Cadernos BAD

Using Metadata Record Graphs to understand controlled vocabulary and keyword usage for subject representation in the UNT theses and dissertations collection

Mark E. Phillips
University of North Texas Libraries
mark.phillips@unt.edu
Hannah Tarver
University of North Texas Libraries
hannah.tarver@unt.edu
Oksana L. Zavalina
University of North Texas College of Information
oksana.zavalina@unt.edu

Abstract

An important function of metadata for electronic theses and dissertations (ETDs) is supporting the discovery of related documents through linking of data values in the fields of metadata records. While benefits of the ETD format allow for full-text searching, metadata is still an important and necessary component of the global ETD infrastructure because often it is not possible to share the full documents in aggregations such as the Global ETD Search for the Networked Digital Library of Theses and Dissertations. The metadata field that has the most potential to assist users in discovery is the subject field used to represent what a resource is about. Over the years there has been much discussion of the value of author-generated keywords versus adding subject terms from controlled vocabularies by information professionals as documents are submitted to the University repository. This research seeks to explore this problem with the help of network analysis method not used for such analyses before by building and analyzing metadata record graphs for the University of North Texas theses and dissertations. This paper reports on the characteristics of keyword-based and controlled-vocabulary-based metadata record

networks and discusses insights that can be gained from this approach to metadata quality analysis.

Keywords: Metadata Record Graphs, metadata, Metadata quality, Metadata analysis

Introduction

The University of North Texas (UNT) Libraries maintains Digital Collections comprising nearly 2.7 million items. This includes a collection of UNT Theses and Dissertations (UNTETD), currently containing 19,291 publicly visible items, obtained as born-digital files (since 1999) or scanned from hard copies, going back to 1936. The UNT Digital Collections team is engaged in ongoing efforts to evaluate and adjust the quality of metadata records across the system to improve findability for users.

Many existing metadata evaluation metrics use aggregated counts of metadata field data values, but we have been assessing supplementary evaluation methods, including a process we call Metadata Record Graphs, which applies network analysis to graphs generated from data values in a specific metadata field. Not all fields offer equal opportunities for measuring interconnectedness of records based on shared data values. For example, in the UNTETD collection, certain data values are shared by every record (e.g., "English" in the language field and "Thesis or Dissertation" in the resource type field); others are nearly unique (e.g., names in creator field, which would only have shared values in records for publications of students who earned multiple UNT degrees). In either case, network analysis would not provide more information about these fields than basic statistics. For this paper, we focus on the subject metadata field in the UNTETD collection's metadata records for in-depth analysis, as the only field that can reasonably be modified to adjust network values based on any findings.

The UNTETD collection has allowed self-submission of ETD documents since 1999, so some keywords are assigned by the authors; previously students would submit information that was added to records manually with some mediation (e.g., standardization of capitalization and punctuation). Additionally, metadata creation for items in this collection happens in several stages. Metadata creators often add keywords when completing other fields, but LCSH terms are generally added as items are fully cataloged in the ILS. These factors likely affect the overall consistency and quality of subject values in the collection.

This paper seeks to answer the following questions:

- How can Metadata Record Graphs help to evaluate subject metadata in a collection of theses and dissertations?

- How do Metadata Record Graphs compare when limited to controlled vocabulary terms (LCSH) vs. uncontrolled data values (keywords), including student-submitted subject terms?
- How do normalizations of subject terms (e.g., lowercase or removal of extra whitespace) affect network characteristics?

Literature Review

In the past twenty years, most institutions have transitioned from an analog workflow to the submission of PDFs as the final output of a Master's or Doctoral program (Swain, 2010). Institutions have adapted workflows for acquiring, describing, storing, and providing access to these documents (Lubas, 2009). Repository software such as DSpace or Eprints has led to a framework for access that generally relies on descriptive metadata formats, such as Dublin Core (i.e., DC, in its simple version of DC Metadata Element Set 1.1 or extended version DC Terms) or Metadata Object Description Schema (MODS), as opposed to traditional bibliographic formats (e.g., MARC). The need for ETD metadata guidelines is supported by the following documents: (Networked Digital Library of Theses and Dissertations, 2009; Texas Digital Library ETD Metadata Working Group, 2015; UNT Libraries' Digital Projects Unit, n.d.).

