

Empirical Evaluation of ETD-ms Compliance for Electronic Theses and Dissertations Metadata

Harvested by the NDLTD Union Catalog

Adrian Chisale, Cecilia Kasonde and Lighton Phiri

Department of Library and Information Science, University of Zambia

Author Note

We have no conflicts of interest to disclose.

Correspondence concerning this article should be addressed to Lighton Phiri, DataLab Research Group, Department of Library and Information Science, University of Zambia, P.O Box 32379, Lusaka, Zambia.

Email: lighton.phiri@unza.zm

Abstract

There are presently a number of academic institutions that have set up Institutional Repositories (IRs) aimed, in part, to archive, manage and facilitate access to Electronic Theses and Dissertations (ETDs). Increasingly, descriptive metadata associated with the ETDs produced by most of these institutions are now automatically harvested by National ETD portals and entities such as The Networked Digital Library of Theses and Dissertation (NDLTD) Union Catalog. However, there are growing concerns regarding the quality of ETD descriptive metadata originating from most of these institutions, in particular, adherence to the ETD-ms metadata standard, the de-facto standard used to describe ETDs. This paper presents an empirical analysis of ETD-ms compliance of the metadata associated with the 7,563,684 ETDs in the NDLTD Union Catalog. ETD metadata records were harvested from the NDLTD Catalog using the OAI-PMH protocol and, subsequently, an analysis was conducted to determine compliance to the ETD-ms metadata standard. In addition, to provide a contextual overview of the potential root causes of non-compliance to ETD-ms, a case study was conducted at The University of Zambia (UNZA), in order to understand current practices and procedures employed during ingestion and association of metadata. The results indicate that compliance with the ETD-ms standard is low, with most metadata elements associated with missing values. The case study conducted at UNZA suggests that policy and awareness are crucial in facilitating compliance to metadata standards such as ETD-ms. The study could potentially provide direction on issues to be addressed during the ingestion of ETDs into IRs.

Keywords:

ETD-ms, Metadata, NDLTD Union Catalog, Quality Metadata

Introduction

A number of Higher Education Institutions (HEIs) have set up Institutional Repository (IR) platforms which, in part, are used to store, manage and archive Electronic Theses and Dissertations (ETDs). With the increase in the number of ETDs produced, the use of aggregate services for harvesting ETDs from different HEI IRs is becoming commonplace. The Networked Digital Library of Theses and Dissertations (NDLTD) Union Catalog portal is one of the largest global aggregate services, designed to collect ETD metadata records from institutions around the world and subsequently aggregate them into a single collection (Suleman, 2004).

While there is a significant number of HEIs from where the NDLTD Union Catalog harvests metadata records from, the poor quality of metadata records harvested has been cited as a longstanding issue. For instance, Hussein highlights that while the recommended metadata standard for describing ETDs is ETD-ms, there is low adoption of the standard (Suleman, 2012). In addition, our previous work highlighted implications of poor quality metadata, including a potential solution for automatically classifying ETDs (Phiri, 2021).

This work was aimed at understanding the extent to which conformance to the ETD-ms scheme is adhered by HEIs whose metadata is harvested by the NDLTD Union Catalog. In addition, a case study was conducted at the University of Zambia (UNZA) in order to understand the factors that can potentially result in poor metadata quality.

The remainder of this paper is organized as follows: the “Related Work” section discusses literature related to this work; the “Methodology” section outlines the methodological approach followed when conducting this work; the “Results and Discussion” section presents and discusses the findings and, finally, the “Conclusion” section presents concluding remarks and potential future work.

Related Work

Metadata Quality

Electronic theses and dissertations represent a significant part of academic publications. Because of this, it is important that the way these vital elements of a university's collection are properly stored and they are readily accessible to the institution's researchers as well as members of the broader public. By utilising metadata standards, institutions can ensure the availability of these key resources

Metadata quality is, obviously, not only an issue for the cultural heritage sector. A large body of research, development and tools has been developed throughout the 1990's within the computer science field, the corporate world and public administrations to examine the notion of data or information quality. A multitude of other denominators and sub activities, such as data cleaning, profiling and standardisation exist (Batini & Scannapieca, 2006).

