 0 and 1. For example, a network with a density of 0 would represent a situation in which metadata records do not share any data values and therefore do not link together, i.e., clicking on a data value in one record would retrieve only that same record in a search. Networks with a density of 1

would characterise a situation when all metadata records share a common data value, such as the language data value for a collection entirely consisting of English-language materials. Either of these cases is so extreme as to represent a useless network for the purposes of analysis but could, depending on the characteristics of the collection and the particular element, provide a metric for quality (e.g., if some of the materials in the collection were known to be in a language other than English, but the language network had a density of 1).

### *Degree*

Another set of metrics calculated from the generated graphs is the degree and the degree distribution. The degree of a node is the number of edges that intersect with a node. Once the degree for each of the nodes in the Metadata Record Graphs is known, we calculate the degree distribution of the graph itself. Stated simply the degree distribution is the probability of a given degree (i.e., number of connections) occurring in the network. In addition to the degree distribution, we generated standard descriptive statistics for the degree of nodes in the network, calculating the mean, mode, and mode frequency, as another way to understand network characteristics.

### *Gini Coefficient of Degree Distribution*

While visualisations are helpful to understand the shape of a degree distribution, it is usually challenging to directly compare these graphics in a meaningful way. Another approach is to use a metric such as the Gini coefficient as a single metric that can be used to compare different graphs. The Gini coefficient for degree distribution is a statistical measure that provides a mechanism to compare distributions using a single number; it was initially developed to gauge economic inequality, but has been suggested as an appropriate measure for degree distributions (Badham, 2013). Essentially, the Gini coefficient is used as a metric for the histogram of degree distribution and is represented as a single integer from 0 to 1.

A Gini coefficient near 0 represents a very uneven distribution – i.e., there is a wide range of frequency distributions within the network. On the other side of the spectrum, a coefficient near 1 would represent an extremely even distribution – i.e., roughly the same percentage of nodes have each number of connections. However, a value of 0.0 will also occur when every node (metadata record) contains the same data value for a particular metadata field, or when a field does not contain any data values in a record.

## 5 Findings

The first step was to examine the entire public holdings of the UNT Digital Library (UNTDL) – containing 698,422 individual item-level metadata records – to gain a better sense of the context and the interconnectedness of data values across the system. For each of the 15 Dublin-Core-based metadata fields, we started with traditional count-based metrics to discern usage of those fields in the UNTDL metadata records before calculating networking metrics (see Table 1). Six Dublin-Core-based metadata fields are required for every UNTDL metadata record – title, language, description, subject, resource type, and format – and all of these fields have existing data values in at least 96% of records. Non-required metadata fields exhibit a much wider range of usage from 2% (relation) to 99% (date and identifier).

Next we examined network analysis data for the same 15 metadata elements across the UNTDL (see Table 2). This includes the number of connected nodes (metadata records connected to at least one other record by a shared data value), unconnected nodes (records with completely unique or no data values for that metadata field), total edges (number of connections), density, average degree (average number of records connected by a particular data value), degree mode (most common number of links to any particular node), frequency of mode degree (percentage of records containing the most common degree), and the degree distribution Gini coefficient.

### *5.1 Selected Collections*

We also analysed subsets of metadata records from two individual collections within UNTDL to examine how network analysis findings might differ among collections that have some level of similarity or theme. The first collection is the Scholarly Works collection (with 5,839 publicly-available items), which serves as the institutional repository. As such, it comprises a wide range of item types and academic disciplines, but we would expect to see overlap in metadata fields such as creator, since faculty members often submit multiple items to the institutional repository. General statistics for this collection are presented in Table 3.

Additionally, we have calculated the same network statistics for the Scholarly Works collection as those calculated for the entire Digital Library (see Table 4).

The second subset of metadata records we examined is for the Professor Ray Gough Slide Collection, which contains a smaller number of items (only 1,008 publicly-available metadata records) representing a single academic field. These are photographs taken by a professor for use in his classes to illustrate architecture. In this case, some information is the same in every metadata record. As shown in count-based statistics (see Table 5), this includes the data values in the following fields: resource type, format, and creator. However, data values representing specific topics, locations, dates, etc. may be different.