One area of ongoing research is subject metadata management. There are usually two sources of subject terms: author-supplied keywords and librarian-supplied controlled-vocabulary subject headings: for example, Library of Congress Subject Headings (LCSH), Medical Subject Headings (MESH), etc. There has been much discussion of the value of author-generated keywords versus adding subject terms from controlled vocabularies by information professionals. Over the years, a number of studies of library online catalogs demonstrated that natural language/keyword search produce effective results but controlled-vocabulary search is much more effective, and that users tend to search more often by keyword than by any other type of search (e.g., Fidel, 1988, 1992; Curl, 1995; Hildreth, 1997; Muddamalle, 1998).

Bates (2002) warned against ignoring the size of databases in choosing subject controlled-vocabulary for a database of a digital library or a repository and pointed out that with the rapid expansion of databases, small-scale subject controlled vocabularies and classification schemes fail, and that the larger the collection is (or is projected to be in future) the more sophisticated controlled vocabularies of subject terms it requires. From this point of view, as the world's most extensive controlled vocabulary of subject terms, LCSH holds promise for describing large collections. However, complexities of LCSH controlled vocabulary and its application in representing information objects (e.g., heading structure variations – inversed or direct phrase entries, inconsistency in subdivision practice, synonymous terms used in different headings, etc.) often negatively affect search performance in databases (e.g., Larsen, 1991A). Polyrepresentation of information objects (Ingwersen, 1994) where the system

contains multiple sets of metadata (e.g., both controlled-vocabulary terms and author- or user-generated keywords or tags) has been viewed as a possibility for improving subject access to large databases. The 2008 report by the US Library of Congress on the future of information representation recommended an integration of the user-contributed data (tags or keywords) into metadata records (Working Group on the Future of Bibliographic Control, 2008, p. 32).

ETD developers try to find answers to questions such as what type of subject headings are important to include (Baker, 2017), and "is it even necessary to include subject information in records when there is direct access to the full-text of the publication?" (Alemneh and Phillips, 2016; Waugh, Tarver, Phillips and Alemneh, 2015).

We believe high-quality subject metadata for ETDs is still important to facilitate aggregation of ETD records in national and international portals such as Global ETD Search (http://search.ndltd.org/) by NDLTD or EBSCO's Open Dissertations tool (https://biblioboard.com/opendissertations/). These systems rarely have functionality to provide full-text search and rely on metadata, particularly subject metadata for topical searching. Moreover, controlled-vocabulary subject metadata has been found crucial for facilitating access to relevant information objects in English language and especially in other languages even in full-text environments (e.g., Gross & Taylor, 2005; Gross, Taylor, & Joudrey, 2015; Garrett, 2007).

One of the main functions of databases providing access to information objects (including ETDs) is to represent relationships between information objects, based on various factors (most importantly subjects and creators or contributors shared by these information objects). Since 1980s-1990s, research suggested that information systems should be judged by success in answering questions through supporting browsing, and that exploratory design models are needed (Borgman, 1996; Hildreth, 1995). For example, the United States Council on Library Research, based on the results of its nationwide catalog use survey (Matthews, Lawrence, & Ferguson, 1983), recommended increasing the amount of subject information in bibliographic records, and restricting the number of possible search terms "either by rigorously controlling the vocabulary or by automatically linking the user's search terms with synonymous and related terms that appear in subject headings" (p. 178-179). According to Bawden & Vilar (2006), followability of data (e.g., ability to quickly get access to related objects through retrieved results) in library catalogs and digital libraries is an important part of changing user expectations that are shaped by experiences with major search engines and transactional sites (e.g., Google and Amazon) and societal changes in general (e.g., perceived need for more information-rich environment). It has been observed that users find system functions supporting user tasks involved in resource discovery by subject, including collocation by subject options, helpful in searching (Zhang & Salaba, 2007). Library and information science community at large agrees that important objectives of navigation support

and information use support should be included in the conceptual models that serve as frameworks for ensuring metadata functionality and as a result, the "explore" user task is now part of Library Reference Model (Consolidation Editorial Group of the IFLA FRBR Review Group, 2017).

Our analysis of the literature demonstrates the need for studies examining how well the metadata supports the functions of navigation, exploration, and use through providing links between information objects based on the data values in metadata records representing them (especially in the area of subject representation). This work seeks to understand how data values of subject fields in metadata records (including topical terms and names) connect ETD documents in repositories by creating networks or graphs that treat metadata records as nodes, connected through the subject data values that they share. These networks often manifest functionally in digital repositories that allow users to click on a link for a subject term to find all metadata records containing that value.

