Building A Large Collection of Multi-domain Electronic Theses and Dissertations

Sami Uddin[*], Bipasha Banerjee[†], Jian Wu[*], William A. Ingram[‡], Edward A. Fox[†]
[*]*Computer Science, Old Dominion University, Norfolk, VA, United States*
[†]*Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA, United States*
[‡]*University Libraries, Virginia Polytechnic Institute and State University, Blacksburg, VA, United States*
{muddi004,j1wu}@odu.edu, {bipashabanerjee,fox,waingram}@vt.edu

***Abstract*—In this work, we report our progress on building a collection containing over 450k Electronic Theses and Dissertations (ETDs), including full-text and metadata. Our goal is to close the gap of accessibility between long text and short text documents, and to create a new research opportunity for the scholarly community. For that, we developed an ETD Ingestion Framework (EIF) that automatically harvests metadata and PDFs of ETDs from university libraries. We faced multiple challenges and learned many lessons during the process, that led to proposed solutions to overcome/mitigate the limitations of the current data. We also described the data that we have collected. We hope our methods will be useful for building similar collections from university libraries and that the data can be used for research and education.**

***Index Terms*—ETD, OAI-PMH, Big data**

I. Introduction

In the past decade, there has been increasing interest in studying a vast volume of scholarly articles, usually referred to as scholarly big data [7]. Many academic big datasets emerged and became available for researchers, such as CiteSeerX [5], PubMed, DBLP, Semantic Scholarly Open Research Corpus (S2ORC) [4], and its subset CORD-19. Most documents in these datasets are papers published in journals or conferences. One understudied type of scholarly document is electronic theses and dissertations (ETDs). We define ETDs as written documents that describe the graduate student's research, usually as partial fulfillment towards a degree. Fig. 1 shows the number of doctoral degrees conferred each year in the United States from 1945 to 2019, which reflects the growth of ETDs of all types.

Existing ETD collections include the Networked Digital Library of Theses and Dissertations (NDLTD [2]), containing 6+ million records worldwide, and ProQuest Dissertation & Theses Global (PDTG), a subscription-based portal indexing 5+ million ETDs. Neither of these allows full-text access. In this paper, we describe our effort to build a collection containing 450,000 US ETDs, as part of research on a digital library paving the path for analyzing book-length documents. The crawling took advantage of the Open Archives Initiative Protocol (OAI-PMH) provided by university libraries. We propose an ETD ingestion framework (EIF), which ingest metadata into a MySQL database and PDFs into a repository. We highlight the challenges faced and lessons learned in the process and then describe the data we collected.

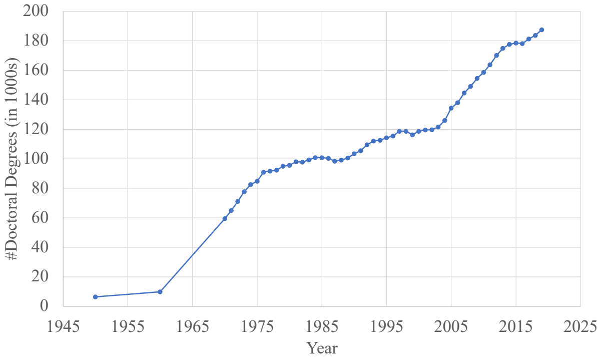

Fig. 1: The number of doctoral degrees earned in the United States from 1950 to 2019. Data was from statista.com.

II. Harvesting ETDs and metadata

A. Focused Web Crawling

We followed two ways to harvest ETDs and scrape their metadata. The first is to start from sitemaps. The sitemap files can usually be found inside the robots.txt file placed directly under the document root of a website. The robots.txt file was proposed to tell bots/crawlers (e.g., googlebot) what is allowed to do and what is not. It also defines how much delay a bot needs to obey between each hit.

After finding the sitemap and collecting the URLs, we start going over them one by one, following the crawl-delay. One problem is that many URLs on the library website do not point to an ETD landing page. Instead, they point to other types of PDFs, such as regular papers or schedules. Therefore, we used methods such as reading the breadcrumbs (e.g., SMARTech Home/Georgia Tech Theses and Dissertations/View Item) on top of the webpage or checking the document type (if available) from the webpage before scraping.

To track the sitemap and go through every URL available is time-consuming. For example, when we downloaded ETDs for the Georgia Institute of Technology (GTech), there are more than 60,000 URLs in the sitemap of the GTech library. Among which, around 22,000 URLs were pointing to ETDs. To avoid getting blocked by the server, we used a 10-second delay between two requests, so it would take at least 166 hours or nearly a week to collect all ETDs from this repository.

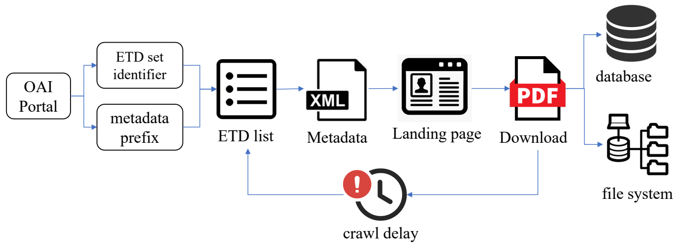

Fig. 2: Crawl pipeline.