Bruce and Hillman offer a useful examination of characteristics of metadata quality particularly in light of its importance to aggregated collections (Bruce & Hillmann, 2004). They outline seven general characteristics of metadata quality: completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility. In addition, they offer some possible criteria and compliance indicators for each, noting that shared metadata may require additional quality efforts.

A good metadata enhances the value of a resource, provided that it is based on good analysis of the resource. A high quality metadata helps users find what they need, even when they are not sure themselves about what they need (Alemneh, 2009).

Park notes that quality metadata reflects the degree to which the metadata in question performs the core bibliographic functions of discovery, use, provenance, currency, authentication, and administration. Further, the functional perspective is closely tied to the criteria and measurements used for assessing metadata quality. Accuracy, completeness, and consistency are the most common criteria

used in measuring metadata quality in the literature (J.-R. Park, 2010). Results of the study indicate a pressing need for building of a common data model that is interoperable across digital repositories.

The metrics used to measure quality of metadata include completeness, accuracy, consistency, objectiveness, appropriateness and correctness.

Ingestion of High Quality Electronic Theses and Dissertations

Metadata is becoming more than a tool to facilitate access and retrieval; librarians and other metadata professionals and their users are expecting metadata to perform multiple and diverse purposes and functions: as much as metadata helps connect users to resources, it is expected to appropriately situate resources in relationship to other resources, and within historical and contemporary social contexts (Jaffe, 2020).

In Chassanoff's study forty respondents provided results that offer a state of the art view into the current metadata quality practices in place at IRs. Survey results indicated that metadata activities may not yet be streamlined into institutional workflow. For most institutions, metadata quality checking is a manual process, with only a small percentage (4%) employing the use of automated tools. Additionally, institutions rely on users as much as repository staff to discover quality problems. Other results indicated that the majority of institutions surveyed were maintaining documentation relating to metadata policies. For example, 75% of respondents reported that their institution had developed either minimum metadata requirements or metadata submission guidelines for contributors. Overall, these results reflect the challenges and growing pains facing institutions as they adapt to managing materials in the digital world (Chassanoff, 2009).

Electronic Theses and Dissertations Metadata Standard

Metadata standards provide a predefined and structured way of describing data. While there are a number of metadata standards available (Hillmann et al., 2008), the decision to adopt a standard is generally dependent on the domain and use case.

The NDLTD has developed a metadata. ETD-ms is a metadata standard that is based upon Dublin Core (Pavani et al., 2009). It was developed by the NDLTD to act as a comprehensive ETD metadata standard. This standard provides instructions for the use of Dublin Core elements for the cataloging of ETDs. According to Park and Richard, the ETD -ms standard comprises 21 elements—14 basic elements and an additional seven supplemental elements (E. G. Park & Richard, 2011). Of the basic ETD-ms elements, 13 correspond to basic Dublin Core elements—title, creator, subject, description, publisher, contributor, date, type, format, identifier, language, coverage, and rights. Furthermore, two Dublin Core elements—relation and source—are not used by the ETD-ms standard. Having a set of standards to follow is very helpful, but not every IR follows these conventions.

The development and application of ETD metadata standards, and the resulting quality, consistency, and interoperability of the metadata produced and exchanged, incur major implications for the discovery and long-term preservation of these unique student works, as one of the major goals of interoperability is enabling the implementation of coherent services for users, from components that are different and inherently hosted by different organisations (Potvin & Thompson, 2016).

Methodology

Research Design

The research used a quantitative approach and the design was experimental design. This design is a form of quantitative research approach which seeks to determine whether a specific treatment influences an outcome.

NDLTD Union Catalog Metadata Quality Completeness

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) (Lagoze et al., 2002) was used to harvest metadata from the NDLTD Union Catalog. The metadata harvested is as of June 13, 2022. Specifically, the “oai_dc” prefix was used to harvest metadata encoded using the Dublin Core

metadata standard. The “oai_dc” prefix was used in preference to existing ETD-ms specific prefixes¹ as the “ETD-ms” prefixes returned an incomplete total number of metadata records.