Methodology

As a target for this study, we selected the UNTETD collection, which allows for comparisons of controlled-vocabulary subject metadata and uncontrolled subject terms due to the structure of the UNT Libraries (UNTL) metadata scheme used for representing resources in this repository. The UNTL metadata allows both uncontrolled keywords and controlled subject terms from the US Library of Congress Subject Headings (LCSH) and metadata implementation guidelines for the ETD collection require inclusion of at least two terms of any type. We generated both traditional, count-based statistics and network analysis statistics to provide comparisons and context for understanding metadata in the UNTETD collection. Standard statistics came from raw (native format) metadata harvested May 10, 2019 using the Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH) from the repository endpoint (https://digital.library.unt.edu/explore/collections/UNTETD/oai/).

A Metadata Record Graph was created for the subject field in the UNTETD collection by building a list of records sharing a subject field data value between them. Combination pairs of connected records were generated and grouped to create an adjacency list with a metadata record (node) identifier as the key, paired with identifiers for other metadata records connected by any shared data value in the subject field. This adjacency list is used to generate network statistics for the subject Metadata Record Graph.

Analysis of the traditional, count-based and network statistics include a number of calculations. Entropy calculates the level of similarity in terms as a probability that a new term in the field will be unique. Density measures connectedness as the number of edges (connections) versus possible edges. Compression represents the amount of change in unique values from normalization. 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).

Results and Discussion

We reviewed count-based statistics for the UNTETD collection first. Nearly every publicly visible record contains at least one instance of a subject field, and the most common number of subject terms is 5 per record (see Table 1).

	Total Values	Unique Values	Maximum Entries	Minimum Entries	Mode	%Mode
All Subjects (both controlled and uncontrolled)	104,341	62,678	73	0	5	18%
Controlled: LCSH terms	34,490	21,174	13	0	0	32%
Uncontrolled: Keywords	69,869	41,530	73	0	3	38%

Table 1

Count-Based Metrics for the Subject Metadata Field Data Values

The diversity in the application of controlled terms (LCSH) versus uncontrolled keywords is much wider: though there are more total keywords, a slightly larger percentage of LCSH terms used in the records (61%) are unique compared to keywords (59%). This is explainable by the fact that LCSH subject headings are often represented or longer phrases (e.g., "work in literature") as subject strings (e.g., "Brontë, Emily, 1818-1848 -- Criticism and interpretation") as opposed to generally much shorter uncontrolled terms in most cases consisting of a single word (e.g., "work"). We also looked at the distribution of subject terms by ETD publication year (Figure 1) to see if there were disparities in coverage. While it did not provide definitive information of use to this research, it is clear that there are large variances in the assignment of LCSH terms, in particular, depending on the year of publication.

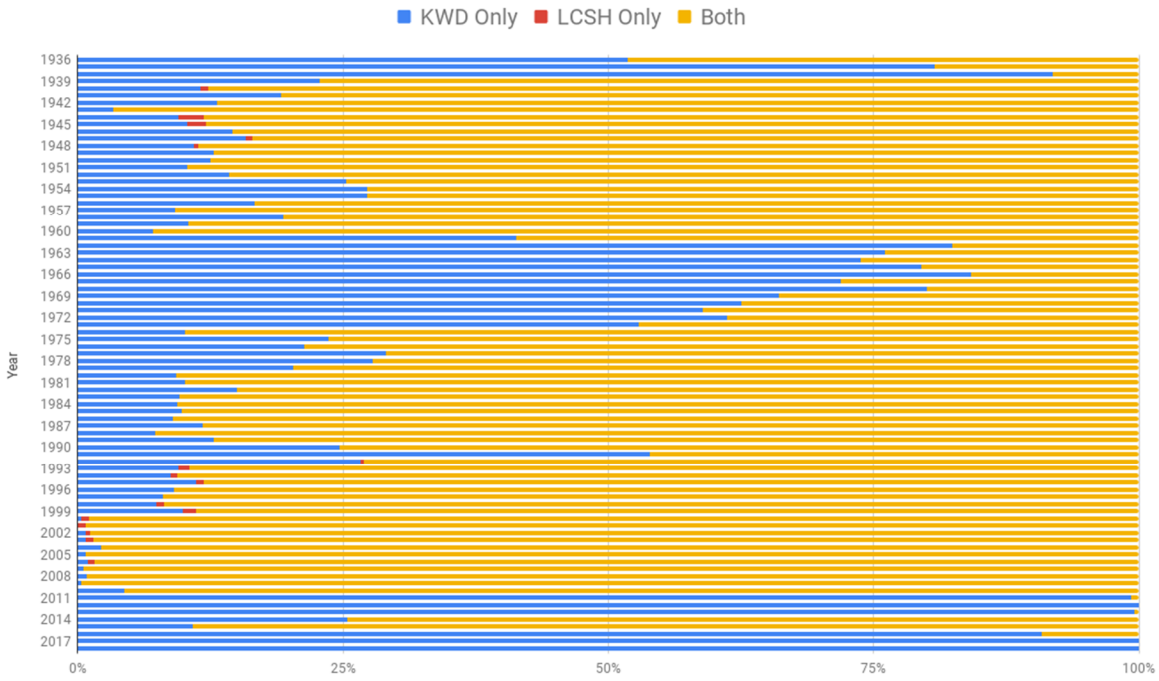

Figure 1

Distribution of subject field data values by the year of creation of ETD documents