In Table 6, we present the network statistics for this smaller collection. The same metadata fields that have a high level of uniformity e.g., creation dates that often

| Field Name | Records with Field Instances | % of Records with Field Instances | Unique Data Values in Field | Mean Field Instances Per Record | Mode Field Instances Per Record | Frequency of Mode Instances Per Record | Entropy |
|---|---|---|---|---|---|---|---|
| title | 698,422 | 100% | 618,703 | 1 | 1 | 61% | 0.832 |
| creator | 601,211 | 86% | 365,444 | 2 | 1 | 64% | 0.850 |
| contributor | 318,151 | 46% | 41,226 | 1 | 1 | 84% | 0.561 |
| publisher | 471,495 | 68% | 26,953 | 1 | 1 | 95% | 0.544 |
| date | 690,343 | 99% | 42,421 | 1 | 1 | 98% | 0.832 |
| language | 698,195 | 100% | 29 | 1 | 1 | 100% | 0.151 |
| description | 670,361 | 96% | 510,814 | 2 | 2 | 85% | 0.754 |
| subject | 678,649 | 97% | 469,261 | 11 | 9 | 9% | 0.675 |
| coverage | 303,482 | 43% | 23,426 | 2 | 3 | 41% | 0.438 |
| source | 249,868 | 36% | 78,068 | 1 | 1 | 100% | 0.842 |
| relation | 12,115 | 2% | 12,415 | 1 | 1 | 85% | 0.974 |
| rights | 292,073 | 42% | 18,365 | 3 | 3 | 64% | 0.271 |
| resourceType | 698,422 | 100% | 36 | 1 | 1 | 100% | 0.546 |
| format | 698,422 | 100% | 6 | 1 | 1 | 100% | 0.413 |
| identifier | 689,503 | 99% | 1,542,518 | 3 | 1 | 36% | 0.917 |

**Table 1** Count-Based and Data-Value-Based Statistics for UNTDL Metadata (n=698,422)

| Field Name | Connected Nodes | Unconnected Nodes | Total Edges | Density | Average Degree | Degree Mode | Frequency of Mode Degree | Degree Distribution Gini Coefficient |
|---|---|---|---|---|---|---|---|---|
| title | 339,128 | 359,294 | 2,838,151,740 | 0.0116 | 8,127 | 0 | 51% | 0.843 |
| creator | 539,660 | 158,762 | 3,411,994,070 | 0.0140 | 9,771 | 0 | 23% | 0.835 |
| contributor | 314,938 | 383,484 | 11,312,257,586 | 0.0464 | 32,394 | 0 | 55% | 0.78 |
| publisher | 460,787 | 237,635 | 2,990,677,377 | 0.0123 | 8,564 | 0 | 34% | 0.708 |
| date | 679,376 | 19,046 | 176,472,909 | 0.0007 | 505 | 0 | 3% | 0.738 |
| language | 698,194 | 228 | 177,288,288,819 | 0.7269 | 507,682 | 585,613 | 84% | 0.129 |
| description | 568,605 | 129,817 | 3,457,039,635 | 0.0142 | 9,900 | 0 | 19% | 0.803 |
| subject | 674,299 | 24,123 | 14,943,268,364 | 0.0613 | 42,792 | 0 | 3% | 0.49 |
| coverage | 302,432 | 395,990 | 23,949,368,821 | 0.0982 | 68,581 | 0 | 57% | 0.679 |
| source | 191,898 | 506,524 | 26,124,883 | 0.0001 | 75 | 0 | 73% | 0.927 |
| relation | 4,298 | 694,124 | 16,621 | 0.0000 | 0 | 0 | 99% | 0.997 |
| rights | 292,038 | 406,384 | 22,170,350,459 | 0.0909 | 63,487 | 0 | 58% | 0.684 |
| resourceType | 698,421 | 1 | 51,741,699,001 | 0.2121 | 148,167 | 258,884 | 37% | 0.347 |
| format | 698,422 | 0 | 146,198,125,295 | 0.5994 | 418,653 | 523,914 | 75% | 0.191 |
| identifier | 358,572 | 339,850 | 1,337,151,920 | 0.0055 | 3,829 | 0 | 49% | 0.805 |

**Table 2** Network Statistics for UNTDL Metadata (n=698,422)

apply to many items also have a high average degree and extremely low Gini coefficient.