The Dublin Core encoded metadata elements were then crosswalked to equivalent ETD-ms elements. In order to determine the relative completeness of metadata, analyses were conducted to determine the proportion of metadata records that were missing crucial ETD-ms metadata element values.

Case Study: Electronic Thesis and Dissertation at The University of Zambia

In order to assess metadata quality and, specifically, compliance with the ETD-ms standard, ETD metadata records were harvested from The UNZA IR using the OAI-PMH protocol. In addition, a document analysis of the UNZA IR Policy was performed, to understand processes and procedures involved during depositing of content into the IR. Furthermore in order to understand the quality issues with metadata, online interviews were conducted with four (4) participants that were involved in the development of the IR policy, two (2) Library Assistants that are responsible for depositing content into the IR and, additionally, The Head of The Special Collection department of the Library. The participants were sampled using purposive sampling.

Results and Discussion

NDLTD Union Catalog: Empirical Analysis of Electronic Theses and Dissertations

A total of 7,563,684 metadata records were harvested, originating from a total of 13,954 collections², with most of the collections originating from North America, Brazil and South Africa, as shown in Figure 1. The harvested metadata records were then pre-processed in order to filter out metadata records that did not have corresponding harvesting identifiers conforming to the “oai:union.ndltd.org” pattern. The total metadata records that were ultimately used during data analysis was 4,901,643.

¹ <http://union.ndltd.org/OAI-PMH/request?verb=ListMetadataFormats>

² <http://union.ndltd.org/OAI-PMH/request?verb=ListSets>

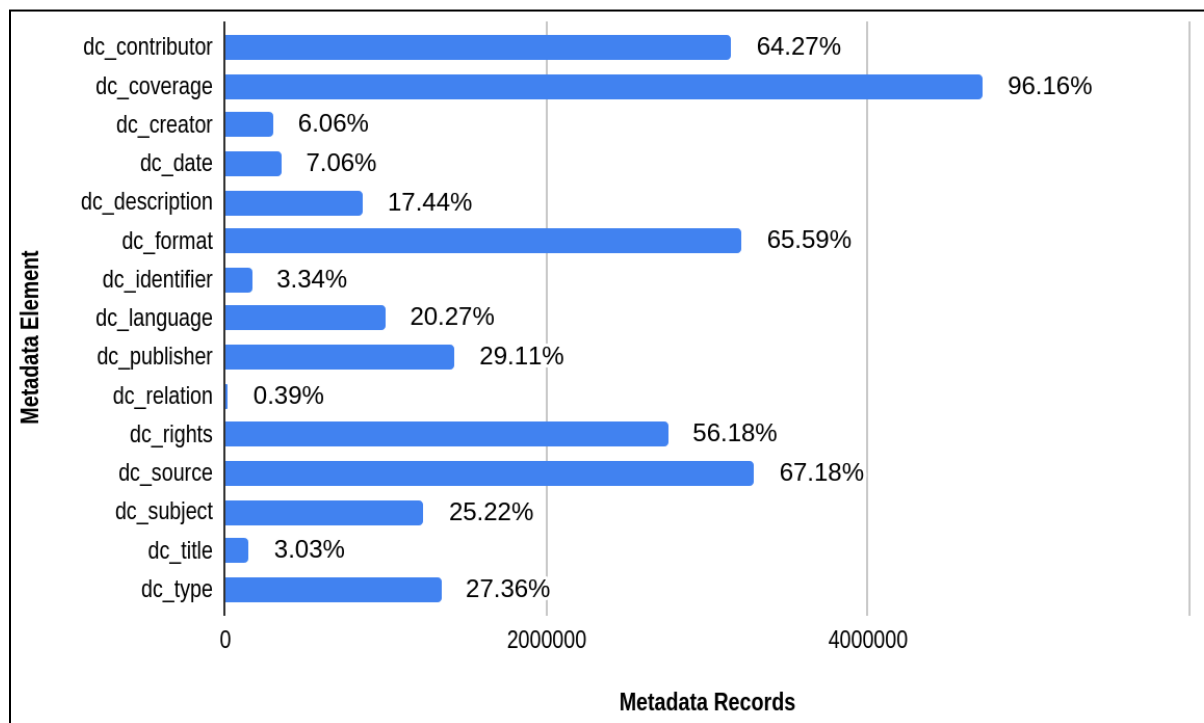
In order to determine compliance with the ETD-ms specification, an analysis of Simple Dublin Core elements was performed, as outlined in Section 4.1.1 and, additionally, an analysis of DC Terms associated with ETD-ms specification, as outlined in Section 4.2.2.

