Original Article

Indian Journal of Library and Information Science
Volume 18 Number 1, January - April 2024
DOI: http://dx.doi.org/10.21088/ijlis.0973.9548.18124.3

Shaping the Smart Search Possible: An Inside Story of ETD Metadata

Shiva Kanaujia Shukla[1], Shipra Awasthi[2]

How to cite this article:
Shiva Kanaujia Shukla, Shipra Awasthi. Shaping the Smart Search Possible: An Inside Story of ETD Metadata. Indian J Lib Inf Sci 2024; 18(1):29–33.

Abstract

Metadata is very important in scientific literature. Metadata in electronic dissertations and dissertations (ETDs) requires a modular workflow based on a flexible architecture. Metadata methods must be simplified by adapting procedures for different document formats and structures. The structure of the dissertation document, including Fig.s, tables, footnotes, etc., requires greater attention to the identification and organization of metadata in the ETD for rich textual content. Reliable extraction of appropriate metadata elements, such as title, author, abstract, keywords, etc., is essential for future citation capabilities. Such mechanisms are crucial for evaluating processes and adjusting ETD metadata to facilitate architecture expansion. Most modern ETDs rely on the organization of quality text documents, ranging from raw PDF documents or semi-structured XML documents to uniform standard TEI (Text Encoding Initiative) documents. The present study discusses metadata creation and harvesting, various extraction methods and digital library search, metadata for MARC records in ETD, and access to ETD metadata in the context of the current scenario. It also highlights the tags adopted for the creation of the MARC records in JNU Central Library ETD System for simple-level learning.

Keywords: Electronic Thesis & Dissertation (ETD); Metadata; Digital Library; MARC Records; JNU Library.

INTRODUCTION

Electronic Theses & Dissertations are considered a rich source of information for scholars. In the past, such collections remained unused and were considered a piece to be stacked in the libraries. Despite being considered a valuable resource, the use of the collection is limited to a few scholars. The digital libraries were built to digitize the collection and maximize its use among the research community. The ETD collection is available through institutional repositories, national libraries, and archives to access freely anytime, anywhere (Fox, 2021).[1] The literature is significant in building new research and enriching the knowledge of the academic community in a particular area.

The myriad initiatives by universities and concerned scholars have come forward with innovative research to produce original, innovative,

Authors Affiliation: [1]Deputy Librarian, [2]Assistant Librarian, Jawaharlal Nehru University, New Delhi 110067, India.

Coressponding Author: Shipra Awasthi, Assistant Librarian, Jawaharlal Nehru University, New Delhi 110067, India.

E-mail: shipranit@gmail.com

Received on: 25.08.2023

Accepted on: 04.10.2023

and future-oriented outcomes. Bringing out new research and ensuring quality are a few significant research concerns. Avoiding duplication of research is possible with the proper use of ICTs and using the best research outcomes. The modern version of the electronic submission of the thesis is considered an Electronic Theses and Dissertations (ETDs). In the best exploration of previously developed ETD, the metadata structure and its utility in the access is very much sought out action.

ETD repositories benefit students and institutions by enhancing education, expanding research, and increasing a university's visibility and use, thereby contributing to the impact and ranking of its parent institutions (Ahmed, Alreyaee and Rahman, 2014).[2] Some organizations and projects have supported creating an ETD database using open source software such as D space and E prints. Electronic theses and dissertations enhance not only the visibility of the research but also the prestige of an organization.

Metadata Structure, Standards and Approaches

To facilitate wider access and reach to each minor component of ETD, the necessity of relevant keywords, domain specific vocabularies, and the role of authors is important. The development of a union catalog of ETDs is another example of enhancing the ETD reach among users. The use of standard language, controlled vocabulary, and metadata standards have also been associated with the metadata tagging of ETDs. Well, the tailored approaches tune into the metadata harvesting, and the transition from digital commons to the OCLC components has led to high quality records. (Banach, 2011; Potvin & Thompson, 2016; Han, *et al.* (2016); Lamba & Madhusudhan, 2018; Veve, 2016).[29-33] Surratt and Hill (2004)[34] described a semiautomated workflow for cataloging ETDs. A metadata query can be made using Perl script in an institutional database, and a MARC record is created for each ETD. The Connexion service imports the records into the Online Computer Library Center (OCLC) WorldCat database.