One way to reduce the crawling time is to use OAI-PMH; see Fig. 2. Many university libraries use DSpace, an open-source repository software package, which supports OAI-PMH. In this approach, the first step is to find the DSpace data provider link for a university's digital repository. A DSpace repository of a university contains various sets of records including but not limited to theses, reports, and newsletters. These records can be identified by several metadata prefixes in OAI-PMH records such as *dim*, *oai_dc*, *etdms*, etc. After finding the sets for theses and dissertations and detecting the metadata format which contains the most detailed information (*dim* for our case), the metadata in XML format has URLs linking to ETD landing pages from the XML. Our crawler visits the URLs to download the PDF files for each ETD's PDFs. The crawler obeys the crawl-delay specified in the robots.txt file. We use a lightweight OAI-PMH client library named Sickle, written in Python.

B. Developing Database and Repository

We organize our ETD collection using a MySQL database and a local repository. The database is used for storing metadata while the repository is used for storing PDF and XML files. The database includes a main table containing metadata of all ETDs. We create other tables that link to the main table and host extended metadata or data derived from the ETDs (Fig. 3). For example, the *subjects* table contains subject terms from the library provided metadata and the *figure_tables* table contains properties of tables and figures extracted from ETDs. A framework has been developed to extract figures and tables from ETDs [3].

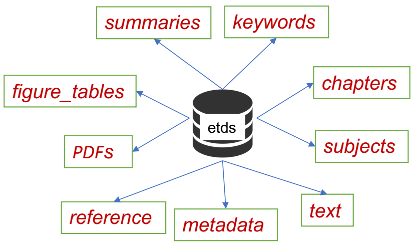

Fig. 3: Tables associated with our ETD records.

We adopt a hierarchical repository structure as in CiteSeerX [6]; see Fig. 4. Each ETD has a unique ID, which maps to its path in the repository. Each leaf folder contains the PDFs and an XML of an ETD.

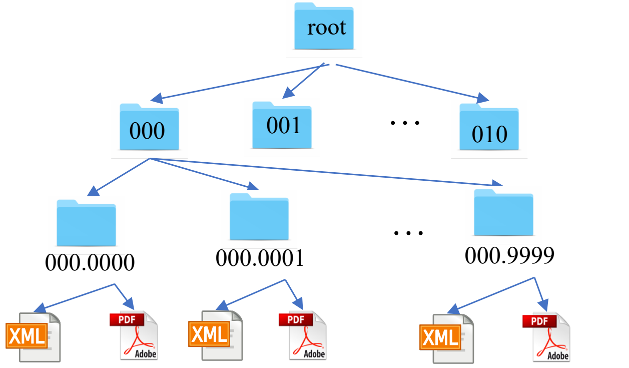

Fig. 4: The hierarchical structure of the ETD repository. Each first-level directory contains 10,000 sub-directories.

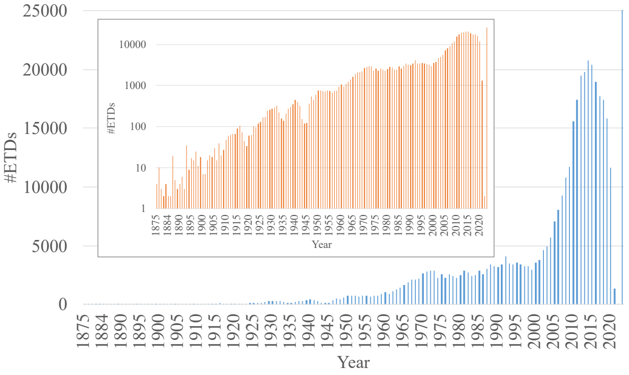

Fig. 5: The distribution of collected ETDs over years as of mid-November 2021. Many ETDs do not have dates in library provided metadata; see the long bar on the right. The inset shows the same distribution with y-axis in logarithmic scale.

C. Properties of the ETD Collection

We have crawled from over 42 universities. The number of ETDs available from these universities ranges from 3000 to 50,000+. In some cases, one repository may host ETDs from different institutions, e.g., OhioLink, which hosts ETDs from Ohio institutions. The number of ETDs collected from top universities is shown in Table 1. The total size of the repository is 3.4 terabytes. The ETD collection is hosted by Old Dominion Computer Science and mirrored in Virginia Tech's University Libraries. The distribution over years (Fig. 5) reveals a prominent increase after around 1945 and a surge after 1997, which is consistent with the time when many universities started adopting ETDs.

Based on the harvested metadata, the ETD dataset contains 2000+ department names (before resolving near duplicates) and at least 300,000+ pairs of advisor-student information. The proportions of doctoral dissertations, master's theses, and bachelor's theses are 56%, 42%, and 2%, respectively. We have 451358 records in our database.

TABLE I: ETDs from top 10 universities in our repository.