Dublin Core Elements

In order to assess the metadata elements that are populated during ingestion of ETDs, an analysis was performed to determine metadata records with null values. Figure 2 shows the proportion of null values for each of the 15 Simple Dublin Core elements.

Figure 1

Distribution of Null Metadata Records in NDLTD Union Catalog

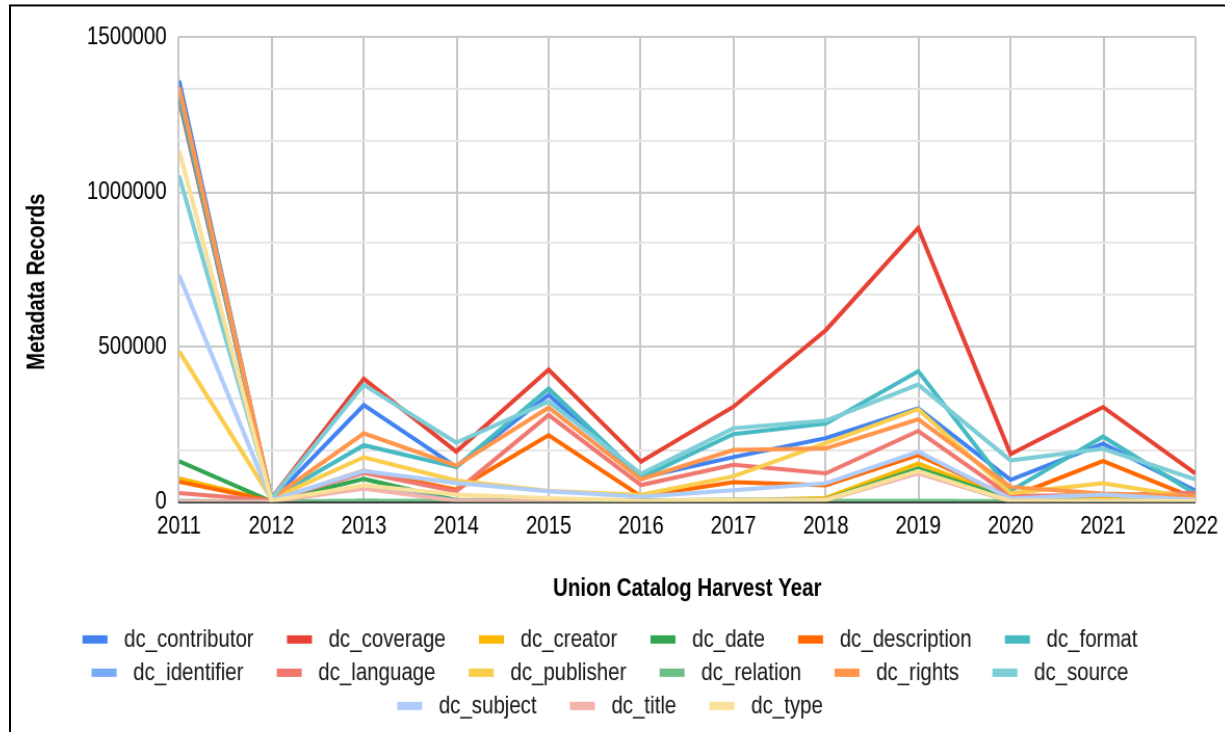


The results indicate that metadata elements with the least missing values are “dc:identifier”, “dc:title”, “dc:creator”, “dc:relation” and “dc:date”, with all of them associated with less than 10%

records with such missing entries. A random analysis of metadata records with null entries suggests that some of the records are associated with records that might have been deleted.