Metadata Generation and Harvesting

There are multiple benefits of ETDs compared with printed theses for certain reasons such as greater visibility of research and access for a wider audience. Other factors also contribute to the proliferation of ETDs; the preservation and promotion of access towards finding a few cost effective methods to create metadata. Access to unique research contributions is possible due to technological advances based on data standards. Metadata generation methods are observed in different natures such as automated and semi-automated approaches. These methods are utilized to harvest ETD metadata from libraries' institutional repositories (IR). Various metadata generation approaches include mechanisms based on: completely automated approaches or Pro Quest Services based semi-automated approaches. There are a few more examples of semi-automated approaches, some of which are based on highly technical tools or the Marc Edit OAI Harvester.

Metadata Extraction and Digital Library Searching

ETDs are rich in domain knowledge for digital library tasks. The other usages of ETDs can be found in the analysis of citation networks and the prediction of research trends. To create a scalable digital library search engine, automatic metadata extraction seems helpful. Various methods which support born digital documents (such as GROBID, CERMINE, and ParsCit) are usually observed to have limitations in extracting metadata from scanned documents. Traditional sequence tagging methods (based on text based features) and resource discovery include typographical and structural error improvements. It is observed that user supplied keywords may be useful for discovery. Still, theuse of controlled vocabulary can be found to fill the gap between the searcher's expectations and the author supplied vocabulary.

The policy of well cataloged theses/dissertations with the inclusion of "Library of Congress Subject Heading (LCSH) subject analysis and Library of Congress Classification (LCC)" can be useful in the case of ETDs for web based researchers. Also, text mining issues are related to extracting metadata. Few open source tools such as GROBID (Lopez and Romary, 2015)[3] may be useful for born digital information material but are helpless in the case of scanned documents usually related to ETDs. Extracting metadata from the scanned ETDs, it has been a practice to convert scanned pages into images and later text files with the application of OCR tools. This process captures patterns for seven metadata fields: titles, authors, years, degrees, academic programs, institutions, and advisors. The method is evaluated on a ground truth dataset comprised of rectified metadata.

Metadata for MARC Record in ETD

While academic libraries have responsibilities for access and usage stories depending on the

multiple options and mechanisms for access management, the metadata input emerges. Implications for digital libraries and customized mapping in context with metadata for ETDs have faced many models. Software and best practices (Vijaya kumar, Murthy & Khan, 2006; Deng & Reese, 2009; Lubas, 2009; Yañez, 2009).[4,5,6,7] The metadata generation, selection arrangements, etc. Are requisites for metadata workflow. The examples include "repurposing Pro Quest metadata for batch ingesting ETDs" (Averkamp & Lee, 2009)[8] whereas author supplied metadata, OAI-PMH, and XSLT to Catalog ETDs are few other significant aspects (Boock & Kunda, 2009; Robinson, Edmunds & Mattes, 2016).[9,10]

Access to ETD Metadata

The basic questions related to electronic theses and dissertation (ETD) repositories are about their structure, development, and the extent to which their capability is assumed in university research management (McMillan, 2002; Yiotis, 2008).[20,21] The process and examples include the development of a heuristic baseline method for metadata extraction in the case of scanned electronic theses and dissertations (Choudhury *et al.*, 2020).[22] While metadata harvesting and extraction are other areas of routine concern (Veve, 2016; Choudhury *et al.*, 2021).[23,24]

The university libraries have observed dramatic changes in the growth and development of ETDs, yet building a distributed and heterogeneous system as well as (ETD metadata remediation has been in demand (Zhao & Jiang, 2004; Thompson *et al.*, 2019).[25,26] The formal models of ETDs have taken various aspects into account; several disciplines, structures, digital spaces, etc. are a few of them (Gonçalves *et al.*, 2004).[27] The proliferation of ETDs in various areas has been experimented with. The technology development and models provide ways for implementation. Example sinclude publishing music related ETDs (Yang *et al.*, 2016).[28]

Selection of Metadata in JNU Central Library

Central Library, JNU has selected the tags available in the virtua and koha software to create MARC records, as several tags are available for each record. The tags used in virtua software are 100 author information, 245 title, 260 for place, organization and year, 502 - type of document, 720 - name of the supervisor, and 856 - URL of the Pdf file (Fig. 1 and 2).