To better understand overlap between controlled and uncontrolled terms, we generated Metadata Record Graphs for the 19,291 nodes in the subject element (see Table 2), including one for all subject field data values, and one each specifically for LCSH terms and keywords.

	Con-nected Nodes	Uncon-nected Nodes	Total Edges	Density	Average Degree	Degree Mode	Frequen-cy of Mode Degree	Degree Distri-bution Gini Coeffici ent
All Subject terms (both controlled and uncontrolled)	17,616	1,675	425,665	0.0023	44	0	9%	0.65
Controlled: LCSH terms	9,099	10,192	88,412	0.0005	9	0	53%	0.84
Uncontrolled: Keywords	16,096	3,195	345,551	0.0019	36	0	17%	0.71

Table 2

Metadata Record Graphs Based on the Subject Metadata Field Data Values, n=19,291

Perhaps the easiest metric to interpret in network statistics is connected vs. unconnected nodes. There is much more overlap (i.e., connections between records) among uncontrolled data values (keywords) versus controlled vocabulary terms (LCSH). However,

more records (32%) do not contain LCSH data values (see Table 1); identifying these records and adding LCSH terms would provide greater coverage and connect additional nodes.

There are fewer connected nodes (records) among specific types of subject terms (LCSH and keywords) than overall for all subject terms. Unconnected nodes in these graphs have unique subject data values with no connections to other records (i.e., a degree of zero). The degree mode for all of these graphs is 0, but the LCSH graph has the highest percentage of nodes with a degree of 0 (53%). Figure 2 provides the degree distribution on both linear and logarithmic scales.