Figure 2

Annual Distribution of Null Metadata Element Values in NDLTD Union Catalog



While all elements associated with significantly more missing metadata elements are classified as being optional in the ETD-ms specification, some of the elements, such as “dc:contributor”, “dc:publisher” and “dc:type” are instrumental in comprehensively describing ETDs. Specifically, the “dc:contributor” element is primarily used for specifying the ETD Supervisor/Advisor, while “dc:publisher” and “dc:type” are used for specifying details about the ETD originating institution and degree classification, respectively.

A detailed analysis of the Simple Dublin Core elements was further conducted in order to analyse the annual trend of missing metadata elements. Figure 3 shows a negligible change in the

proportion of records with missing metadata elements. The results suggest that the metadata records harvested by the NDLTD Union Catalog will generally have crucial metadata elements missing.

ETD-ms Elements

In order to analyse distinct ETD-ms elements associated with the same Simple Dublin Core element, ETD-ms elements were mapped on to their equivalent Dublin Core elements, as shown in Table 1—the “ETD-ms” column shows results rendered³ from the DSpace-powered IR at The University of Cape Town, while the “Dublin Core” column shows the same record⁴, as stored in the NDLTD Union Catalog.

Table 1

Sample Elements Showing ETD-ms to Dublin Core Mapping

IR Element	Dublin Core	ETD-ms	Example Value
dc.contributor.advisor	dc.contributor	dc.contributor.role	Suleman, Hussein
dc.contributor.author	dc.contributor	dc.creator	Phiri, Lighton
dc.publisher.institution	dc.publisher	dc.publisher	University of Cape Town
dc.publisher.faculty	dc.publisher	dc.publisher	Faculty of Science
dc.publisher.department	dc.publisher	dc.publisher	Department of Computer Science
dc.type.qualificationlevel	dc.type	thesis.degree.level	Doctoral
dc.type.qualificationname	dc.type	thesis.degree.name	PhD

A further analysis was performed on metadata records with non-null entries, in order to determine the relative number of multiple values associated with repeatable Dublin Core elements.

Table 2 shows the number of multiple values associated with repeatable fields.

³ <https://open.uct.ac.za/handle/11427/29435>

⁴ <https://bit.ly/3cvM53U>

Table 2

Distribution of Non-Null Values

	1	2	3	4	5+
dc.creator	93.00%	5.00%	1.00%	0.00%	1.00%
dc.contributor	42.00%	24.00%	15.00%	7.00%	11.00%
dc.date	70.00%	5.00%	13.00%	10.00%	3.00%
dc.description	45.00%	36.00%	6.00%	5.00%	8.00%
dc.publisher	80.00%	8.00%	3.00%	4.00%	4.00%
dc.type	59.00%	29.00%	11.00%	1.00%	0.00%

The results indicate that with the exception of the “dc.contributor” and “dc.description” elements, which both have single value entries less than 50%, all other elements are associated with the highest proportion of single-value entries.