Central Library has used the following tags to create MARC records in the Koha software: 100 - author, 245 - title, 502 - type of document, 700 - name of the supervisor, 710 - details of the department, 856 - URL of the Pdf file, 942 - theses/dissertations and name of the classification scheme (Fig. 3 and 4).

Developments in ETD

The beginning of the new millennium witnessed the development of a Networked Digital Library of Theses and Dissertations (Suleman *et al.*, 2001).[11] Jin (2004)[12] also brought the visions of access through the smart development of ETDs (Mikeal *et al.*, 2007)[13] shortly. The digital library perspectives encompassed certain components such as *Content* dm as well as OCLC Systems (Howard & Goldberg, 2011)[14] for enhanced access to ETDs. In the same year, almost a decade ago the benefits and ETDs publishing through the open access repository was brought into the process.

The development of ETDs continued in the last decade with the proliferation of technology and the intimidating need for research

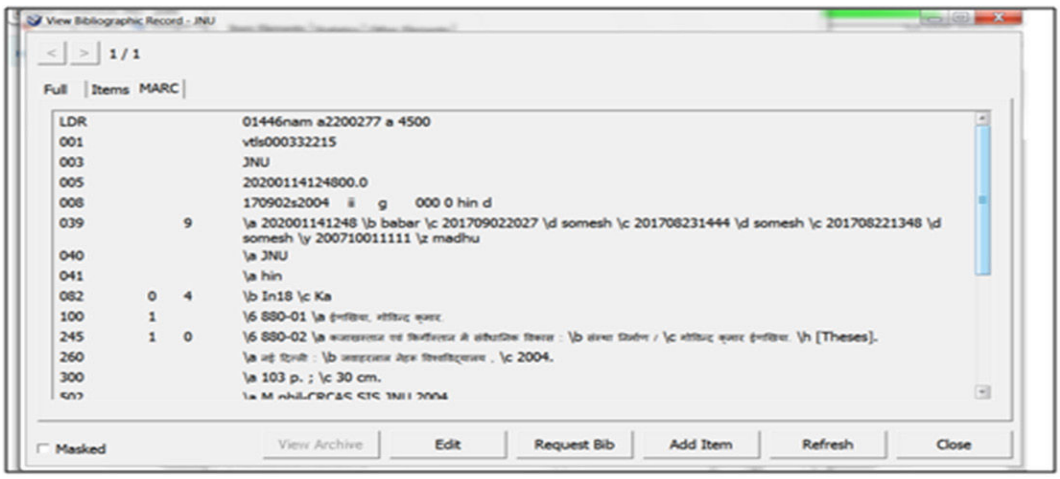

Fig. 1

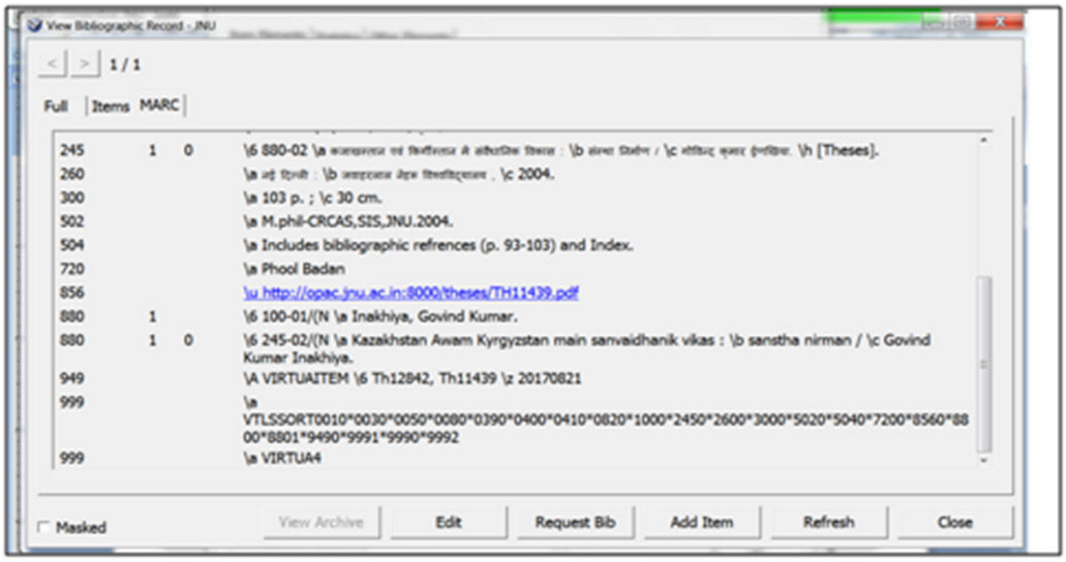

Fig. 2

transparency.Creating digital libraries for better management of ETDs (Fox, 2021)[1] also included the common submission system for ETDs. The ETDs development and management always envisioned using metadata, which required regular assessment and formulation of policies for a better service (Ghosh, 2009; Park & Richard, 2011; McCutcheon, 2011).[15,16,17] The simultaneous development of institutional repositories and digital libraries has opened new vistas in the shape of Open access which had its own issues and challenges. The quality of the thesis in the context of visibility, readability, and access were also concerns of the usual collection processes of libraries; ETDs have also observed the open asccess movements from the perspective of various issues and challenges for Research and Development (Khaparde & Ambedkar, 2014; Gunjal & Gaitanou, 2015).[18,19]

CONCLUSION