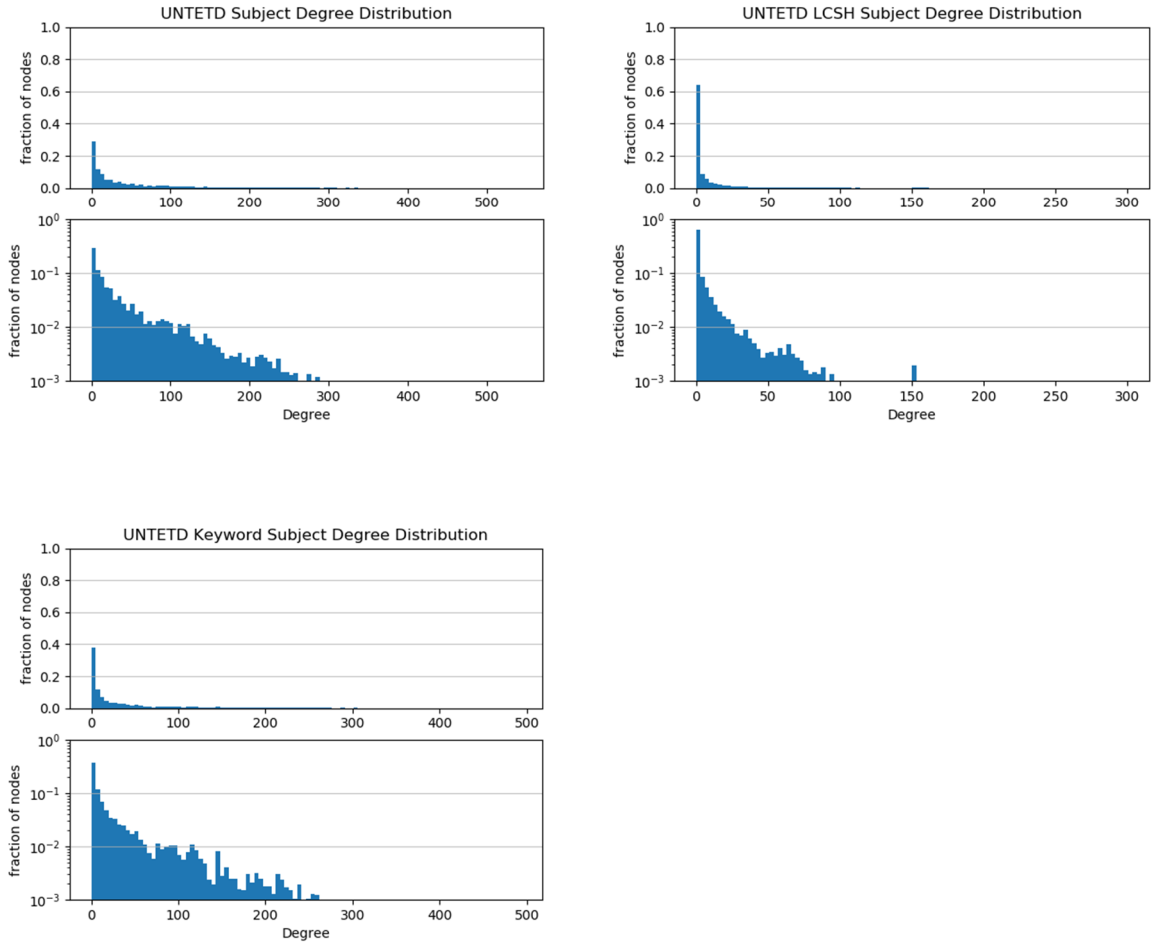

Figure 2

Degree Distributions for Metadata Record Graphs

The distribution we see in the plots of the degree distributions in Figure 2 are what we can expect from real world networks and shows a highly right-skewed distribution which means that a large majority of the nodes have a low degree but there are a small number of nodes that have a high degree. This shows a diversity in subject term used in metadata records for ETDs. This makes sense in thinking about the diversity of topics that are chosen in a large university with a wide range of colleges and departments.

Normalizations of Data Values in Subject Metadata Fields

We then generated the same network statistics for normalized versions of the subject fields' data values using a number of algorithms. These numbers attempt to simulate the potential connectivity if editors were to change subject values to account for mistakes and inconsistencies in formatting (e.g., extra spaces or punctuation, lowercase vs. uppercase) or in authority (e.g., LCSH authorized subjects or names vs. keyword versions).

For the purposes of this research, we applied the following kinds of algorithms, singly, or in combination (see Table 3). Note that, in this context, a "token" is an individual word or component of a term.

Algorithm	**Normalization process**	**Normalized Version/"Key"**
n/a	None: original data value	Brontë, Emily, 1818-1848 -- Criticism and interpretation.
Lowercase	All letters are made lowercase	brontë, emily, 1818-1848 -- criticism and interpretation
Punctuation	Removes all punctuation but leaves spacing around tokens	Brontë Emily 1818 1848 Criticism and interpretation
Whitespace	Collapses multiple spaces to a single space between tokens	Brontë, Emily, 1818-1848 -- Criticism and interpretation.
Convert Accents	Downgrades Unicode characters to an ASCII equivalent	Bronte, Emily, 1818-1848 -- Criticism and interpretation.
NACO	Lowercase, remove leading/trailing spaces and diacritics, convert super- and subscript numbers to digits, convert symbols (except # & +) to blanks, and convert some characters to ASCII values	bronte emily 1818 1848 criticism and interpretation
Fingerprint	All basic normalizations (lowercase, punctuation, whitespace, convert accents); then tokens are alphabetized and de-duplicated	1818 1848 and bronte criticism emily interpretation

Table 3

Description of algorithms used to normalize subject values in the UNTETD Collection

In addition to string-based normalizations we performed four additional transformations aimed at converting words to their base form, including lemmatization using WordNet, and three stemmers, Porter, Snowball-English, and Lancaster. These transformations allow us to see what might happen if the subject values are controlled for the use of plural and singular versions of subject.

After running these normalizations, we re-evaluated count-based statistics (see Table 4), including the number of unique values and entropy.

	Keywords		LCSH		All Subjects	
	Unique Values	Entropy	Unique Values	Entropy	Unique Values	Entropy
Original data values	41,531	0.94621	21200	0.95748	62,705	0.95264
Normalized data values:						
Lowercase	39,241	0.94066	21,186	0.95745	60,390	0.94901
Punctuation	41,403	0.94603	21,153	0.95736	61,736	0.95148
Whitespace	41,529	0.94621	21,196	0.95748	62,699	0.95263
Convert Accents	41,519	0.94621	21,200	0.95748	62,693	0.95263
NACO	38,935	0.94014	21,033	0.95723	57,437	0.943
Fingerprint	38,764	0.93952	20,953	0.95717	57,036	0.94247
NACO+Fingerprint	38,768	0.93952	20,951	0.95717	57,037	0.94246
Lemmatize	37,700	0.93691	21,028	0.95722	56,004	0.94056
Porter	36,759	0.93436	20,986	0.95706	54,944	0.93855
Snowball-English	36,762	0.93441	20,984	0.95706	54,943	0.93859
Lancaster	35,900	0.93193	20,918	0.95668	53,999	0.93659