University	Number of PDF
The Ohio State University	55780
Virginia Polytechnic Institute and State University	29597
Georgia Institute of Technology	22400
Texas Tech University	21702
Kansas State University	19299
The University of Texas at Austin	18283
Oklahoma State University	17746
North Carolina State University	15365
University of Illinois at Urbana-Champaign	14281
Rice University	13151
Others	223754

III. Challenges and Lessons

Challenges we faced when building the collection include:

1) **Crawl-delay:** Even when we followed the crawl-delay specified in `robots.txt`, our request sometimes was blocked. It was often necessary, through trial-and-error, to find the best delay between two consecutive requests.
2) **Embargoed ETDs:** Unfortunately, it was usually impossible to find, from the OAI-PMH metadata or on the landing pages, which ETDs are embargoed, until we hit the links. Empirically, a large fraction of embargoed ETDs were published in recent years. Also, many embargoed ETDs are only allowed for member access at that particular library and disallow public access. It is a common scenario to not be able to download every PDF for each metadata record available.
3) **Non-uniform DOM structures:** The HTML DOM structure varies across university repositories, which required us to customize HTML parsers to extract the target metadata fields.
4) **Inconsistent and incomplete metadata:** A significant fraction of ETD metadata fields have missing data; the most common ones are department, discipline, and subjects. Available fields, such as "year issued", may have inconsistent metadata formats across different university repositories, such as "mm-dd-yyyy" or "yyyy-mm-dd". Even within the same repository, metadata fields may have inconsistent values. Common cases include using synonyms (e.g., "jhu" vs. "Johns Hopkins University") and extra spaces (e.g., "Texas A&M University" vs. "Texas A & M University").
5) **Other issues:** One university banned any crawlers to visit URLs containing /oai, specified in the robots.txt file, so no ETDs could be crawled from that site. We also encountered a case where an API call returned multiple records, not just one.

As mentioned above, one problem that needs to be solved is that many repositories have incomplete and/or inconsistent metadata. As shown in Table II, data was missing in many fields. To mitigate the challenges of obtaining complete and consistent metadata, we developed a metadata extraction framework [1], trained on a set of human annotated ETDs. It considers both textual and visual features. This framework achieves F1 of 81%–97% for seven key metadata fields, extracted from ETD cover pages. We will apply it to complete/improve the metadata of ETDs.

We are working on improving metadata quality and building a web interface to make the dataset more accessible and usable. The dataset will also facilitate training and improving existing language models for scholarly documents.

TABLE II: Missing fields across the database

Field	Missing count
Year	65,955
Advisor	112,748
Deparment	232,653
Discipline	166,690
Subjects	52,579
Abstract	99,583

To deal with a stalled crawler, we set a timeout after which the crawler is restarted. To tackle the problem of unavailable PDF, we implemented a filter to download only available PDFs. We skipped the fields that did not have values while ingesting them into the database. For fields which have combined data into one field, we developed a robust checker to handle that.

IV. Conclusions

We gathered around 450,000 ETDs by crawling from over 42 universities in the United States. The collection covers a wide spectrum of academic domains and a large time range, from 1875 to the present. We currently have not made the data publicly available, but we have built a search interface based on the collected data. We expect this web interface will make the dataset accessible and usable for researchers and other users, and hope to hear from those with additional needs.

Support was made in part by the Institute of Museum and Library Services through grant LG-37-19-0078-198.

References

[1] M. H. Choudhury, H. R. Jayanetti, J. Wu, W. A. Ingram, and E. A. Fox. Automatic metadata extraction incorporating visual features from scanned electronic theses and dissertations. *CoRR*, abs/2107.00516, 2021.

[2] E. A. Fox, M. A. Gonçalves, G. McMillan, J. Eaton, A. Atkins, and N. Kipp. The Networked Digital Library of Theses and Dissertations: Changes in the university community. *Journal of Computing in Higher Education*, 13(2):102–124, 2002. https://doi.org/10.1007/BF02940968.

[3] S. Y. Kahu, W. A. Ingram, E. A. Fox, and J. Wu. Scanbank: A benchmark dataset for figure extraction from scanned electronic theses and dissertations. *CoRR*, abs/2106.15320, 2021.

[4] K. Lo, L. L. Wang, M. Neumann, R. Kinney, and D. Weld. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online, July 2020. Association for Computational Linguistics.

[5] J. Wu, B. Kandimalla, S. Rohatgi, A. Sefid, J. Mao, and C. L. Giles. Citeseerx-2018: A cleansed multidisciplinary scholarly big dataset. In *IEEE International Conference on Big Data, Big Data 2018, Seattle, WA, USA, December 10-13, 2018*, pages 5465–5467, 2018.

[6] Z. Wu, J. Wu, M. Khabsa, K. Williams, H.-H. Chen, W. Huang, S. Tuarob, S. R. Choudhury, A. Ororbia, P. Mitra, and et al. Towards building a scholarly big data platform: Challenges, lessons and opportunities. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL 2014, pages 117–126. IEEE Press, 2014.

[7] F. Xia, W. Wang, T. M. Bekele, and H. Liu. Big scholarly data: A survey. *IEEE Trans. Big Data*, 3(1):18–35, 2017.