Case Study: Electronic Thesis and Dissertation at The University of Zambia

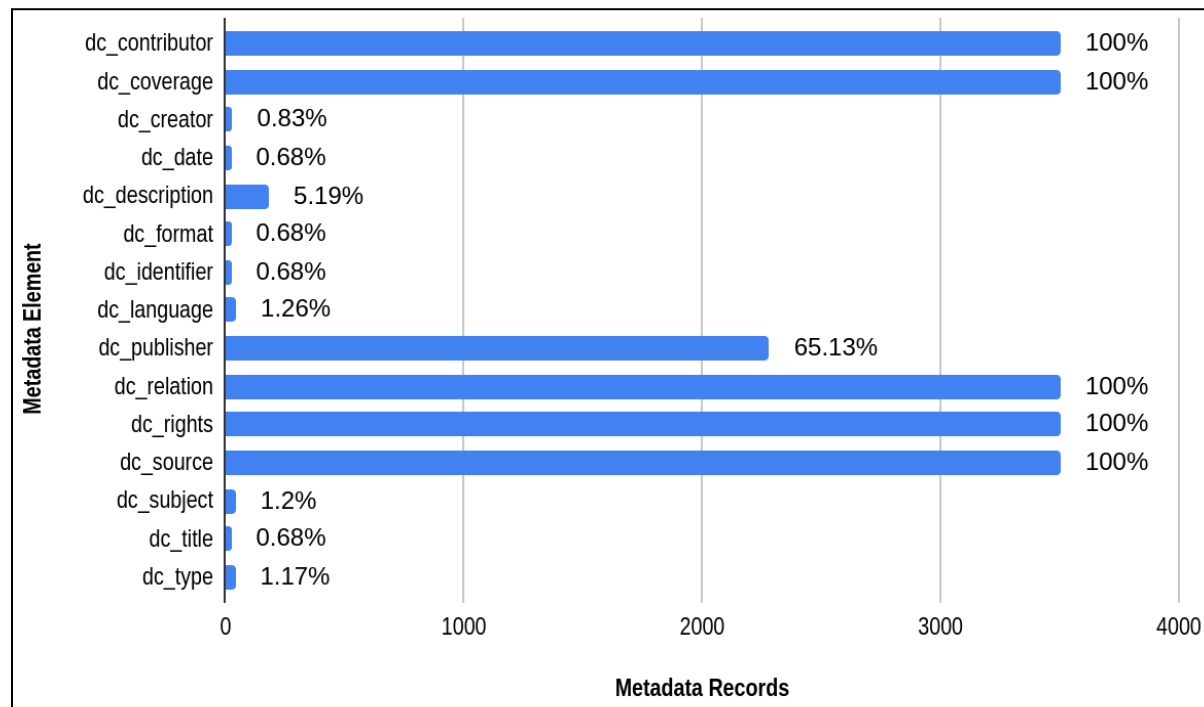
Empirical Analysis of Electronic Theses and Dissertations

6,716 metadata records were harvested from the UNZA repository. A total of 3,504 ETD metadata records were analysed from the original batch of harvested records, since The UNZA repository is used for archiving a variety of publication types, in addition to ETDs.

Figure 3 shows the distribution of null Simple Dublin Core null element values. The results indicate that only 10 Simple Dublin Core elements are used to describe ETDs, with the “dc:publisher” element associated with significantly more missing values. A further analysis of ETD-ms elements, focused on analysing multi-valued elements indicates the lack of use of metadata for describing “ETD Advisors”, “Faculty” and “Department” associated with the ETDs.

Figure 3

Distribution of Null Metadata Records in The University of Zambia Institutional Repository

***Institutional Repository Policy***

An analysis of The UNZA IR policy on metadata records indicates that content to be deposited into the IR should have the following mandatory elements: “Author”, “Title”, “Keywords (subject headings)”, “Abstract”, “Full-text”, “Imprint (Publisher, place and year of publication)” and “Item type”. This policy direction explains the distribution of the 10 Simple Dublin Core elements, described in Section 4.2.1.. This policy direction further explains why crucial metadata elements such as “Supervisor/Advisor” and “Faculty/Department” are not specified during ingestion of content into the IR.

An important point to note is that the UNZA policy does not explicitly distinguish how different content types—ETDs, Journal Articles, Conference Papers—should have metadata specified. In addition, the policy does not prescribe or recommend a specific type of metadata standard to be used.

ETD-ms Awareness

The interviews revealed that two (2) out of the four (4) Policy Makers interviewed were aware of ETD-ms. Furthermore, when interviewed, The Head of The Special Collection department in the Library confirmed his awareness of the existence and importance of ETD-ms. However, none of the Library Assistants were aware of the ETD-ms standard.

The results from interviews conducted with policy makers suggest that ETD-ms was not adopted as the prescribed metadata format as it was not convenient. Specifically, the choice for Dublin Core was influenced by it being the default metadata schema in DSpace. When asked why important metadata elements such as "Advisor/Supervisor" and "Faculty/Department" were not specified as mandatory elements, two (2) out of four (4) respondents indicated that there was no particular reason as to why the elements were omitted, while the other two (2) stated that it was not considered important at the time the IR Policy was being drafted.

Summary

The case study conducted at The UNZA suggests that policy direction and adoption of recommended metadata standards is crucial in ensuring that metadata is comprehensively associated with content deposited into IRs. More importantly, an understanding of the importance of prescribed and recommended metadata standards, by policy makers, is vital in ensuring that processes and procedures used during metadata preparation activities are properly performed.