## 6 Discussion

The Metadata Record Graphs are complementary to the count- and value-based metrics (e.g., Tables 1, 3, and 5). Each set of metrics provides different information; when these are used together it is possible to develop a better understanding of metadata characteristics in a collection or across a whole digital library.

A number of aspects of the Metadata Record Graphs may interest metadata professionals. Perhaps the simplest measure is the number of connected and unconnected nodes in the graph – i.e., the number of metadata records that have a connection to at least one other metadata record in the collection (for a particular field) vs. the number of records that have no connections to the rest of the collection based on those fields' data value(s). Unconnected nodes in a Metadata Record Graph represent metadata records that do not contain a data value for that particular metadata field, or a data value that does not connect to another record, i.e., it is unique within the collection. This could be due to variations in spelling or formatting – e.g., for a name in the creator field – or the use of an extremely specific term (usually a subject term) that is not used in any other records. In the data from this paper, there are a range of numbers related to unconnected nodes. The most straightforward data is from the Gough collection, where there is a direct correlation between usage of metadata field and connected nodes (e.g., contributor, publisher, source, relation, and identifier have 0% usage and 100% unconnected nodes). In comparison, the Scholarly Works

| Field Name | Records with Field Instances | % of Records with Field Instances | Unique Data Values in Field | Mean Field Instances Per Record | Mode Field Instances Per Record | Frequency of Mode Instances Per Record | Entropy |
|---|---|---|---|---|---|---|---|
| title | 5,839 | 100% | 5,780 | 1 | 1 | 83% | 0.944 |
| creator | 5,823 | 100% | 7,462 | 3 | 1 | 31% | 0.897 |
| contributor | 2,080 | 36% | 883 | 1 | 1 | 85% | 0.781 |
| publisher | 3,348 | 57% | 591 | 1 | 1 | 100% | 0.757 |
| date | 5,711 | 98% | 2,987 | 1 | 1 | 86% | 0.911 |
| language | 5,839 | 100% | 10 | 1 | 1 | 100% | 0.147 |
| description | 5,839 | 100% | 6,580 | 2 | 2 | 99% | 0.833 |
| subject | 5,839 | 100% | 11,228 | 4 | 3 | 42% | 0.919 |
| coverage | 453 | 8% | 341 | 2 | 1 | 59% | 0.864 |
| source | 4,873 | 83% | 3,432 | 1 | 1 | 100% | 0.948 |
| relation | 699 | 12% | 694 | 1 | 1 | 88% | 0.989 |
| rights | 5,602 | 96% | 282 | 1 | 1 | 69% | 0.299 |
| resourceType | 5,839 | 100% | 24 | 1 | 1 | 100% | 0.581 |
| format | 5,839 | 100% | 6 | 1 | 1 | 100% | 0.352 |
| identifier | 2,885 | 49% | 2,847 | 1 | 1 | 96% | 0.988 |

**Table 3** Count-Based and Data-Value-Based Statistics for the UNTSW Collection Metadata(n=5,839)

| Field Name | Connected Nodes | Unconnected Nodes | Total Edges | Density | Average Degree | Degree Mode | Frequency of Mode Degree | Degree Distribution Gini Coefficient |
|---|---|---|---|---|---|---|---|---|
| title | 1,227 | 4,612 | 96,320 | 0.0057 | 33 | 0 | 79% | 0.917 |
| creator | 5,585 | 254 | 229,477 | 0.0135 | 79 | 371 | 6% | 0.625 |
| contributor | 1,867 | 3,972 | 121,548 | 0.0071 | 42 | 0 | 68% | 0.88 |
| publisher | 3,008 | 2,831 | 146,164 | 0.0086 | 50 | 0 | 48% | 0.793 |
| date | 4,471 | 1,368 | 44,238 | 0.0026 | 15 | 0 | 23% | 0.752 |
| language | 5,839 | 0 | 14,561,709 | 0.8544 | 4,988 | 5,383 | 92% | 0.073 |
| description | 5,261 | 578 | 410,930 | 0.0241 | 141 | 0 | 10% | 0.441 |
| subject | 5,329 | 510 | 181,596 | 0.0107 | 62 | 0 | 9% | 0.696 |
| coverage | 402 | 5,437 | 6,394 | 0.0004 | 2 | 0 | 93% | 0.971 |
| source | 1,910 | 3,929 | 15,513 | 0.0009 | 5 | 0 | 67% | 0.897 |
| relation | 147 | 5,692 | 154 | 0 | 0 | 0 | 97% | 0.983 |
| rights | 5,574 | 265 | 14,837,427 | 0.8705 | 5,082 | 5,444 | 67% | 0.067 |
| resourceType | 5,834 | 5 | 4,461,700 | 0.2618 | 1,528 | 2,712 | 46% | 0.392 |
| format | 5,837 | 2 | 10,178,507 | 0.5972 | 3,486 | 4,235 | 73% | 0.157 |
| identifier | 206 | 5,633 | 802 | 0 | 0 | 0 | 96% | 0.983 |