Table 4

Count-based statistics for data values in subject metadata fields after normalization

Overall, entropy in this data tends to decrease in direct relation to the number of unique terms. The amount of variability in LCSH terms is relatively low, especially compared to similar normalizations among uncontrolled keywords; this is expected given the different origins of author-provided keywords versus staff-supplied LCSH terms by trained catalogers. The difference in both unique values and entropy is most significant for all subjects combined, for example dropping by more than 5,000 unique terms (62,705 to 57,437) with NACO normalization. however, counts only provide one facet of information, so we also generated network statistics to see how changes in unique values manifested as connected nodes (see Table 5).

	Keywords		LCSH		All Subjects	
	Connected Nodes	Unconnected Nodes	Connected Nodes	Unconnected Nodes	Connected Nodes	Unconnected Nodes
Original data values	16,096	3,195 (7.7%)	9,099	10,192 (48.1%)	17,616	1,675 (2.7%)
Normalized data values:						

Lowercase	16,502	2,789 (6.7%)	9,102	10,189 (48.1%)	17,834	1,457 (2.3%)
Punctuation	16,122	3,169 (7.6%)	9,117	10,174 (48.0%)	17,708	1,583 (2.5%)
Whitespace	16,096	3,195 (7.7%)	9,104	10,187 (48.1%)	17,616	1,675 (2.7%)
Convert Accents	16,099	3,192 (7.7%)	9,099	10,192 (48.1%)	17,618	1,673 (2.7%)
NACO	16,573	2,718 (6.5%)	9,185	10,106 (47.7%)	18,047	1,244 (1.9%)
Fingerprint	16,629	2,662 (6.4%)	9,222	10,069 (47.5%)	18,113	1,178 (1.9%)
NACO+Fingerprint	16,624	2,667 (6.4%)	9,224	10,067 (47.5%)	18,107	1,184 (1.9%)
Lemmatize	16,802	2,489 (6.0%)	9,187	10,104 (47.7%)	18,168	1,123 (1.8%)
Porter	16,928	2,363 (5.7%)	9,197	10,094 (47.6%)	18,236	1,055 (1.7%)
Snowball-English	16,927	2,364 (5.7%)	9,198	10,093 (47.6%)	18,238	1,053 (1.7%)
Lancaster	17,011	2,280 (5.5%)	9,210	10,081 (47.6%)	18,268	1,023 (1.6%)

Table 5

Network statistics for data values in the subject metadata field after normalization

In terms of connections, many of the basic normalizations have relatively little effect, particularly within the keyword and LCSH subsets. This does not seem unexpected since each of those normalizations would only create new connections between terms with specific types of differences -- i.e., terms that are identical except for a single difference in capitalization, punctuation, spacing, *or* diacritics. However, the combination normalizations (NACO, fingerprint, and NACO with fingerprint) show a fairly significant change in the number of connected nodes, up to 98% connectivity across all subjects.

For additional comparison, we have included more detailed count-based and network statistics for both types of data values in subject metadata fields separately and overall in Table 6.

	Count-Based Statistics			Network Statistics		
	Unique	Compression	Entropy	Connected Nodes	Unconnected Nodes	Density
Original data values	62,705		0.95263	17,616	1,675	0.00229
Normalized data values:						
Lowercase	60,390	0.04	0.94901	17,834	1,457	0.00258
Punctuation	61,736	0.02	0.95148	17,708	1,583	0.00233
Whitespace	62,699	0.01	0.95263	17,616	1,675	0.00229
Convert Accents	62,693	0.01	0.95263	17,618	1,673	0.00229

NACO	57,437	0.09	0.94300	18,047	1,244	0.00290
Fingerprint	57,036	0.1	0.94247	18,113	1,178	0.00294
NACO+Fingerprint	57,037	0.1	0.94246	18,107	1,184	0.00294
Lemmatize	56,004	0.11	0.94056	18,168	1,123	0.00310
Porter	54,944	0.13	0.93855	18,236	1,055	0.00323
Snowball-English	54,943	0.13	0.93859	18,238	1,053	0.00322
Lancaster	53,999	0.14	0.93659	18,268	1,023	0.00336

Table 6

Count-based and network statistics for all subject field data values after normalization