The archiving of electronic documents is the latest trend due to long-term retention. ETDs repositories play a vital role in managing scholarly output at the university level and assist scholars in accessing it anytime, anywhere. The *Shodhganga* national repository presently contains 411187 full-text theses which facilitate the academic community in pursuing their scholarly projects. With the change in scenario, the role of library professionals is also changing in submitting the ETDs and creating metadata processes.

REFERENCES

1. Fox E A. Building and using digital libraries for ETDs. The Journal of Electronic Theses and Dissertations. 2021;1(1):5.doi: https://doi.org/10.52407/LJQF5826.
2. Ahmed A, Alreyaee S, Rahman A. Theses and dissertations in institutional repositories: an Asian perspective. New Library World. 2014; 115(9/10):438-451.
3. Lopez P, Romary L. GROBID - Information Extraction from Scientific Publications. Research and Innovation. 2015. 02 January 2015. doi: https://ercim-news.ercim.eu/en100/r-i/grobid-information-extraction-from-scientific-publications.
4. Vijayakumar J K, Murthy T A V, Khan M T M. Experimenting with a model digital library of ETDs for Indian universities using D-space. Library Philosophy and Practice.2006;9(1).
5. Deng S, Reese T. Customized mapping and metadata transfer from D Space to OCLC to improve ETD workflow. New Library World. 2009. doi:10.1108/NLW-04-2014-0035.
6. Lubas R L. Defining best practices in electronic thesis and dissertation metadata. Journal of Library Metadata. 2009;9(3-4): 252-263.
7. Yañez I. Metadata: Implications for academic libraries. Library Philosophy and Practice. 2009:1-8.
8. Averkamp S, Lee J. Repurposing Pro Quest metadata for batch ingesting ETDs into an institutional repository. In: Mc Intosh, Joyce (Ed). Cataloging and Indexing: challenges and solutions. Apple Academic, USA:2009.
9. Boock M, Kunda S. Electronic thesis and dissertation metadata workflow at Oregon State

University Libraries. Cataloging & Classification Quarterly.2009;47(3-4): 297-308.