**Table 4** Network Statistics for the UNTSW Collection Metadata (n=5,839)

collection has data values in every metadata field, but also has at least 2 unconnected nodes in every metadata field except language (with 0); this is expected given the greater diversity in materials and topics, but is harder to analyse in regard to which metadata fields or data values might require further scrutiny to correct incorrect values or verify information.

A similar pattern on a much larger scale is clear in the UNT Digital Library statistics, where most metadata fields have at least 50% usage and also have hundreds of thousands of unconnected nodes. Owing to the range of content subjects and metadata sources (e.g., records manually created by staff vs. volunteers, or imported from other databases), this metric may generally be more useful at a collection level, but also seems to be better served for verifying that collections appear to have appropriate amounts of connectedness rather than identifying potential problems. Alternatively, the language, resource type, and format fields all have 100% usage, and also happen to come from strictly-controlled lists, which should result in significant overlap with 0 unconnected nodes (like the format field); however, language has 228 unconnected nodes and resource type has 1. Upon further investigation, the unconnected resource type node was an incorrect value (not from the vocabulary) and has been fixed, but the unconnected language nodes should also be verified.

Degree mode and the frequency of degree mode are useful to understand the most common degrees (i.e., the number of other records that would be retrieved in searching based on a shared data value) in each Metadata Record Graph and what percentage of the nodes have that degree. This is an aspect of network analysis that significantly complements more

| Field Name | Records with Field Instances | % of Records with Field Instances | Unique Data Values in Field | Mean Field Instances Per Record | Mode Field Instances Per Record | Frequency of Mode Instances Per Record | Entropy |
|---|---|---|---|---|---|---|---|
| title | 1,008 | 100% | 815 | 1 | 1 | 100% | 0.986 |
| creator | 1,007 | 100% | 1 | 1 | 1 | 100% | 0 |
| contributor | 0 | 0% | 0 | 0 | 0 | 100% | 0 |
| publisher | 0 | 0% | 0 | 0 | 0 | 100% | 0 |
| date | 1,008 | 100% | 48 | 2 | 1 | 55% | 0.866 |
| language | 1,008 | 100% | 9 | 1 | 1 | 99% | 0.233 |
| description | 1,008 | 100% | 1,009 | 2 | 2 | 100% | 0.601 |
| subject | 1,008 | 100% | 1,171 | 4 | 4 | 29% | 0.896 |
| coverage | 1,008 | 100% | 205 | 1 | 1 | 57% | 0.863 |
| source | 0 | 0% | 0 | 0 | 0 | 100% | 0 |
| relation | 0 | 0% | 0 | 0 | 0 | 100% | 0 |
| rights | 74 | 7% | 1 | 1 | 1 | 100% | 0 |
| resourceType | 1,008 | 100% | 1 | 1 | 1 | 100% | 0 |
| format | 1,008 | 100% | 1 | 1 | 1 | 100% | 0 |
| identifier | 0 | 0% | 0 | 0 | 0 | 100% | 0 |