One noticeable comparison is that the variation in compression (i.e., amount of change in unique terms) does seem to have a direct relation to density (overall connectivity). As compression increases, density does also; however, actual total change in density seems to be fairly minimal. Additionally, the compression shows varying levels of normalization, depending on how aggressively each algorithm strips values down, particularly the stemming algorithms. The most aggressive stemming algorithm -- Lancaster -- reduces total unique terms from 62,678 to only 53,999. While this would increase connectivity (density .00336), it almost certainly introduces errors, matching terms that may be similar, but not significantly different. Similarly, some of the least aggressive algorithms almost certainly miss terms that ought to be the same. For comparison, Table 7 lists compression and density from each of the three Metadata Record Graphs.

	Compression			**Density**		
Normalization algorithm	**KWD**	**LCSH**	**All Subjects**	**KWD**	**LCSH**	**All Subjects**
Lowercase	0.06	0.01	0.04	0.00216	0.000475	0.00258
Punctuation	0.01	0.01	0.02	0.00186	0.000477	0.00233
Whitespace	0.01	0.01	0.01	0.00186	0.000475	0.00229
Convert Accents	0.01	0	0.01	0.00186	0.000475	0.00229
NACO	0.07	0.01	0.09	0.00218	0.000479	0.00290
Fingerprint	0.07	0.02	0.1	0.00221	0.000480	0.00294
NACO+Fingerprint	0.07	0.02	0.1	0.00221	0.000480	0.00294
Lemmatize	0.1	0.01	0.11	0.00237	0.000479	0.00310

Porter	0.12	0.02	0.13	0.00248	0.000480	0.00323
Snowball-English	0.12	0.02	0.13	0.00248	0.000480	0.00322
Lancaster	0.14	0.02	0.14	0.00258	0.000485	0.00336

Table 7

Network statistics for subject field data values after normalization

Changes in density and compression have different characteristics when looking at only keywords, only LCSH terms, or all subject terms combined. In terms of compression, LCSH terms consistently have significantly low numbers, with no compression for ASCII conversion. This may be partially due to the fact that there are many fewer unique LCSH terms than keywords, but also suggests that LCSH terms are already more consistent in formatting. In comparison, keywords are applied in greater numbers, by various people, including authors who are self-submitting information and may not be adhering to formatting rules (such as punctuation and capitalization), which would lead to higher compression when those aspects are normalized. For all subjects, overall density is often higher than the combined density of keywords or LCSH terms individually. This also makes sense if keyword values and LCSH terms in separate records are used to represent similar topics but have different formatting; normalizing between keywords and controlled terms would create connections within those topics that would not be linked using exact string matching. Comparing all subject values would also negate any incorrect qualifiers (i.e., keywords accidentally labelled as LCSH or vice versa).

Conclusion

Metadata Record Graphs provide additional information about subject metadata that supplement traditional count-based statistics. Subject counts can give a general sense of how much information is included in each record and compression can demonstrate change in unique values through normalization, but neither gives an accurate picture of connectivity. For example, in this collection, there are likely some records that may contain nearly identical LCSH terms and keywords, since they are often assigned at different times in our workflows. Formatting normalization would compress these values, but connectivity would not increase, as both values would still be in a single record. As such, compression versus density of combined subjects may also provide a sense of how common this scenario is rather than unique subject counts.

In terms of normalization algorithms, this research used a number of approaches with varying levels of aggression regarding the degree to which subject values were stripped or normalized before matching like values. No automated tool can be completely accurate and each of the algorithms introduce a degree of error, by missing some matches or creating false matches. The most efficient algorithms are likely the ones somewhere in the middle, that

combine several normalizations (e.g., NACO or fingerprint algorithms). Depending on resources, possible matches from multiple algorithms should be reviewed to be most effective, if the goal is to use the data to change values and operationally make subjects more connected with exact string matching.

Network analysis can play a valuable role in understanding metadata in a collection, but it is one of many tools needed to fully understand the complexities and to offer possible avenues of improvement by modifying records.

References

Alemneh, D. G. & Phillips, M. E. (2016). Indexing quality and effectiveness: An exploratory analysis of electronic theses and dissertations representation. *Proceedings of the Association for Information Science and Technology 53*: 1-4. https://doi.org/10.1002/pra2.2016.14505301111

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

Baker, W. (2017). *Controlled Vocabularies in the Digital Age: Are They Still Relevant?* (Doctoral dissertation). Retrieved from https://digital.library.unt.edu/ark:/67531/metadc1011802/

Bates, M. (2002). After the dot-bomb: getting web information retrieval right this time. *First Monday,* *7*(7). Retrieved from http://firstmonday.org/issues/issue7_7/bates/index.html

Bawden, D., & Vilar, P. (2006). Digital libraries: to meet or manage user expectations. *Aslib Proceedings, 58*(4), 346-354.