10. Robinson K, Edmunds J, Mattes S C. Leveraging author-supplied metadata, OAI-PMH, and XSLT to Catalog ETDs: A case study at a large research librarys. Library Resources & Technical Services. (2016);60(3):191-203.
11. Suleman H, Atkins A, Gonçalves M A, France R K, Fox E A, Chachra V., Young J. Networked digital library of theses and dissertations. D-lib Magazine.2001;7(9):1082-9873.
12. Jin Y. The development of the China networked digital library of theses and dissertations. Online Information Review.2004.
13. Mikeal A, Brace T, Leggett J, McFarland M, Phillips S. Developing a common submission system for ETDs in the Texas Digital Library. 2007.doi:https:// oaktrust. library.tamu.edu/handle/1969.1/5679.
14. Howard R I, Goldberg T. Facilitating greater access to ETDs through CONTENT dm. OCLC Systems & Services: International digital library perspectives. 2011;27 (2).
15. Ghosh M. E-theses and Indian academia: A case study of nine ETD digital libraries and formulation of policies for a national service. The International Information & Library Review. 2009;41(1): 21-33.
16. Park E G, Richard M. Metadata assessment in etheses and dissertations of Canadian institutional repositories. The Electronic Library. 2011.
17. Mc Cutcheon S. Basic, fuller, fullest: Treatment options for electronic theses and dissertations. Library Collections, Acquisitions, and Technical Services. 2011;35(2-3): 64-68.
18. Khaparde V, Ambedkar B. Growth and development of electronic theses and dissertation (ETDs) in India. Journal of Library and Information Sciences.2014;2(1): 99-116.
19. Gunjal B, Gaitanou P. (2015). ETDs and Open Access for Research and Development: Issues and challenges. In: 18th International Symposium on Electronic Theses and Dissertations Evolving Genre of ETDs for Knowledge Discovery (ETD 2015 India), New Delhi. 2015. pp.136-143.
20. McMillan G. ETDs and Libraries. Virginia Polytechnic Institute and State University, Virginia; 2002.
21. Yiotis K. Electronic theses and dissertation (ETD) repositories: What are they? Where do they come from? How do they work? OCLC Systems & Services: International digital library perspectives.2008.
22. Choudhury M H, Wu J, Ingram W A, Fox E A. (2020, August). A heuristic baseline method for metadata extraction from scanned electronic theses and dissertations. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries. 2020. pp. 515-516.
23. Veve M. From Digital Commons to OCLC: a tailored approach for harvesting and transforming ETD metadata into high-quality records. Code 4 Lib Journal.2016; (33).
24. Choudhury M H, Jayanetti H R, Wu J, Ingram W A, Fox E A. (2021). Automatic metadata extraction incorporating visual features from scanned electronic theses and dissertations. In: ACM/IEEE Joint Conference on Digital Libraries (JCDL). IEEE; 2021. pp. 230-233.
25. Zhao Y, Jiang A. (2004). Building a distributed heterogeneous CALIS-ETD digital library. In: International Conference on Asian Digital Libraries. Springer, Berlin, Heidelberg; 2021. pp. 155-164.. https://doi.org/10.1007/978-3-540-30544-6_16.
26. Thompson S, Liu X, Duran A, Washington A. A case study of ETD metadata remediation at the University of Houston libraries. Library resources & technical services. 2019; 63(1).
27. Gonçalves M A, Fox E A, Watson L T, Kipp N A. Streams, structures, spaces, scenarios, societies (5s) A formal model for digital libraries. ACM transactions on information systems (TOIS).2004;22(2);270-312.
28. Yang L, Ketner K, Luker S, Patterson M. A complete system for publishing music-related ETDs: Technology development and publishing model. Library Hi Tech. 2016; 34 (1).
29. Banach MGHAN. The benefits of managing and publishing ETDs" in house" using an open access repository. USETDA 201. Available at: http:// works.bepress.com/meghan_banach/4/.
30. Potvin S, Thompson S. An analysis of evolving metadata influences, standards, and practices in electronic theses and dissertations. 2016. doi: http://hdl. handle. net/10657/1341.
31. Han MJK, Harrington P, Black A, Kudeki D. Aligning author-supplied keywords for ETDS with domain-specific controlled vocabularies. 2016.
32. Lamba M, Madhusudhan M. (2018). Metadata tagging of library and information science theses: Shodhganga (2013–2017). In: ETD 2018: Beyond the Boundaries of Rims and Oceans Globalizing Knowledge with ETDs, Taiwan; 2018.
33. Veve M. Harvesting ETD metadata from institutional repositories to OCLC: Approaches and barriers to implementation. Journal of Library Metadata.2016;16(2): 69-79.
34. Surratt B E, Hill D. ETD2MARC: A semiautomated workflow for cataloging electronic theses and dissertations. Library Collections, Acquisitions, and Technical Services. 2004;28(2): 205-223.