**Table 5** Count-Based and Data-Value-Based Statistics for PRGSC Collection Metadata (n=1,008)

| Field Name | Connected Nodes | Unconnected Nodes | Total Edges | Density | Average Degree | Degree Mode | Frequency of Mode Degree | Degree Distribution Gini Coefficient |
|---|---|---|---|---|---|---|---|---|
| title | 326 | 682 | 294 | 0.0006 | 1 | 0 | 68% | 0.781 |
| creator | 1,007 | 1 | 506,521 | 0.998 | 1,005 | 1006 | 100% | 0.001 |
| contributor | 0 | 1,008 | 0 | 0 | 0 | 0 | 100% | 0 |
| publisher | 0 | 1,008 | 0 | 0 | 0 | 0 | 100% | 0 |
| date | 1,006 | 2 | 37,793 | 0.0745 | 75 | 54 | 13% | 0.285 |
| language | 1,008 | 0 | 411,638 | 0.8111 | 817 | 905 | 90% | 0.097 |
| description | 1,006 | 2 | 505,515 | 0.996 | 1,003 | 1005 | 100% | 0.002 |
| subject | 1,005 | 3 | 30,041 | 0.0592 | 60 | 28 | 2% | 0.406 |
| coverage | 983 | 25 | 14,948 | 0.0295 | 30 | 30 | 8% | 0.462 |
| source | 0 | 1,008 | 0 | 0 | 0 | 0 | 100% | 0 |
| relation | 0 | 1,008 | 0 | 0 | 0 | 0 | 100% | 0 |
| rights | 74 | 934 | 2,701 | 0.0053 | 5 | 0 | 93% | 0.927 |
| resourceType | 1,008 | 0 | 507,528 | 1 | 1,007 | 1,007 | 100% | 0 |
| format | 1,008 | 0 | 507,528 | 1 | 1,007 | 1,007 | 100% | 0 |
| identifier | 0 | 1,008 | 0 | 0 | 0 | 0 | 100% | 0 |

**Table 6** Network Statistics for the PRGSC Collection Metadata (n=1,008)

basic counts of unique data values or connected nodes; for example, a metadata field could have 0 unconnected nodes, but the degree mode could show that 90% of the records only connect to 1 other related resource (node), which might suggest problems depending on the expected level of uniqueness in the data values of a metadata field for that collection. An interesting aspect in the Digital Library is that, despite relatively high field usage in general, for all but three of the elements (language, resource type, and format), the degree mode was '0'.

The final measure is the Gini coefficient for the degree distribution which illustrates the homogeneity of the distribution. Within the UNT Digital Library, this distribution ranges among metadata fields from .191 (format) to .997 (relation), but the majority of metadata fields have Gini coefficients of .6 or above; numbers closer to 1 represent even distributions, or relatively flat histograms. The most uneven distributions, aside from format, are observed for the language, subject, and resource type fields. While this seems to hold true for the language metadata field in the collection analyses, there are slightly different patterns of distributions. In the Gough collection, a number of metadata fields are unused (as noted above), resulting in coefficients of 0, but which may be disregarded for this analysis. The other two lowest coefficients for that collection are observed for creator and description metadata fields, while the distribution based on the rights field is the most even in that collection. In the Scholarly Works collection, the most uneven distributions aside from the language field are observed for the rights and format metadata fields, while the fields with the most even distributions are title, coverage, and identifier.

For each of these measures, network analysis allows us to identify areas of incongruity (such as the unconnected language field nodes) or fields that seem under-connected (such as the 9% of records in the Scholarly Works collection that have a degree mode of 0) for review and further enhancement or remediation. Network analysis after making edits to identified records could also help to gauge the accuracy and effectiveness of changes for the purpose of increasing connectivity.

## 7 Further Research

We have identified some areas of improvement that might be useful for generating and analysing these Graphs. At present, analysis uses exact string matches, so differences in capitalisation, inconsistent use of diacritics or accents, and differences in whitespace or punctuation would all cause records not to be joined. Other methods of analysing metadata records would allow for data values that are similar, but not exact matches, to be connected. One benefit of Metadata Record Graphs is that it is possible to measure the effects of various normalisations on a group of data values to see how the graphs change after each modification, e.g., the projected effects of enforcing name authority on creator, contributor, or publisher fields' data values that might create additional connectedness. Traditional count- and value-based metrics provide information about changes to the number of unique data values after normalisation, but the Graphs show how these normalisations affect the connections between the records. It might be helpful to use entropy measurements to determine which metadata fields could benefit most from normalisation of data values. For example, in the Gough collection, data values in the creator field have extremely low entropy (since there is only a single photographer for each item in the collection), but across the UNT Digital Library, the entropy for that metadata field was observed to be much higher (.85), which means there is likely a need for normalisation in other subsets. This kind of analysis also has potential in allocating resources by demonstrating possible effects on specific elements by normalising data values or determining how many data values would likely need to be changed.