Borgman, C. (1996). Why are online catalogs still hard to use? *Journal of American Society for Information Science, 47* (7), 493-503.

Consolidation Editorial Group of the IFLA FRBR Review Group. (2017). *IFLA Library Reference Model: A Conceptual Model for Bibliographic Information.* Retrieved from https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017_rev201712.pdf

Curl, M. (1995). Enhancing subject and keyword access to periodical abstracts and indexes: Possibilities and problems. *Cataloging & Classification Quarterly, 20*(4), 45-55.

Fidel, R. (1988). Factors affecting the selection of search keys. In *Proceedings of the 51st annual meeting of the American Society for Information Science, volume 25, Atlanta, Georgia, 23-27 October 1988* Medford, NJ: Learned Information, pp. 76-79.

Fidel, R. (1992). Who needs controlled vocabulary? *Special libraries, 83*(1), 1-9.

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

Gross, T., & Taylor, A.G. (2005). What have we got to lose? The effect of controlled vocabulary on keyword searching results. *College & Research Libraries*, *66*(3), 212-230. https://doi.org/10.5860/crl.66.3.212

Gross, T., Taylor, A.G, & Joudrey, D. (2015). Still a lot to lose: The role of controlled vocabulary in keyword searching. *Cataloging & Classification Quarterly*, *53*(1), 1-39. https://doi.org/10.1080/01639374.2014.917447

Hildreth, C. (1995). *Online Catalog Design Models: Are We Moving in the Right Direction?* http://www.ou.edu/faculty/Charles.R.Hildreth

Hildreth, C. (1997). The use and understanding of keyword searching in a university online catalog. *Information Technology and Libraries, 16*(2), 52-62.

Ingwersen, P. (1994). Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction. In *Proceedings of the SIGIR 1994,* 101-110.

Larson, R. (1991). Between Scylla and Charybdis: Subject searching in online catalogs. *Advances in Librarianship, 15,* 175-236.

Lubas, R. L. (2009) Defining Best Practices in Electronic Thesis and Dissertation Metadata, *Journal of Library Metadata*, *9*:3-4, 252-263 https://doi.org/10.1080/19386380903405165

Matthews, J., Lawrence, G., & Ferguson, D. (Eds.), (1983). *Using Online Catalogs: A Nationwide Survey*: *A Report of a Study Sponsored by the Council on Library Resources.* New York, NY: Neal-Schumann.

Muddamalle, M. (1998). Natural language versus controlled vocabulary in information retrieval: A case study in soil mechanics. *Journal of American Society for Information Science, 49*(10), 881-87.

Networked Digital Library of Theses and Dissertations. (2009) *ETD-MS V1.1 an Interoperability Metadata Standard for Electronic Theses and Dissertations.* Retreieved from http://www.ndltd.org/standards/metadata

Swain, D. K. (2010). Global Adoption of Electronic Theses and Dissertations. *Library Philosophy and Practice*, vol. Annual, n. August. [Journal article (Unpaginated)] Retrieved from https://digitalcommons.unl.edu/libphilprac/418/

Texas Digital Library ETD Metadata Working Group. (2015) *Dictionary of Texas Digital Library Descriptive Metadata for Electronic Theses and Dissertations, v. 2*. Retrieved from http://hdl.handle.net/2249.1/68437

UNTLibraries' Digital Projects Unit (n.d.) *Metadata Implementation Guidelines for UNT ETDs* Retrieved from https://library.unt.edu/digital-curation-unit/etd/metadata-implementation-guidelines/

Waugh, L., Tarver, H. Phillips, M. E., & Alemneh, D. G. (2015). *Comparison of Full-text Versus Metadata Searching in an Institutional Repository: Case Study of the UNT Scholarly Works*. Retrieved from http://arxiv.org/abs/1512.07193

Working Group on the Future of Bibliographic Control. (2008). *On the Record: Report of the Library of Congress Working Group on the Future of Bibliographic Control*. Retrieved from http://www.loc.gov/bibliographic-future/news/lcwg-ontherecord-jan08-final.pdf

Zhang, Y., & Salaba, A. (2007). User research and testing of FRBR prototype systems: Poster. *70th ASIS&T Annual Meeting* (Milwaukee WI, Oct. 19-24, 2007).