Conclusion

This paper presented an empirical evaluation of ETD metadata records harvested from the NDLTD Union Catalog, with a specific focus on the analysis of common missing ETD-ms metadata elements. In order to understand the potential causes of non-compliance of ETD metadata to the ETD-ms standard, a case study was conducted at The UNZA—an HEI. The case study involved analysing

ETD metadata records in The UNZA IR, analysis of the institution's IR policy and, additionally, interviews with key stakeholders associated with preparation and ingestions of ETD metadata.

The rapid increase in the number of ETD metadata records harvested by the NDLTD Union Catalog and indeed other downstream services, including National ETD Portals such as the South African National ETD Portal (Webley et al., 2011), necessitate the the need for comprehensive and quality ETD metadata for increased discoverability of information, when using searching and browsing services such as the Global ETD Search platform (Suleman, n.d.).

References

- Alemneh, D. G. (2009). *Metadata Quality: A Phased Approach to Ensuring Long-Term Access to Digital Resources*. <https://digital.library.unt.edu/ark:/67531/metadc29318/>
- Batini, C., & Scannapieca, M. (2006). *Data Quality: Concepts, Methodologies and Techniques*. Springer Berlin Heidelberg.
- Bruce, T. R., & Hillmann, D. I. (2004). *The Continuum of Metadata Quality: Defining, Expressing, Exploiting*. <https://hdl.handle.net/1813/7895>
- Chassanoff, A. M. (2009). *Metadata Quality Evaluation in Institutional Repositories: A Survey of Current Practices*. <http://cdr.lib.unc.edu/downloads/t722hd588>
- Hillmann, D. I., Marker, R., & Brady, C. (2008). Metadata Standards and Applications. *The Serials Librarian*, 54(1), 7–21.
- Jaffe, A. (2020). Social Reproduction Theory and the Form of Labor Power. *CLCWeb: Comparative Literature and Culture*, 22(2), 3.
- Lagoze, C., Van de Sompel, H., Nelson, M., & Warner, S. (2002). *Open Archives Initiative - Protocol for Metadata Harvesting - v.2.0*. <http://www.openarchives.org/OAI/openarchivesprotocol.html>
- Park, E. G., & Richard, M. (2011). Metadata assessment in e-theses and dissertations of Canadian institutional repositories. *The Electronic Library*, 29(3), 394–407.
- Park, J.-R. (2010). Metadata Quality in Digital Repositories: A Survey of the Current State of the Art. *Cataloging & Classification Quarterly*, 47(3), 213–228.
- Pavani, A., Suleman, H., McMillan, G., & Hickey, T. (2009). *ETD-MS v1.1: an Interoperability Metadata Standard for Electronic Theses and Dissertations*. Networked Digital Library of Theses and Dissertations. <https://ndltd.org/wp-content/uploads/2021/04/etd-ms-v1.1.html>
- Phiri, L. (2021). Automatic classification of digital objects for improved metadata quality of electronic theses and dissertations in institutional repositories. *International Journal of Metadata, Semantics*

and Ontologies, 14(3), 234–248.

Potvin, S., & Thompson, S. (2016). An Analysis of Evolving Metadata Influences, Standards, and Practices in Electronic Theses and Dissertations. *Library Resources & Technical Services*, 60(2), 99–114.

Suleman, H. (n.d.). *Global ETD Search*. Networked Digital Library of Theses and Dissertations. Retrieved December 26, 2022, from <http://search.ndltd.org>

Suleman, H. (2004). NDLTD Union Catalog Project. In *Electronic Theses and Dissertations* (1st Edition, pp. 89–92). CRC Press.

Suleman, H. (2012, September). The NDLTD Union Catalog: Issues at a Global Scale. *15th International Symposium on Electronic Theses and Dissertations*. <http://hdl.handle.net/10757/622568>

Webley, L., Chipeperewa, T., & Suleman, H. (2011). Creating a National Electronic Thesis and Dissertation Portal in South Africa. *14th International Symposium on Electronic Theses and Dissertations*. <http://pubs.cs.uct.ac.za/id/eprint/748/>