Other analysis could make use of more complicated calculations to gain a sense of how connected similar resources might be if a user does a keyword search on the collection, rather than clicking on hyperlinked string data values for exact matches. This is especially important for text-heavy metadata fields (e.g., title or description), where it might make more sense to use different metrics such as cosine similarity to connect related records, compared to fields such as subject, creator, contributor, or publisher that often rely on standard string normalisation.

Another consideration is that there are limitations to the possible connectedness of most fields in a particular collection. For example, normalising names in the creator field or correcting misspellings might increase the connectedness of the Metadata Record Graph based on this field, but it is impossible to add more creators if all of the information is complete, or to force unconnected nodes to become connected to other records if the creator names do not legitimately match. The one general exception in the Dublin Core element set is the subject element, since there is almost always an opportunity to add data value(s) to an unconnected node, such as a broader subject term that is used in other records within the Graph. Likewise, metadata records that seem too connected with overly broad terms (e.g., collection-level data values applied to every record in a set, such as 'Patents' in the collection of historical patents), could be edited to replace the broad term in a group of records with narrower or more specific terms (e.g., 'Patents – 19th century', 'Furniture – Patents', etc.) to change the underlying Graph and affect the way that end users interface with the records in the system.

## 8 Conclusion

The results of this study provide more information about how network analysis and Metadata Record Graphs might be used as a supplement to traditional count- and value-based statistics for high-level metadata quality assessments. From this early work it seems that both types of analysis are helpful in providing insight for large collections of metadata records. The metrics we identified from the field of network analysis are only a few of the available metrics in that field but they provide a promising set of first steps in identifying common and easily-understandable metrics for analysing aggregations of metadata records.

From this study, we clearly determined that network analysis has value as a measure of quality, though it works best in combination with count-based statistics and has limited usefulness depending on the parameters. For example, across the entire UNT Digital Library, we would expect to see connectedness of subjects – and many other fields – due to the size and variety of materials; however, the UNT Scholarly Works collection represents similar diversity of topics in a smaller subset of materials, making it difficult to determine if records are legitimately less connected or need additional subject analysis. In terms of meaningful metrics, unconnected nodes provide an obvious place to start reviewing records for correctness, while density measures can help to identify records that may benefit from enhancement, to provide more information to users, even if records are already 'correct' and potentially 'complete'.

Although this study focused on metadata records in a single digital library at a specific institution, the workflow and analysis used in this study would be applicable across a variety of metadata collections that use different metadata formats. This outcome is achievable because our analysis does not take into account the underlying languages, formatting, input

guidelines or local practices. Instead, it makes use of data values as they exist in the metadata records without judgement. Additionally, the approach used in our study relies in data collection on the OAI-PMH, which is a common standard for digital library infrastructures.

Gaining insights into how metadata records are connected both explicitly and implicitly in digital library systems is important to understanding the overall discoverability of the resources represented by these records. Metadata Record Graphs can provide a glimpse into the way that these records are implicitly connected or linked. By leveraging common interface features such as connecting one metadata record to other records that share a given data value with hyperlinked instances of data values it is possible to use standard metrics from network analysis to better understand these relationships in our digital library systems. With more information about these metadata networks, we can add or adjust information in metadata records that will increase findability and ultimately improve our users experience with our systems and collections.

## References


Alemu, G. (2016). A theory of metadata enriching and filtering. *International Journal of Libraries and Information Studies*, *66*(4), 251–262. doi: 10.1515/libri-2016-0109

Alemu, G. (2017). A theory of metadata enriching and filtering: Challenges and opportunities to implementation. *Qualitative and Quantitative Methods in Libraries*, *5*(2), 311–334. Retrieved from `http://www.qqml-journal.net/index.php/qqml/article/view/343`

Badham, J. M. (2013). Commentary: Measuring the shape of degree distributions. *Network Science*, *1*(2), 213-225. doi: 10.1017/nws.2013.10

Bates, M. J. (1989). Rethinking subject cataloging in the online environment. *Library Resources & Technical Services*, *33*, 400–412.

Berners-Lee, T. (2006). Linked data. *Design Issues*. Retrieved from `http://www.w3.org/DesignIssues/LinkedData.html`

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. Retrieved from `https://ecommons.cornell.edu/handle/1813/7895`

Chen, Y.-N., Wen, C.-Y., Chen, H.-P., Lin, Y.-H., & Sum, H.-C. (2011). Metrics for metadata quality assurance and their implications for digital libraries. In C. Xing, F. Crestani, & A. Rauber (Eds.), *Digital libraries: For cultural heritage, knowledge dissemination, and future creation* (pp. 138–147). Berlin, Heidelberg: Springer Berlin Heidelberg.

Cutter, C. A. (1904). *Rules for a dictionary catalog* (4th ed.). Washington, DC: Government Printing Office.

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. doi: 10.1080/19386380903095073

Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, *122*(3159), 108–111. doi: 10.1126/science.122.3159.108

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`

Hider, P., & Tan, K.-C. (2008). Constructing record quality measures based on catalog use. *Cataloging & Classification Quarterly*, *46*(4), 338–361. doi: 10.1080/01639370802322515

Hill, J. S. (2008). Is it worth it? management issues related to database quality. *Cataloging & Classification Quarterly*, *46*(1), 5–26. doi: 10.1080/01639370802182885

Hillmann, D. I. (2008). Metadata quality: From evaluation to augmentation. *Cataloging & Classification Quarterly*, *46*(1), 65-80. doi: 10.1080/01639370802183008

Jackson, A. S., Han, M.-J., Groetsch, K., Mustafoff, M., & Cole, T. W. (2008). Dublin Core metadata harvested through OAI-PMH. *Journal of Library Metadata*, *8*(1), 5-21. doi: 10.1300/J517v08n01_02

Király, P. (2015). *A metadata quality assurance framework* (Tech. Rep.). http://pkiraly.github.io/metadata-quality-project-plan.pdf: Gesellschaft fr wissenschaftliche Datenverarbeitung mbH Gttingen.

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). doi: 10.1109/BigData.2018.8622487

Mason, M. K. (2007). *Copy cataloguing: Our quest for the perfect copy* (Tech. Rep.). Retrieved from `http://www.moyak.com/papers/cataloguing-library-congress.html`

Mayernik, M. (2010). The distribution of MARC fields in bibliographic records : A power law analysis. *Library Resources and Technical Services*, *54*, 40–54.

Moen, W. E., Miksa, S. D., Eklund, A., Polyakov, S., & Snyder, G. (2006). Learning from artifacts: metadata utilization analysis. In *Proceedings of the 6th acm/ieee-cs joint conference on digital*

*libraries (jcdl '06)* (p. 270-271). doi: 10.1145/1141753.1141813

Moen, W. E., Stewart, E. L., & McClure, C. R. (1998, April). Assessing metadata quality: 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* (p. 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

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The PageRank citation ranking: Bringing order to the web. In *Proceedings of the 7th international world wide web conference* (pp. 161–172). Brisbane, Australia. Retrieved from `citeseer.nj.nec.com/page98pagerank.html`

Park, J.-R. (2009). Metadata quality in digital repositories: A survey of the current state of the art. *Cataloging & Classification Quarterly*, *47*(3-4), 213-228. doi: 10.1080/01639370902737240

Price, D. J. D. S. (1965). Networks of scientific papers. *Science*, *149*(3683), 510–515. doi: 10.1126/science.149.3683.510

Smith-Yoshimura, K., Argus, C., Dickey, T., Naun, C., Ortiz, L., & Taylor, H. (2010). *Implications of MARC tag usage on library metadata practices.* (Tech. Rep.). OCLC Research in support of the RLG Partnership. Retrieved from `http://www.oclc.org/research/publications/library/2010/2010-06.pdf`

Stvilia, B., Gasser, L., Twidale, M., Shreeves, S. L., & Cole, T. W. (2004). Metadata quality for federated collections. In *International conference on information quality (ICIQ)* (pp. 111–125).

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.

Tarver, H., Phillips, M., Zavalina, O., & Kizhakkethil, P. (2015). An exploratory analysis of subject metadata in the Digital Public Library of America. In *International conference on dublin core and metadata applications* (pp. 30–40). Retrieved from `https://dcpapers.dublincore.org/pubs/article/view/3761`

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