Extracting and Registering References to Improve Scholarly Impact of ETDs

Background

Citation networks are a cornerstone of modern scholarly communication. They provide readers with access to other articles and publications that may be related to an article or paper they are interested in using. Many journal publishers provide easy access to citation information for an article – both the sources cited within the article and information about how often the article is cited by others.

Citation networks have been an important part of scholarly communication since they first began to appear in the 1960s , including the Science Citation Index in 1955 , which became a core part of the "Web of Science" in the online environment. Since these networks first appeared, they have grown in size and complexity as the value has been proven. In the past twenty years there have been several simultaneous innovations that have moved the citation networks to what they are now. First the development of the internet as the primary avenue for disseminating information allowed for direct linking rather than textual indexing. The development of unique identifier strategies – such as the Digital Object Identifier (DOI) – improved methods of linking to reference sources unambiguously. Built on top of these strategies, systems like Crossref connect resources by registering DOIs and also aggregating associated metadata about scholarly output .

The Crossref system, specifically, supports the inclusion of supplemental information associated with a resource DOI, such as the resources that are internally referenced by the scholarly resource. When these citation references are aggregated across a large collection of resources, a network of connections is created. These networks can even reflect citations for resources that do not have a DOI – either because they are beyond the scope of that identifier system or because they have not been registered. For this reason, Crossref encourages its members to register the references contained within a resource as part of the metadata database and provides additional functionality when references are present. Between June 2020 and June 2021, there was a 10% growth in the number of references registered and a 21% growth in the number of publishers registering references in Crossref . Reference information is also aggregated by other organizations such as the Open Citation project and OpenAlex , which use the Crossref dataset as one of the sources for building open citation networks.

Most major publishers register their journal articles, book chapters, and conference papers with Crossref as part of the DOI creation process. Currently Crossref contains more than 162 million registered DOIs and these three types account for over 86% of the total records, particularly representing scholarly literature from across the science and humanities disciplines. Conversely Electronic Theses and Dissertations (ETDs) are not well represented among registered DOIs. Currently, there are over 600,000 theses and dissertations registered with Crossref, but this only represents a fraction of the globally-published ETDs. One recent publication found that ETDs are the second-most-frequent resource type among open-access publications available via OpenDOAR . While there is no exact count of the total number of analog and born-digital ETDs worldwide, there are several large aggregations of ETDs that can be used to gauge the total number. These aggregations include the National Digital Library of Theses and Dissertation (NDLTD), which operates the "Global ETD Search" system containing over 6.5 million records in its database ; the "EBSCO Open Dissertations" database containing over 1.5 million ETDs ; and the "Dissertations & Theses" database from ProQuest containing over 5 million records . With these large aggregations in mind, the ETDs registered with Crossref represent around 10% of the ETDs in the largest aggregation.

The curation of ETDs at colleges and universities often falls to the libraries at those institutions. Many libraries have robust workflows and systems to make them available and discoverable by the public. Over time the collective output of ETDs represents a considerable amount of information about an extremely wide range of subjects. Representing this scholarship within global citation networks provides a major opportunity to extend visibility for ETDs and expanding the scope of citation networks to represent the broadest content and scholarly output. Unfortunately, implementing any of these changes requires additional time and effort, as well as the challenge of adjusting existing workflows.

The first challenge is identifying and extracting references within ETDs. A wide range of work has been done to facilitate the programmatic processing and extraction of information from ETDs, such as the development of test beds and datasets at Virginia Tech in the United States to try out new approaches for information extraction from ETDs. An important first step in this process is identifying references and there are current efforts to extract references from a wide range of publications – including academic articles, legal text and patents – with tools based on modern neural networks and large language models . A tool that is often cited as a baseline for identifying references in scholarly publication is the GROBID tool . GROBID is a machine learning library for extracting, parsing and re-structuring raw documents such as PDFs into structured XML/TEI encoded documents with a particular focus on technical and scientific publications. It has been used extensively to identify references in a wide variety of scholarly resources including ETDs.

Another challenge within the corpus of theses and dissertations is that some documents are digitized from analog sources and the text is processed with Optical Character Recognition (OCR). Digitized text may further complicate the identification and extraction of references due to varying quality of source text and potential errors in accuracy of the recognized text from the OCR process . Variations in reference layout and formatting can also be challenging; although a university may have general style guidelines that result in common patterns of citation components, these may change over time or be inconsistently applied, hindering programmatic extraction.

Objectives

The major objective of this work is to increase the visibility of the ETDs written by students at the University of North Texas (UNT). All ETDs from UNT students are loaded into the UNT Digital Library, which is the permanent digital repository for all scholarly work at the university. These are accessible through public-facing collection available here: https://digital.library.unt.edu/explore/collections/UNTETD/. The UNT Digital Library has a locally-developed technical infrastructure and includes some features to make items findable for users. For example, metadata includes mark-up that exposes it for easier crawling by search engines and many users are directed to items in the UNT Digital Library directly from internet search results. Additionally, the system tracks individual item views and aggregates this at the collection level to show how often ETDs are being used, but it cannot provide information beyond those interactions within the system.

As discussed in the previous section, adding references in the Crossref database when registering ETDs from UNT students will incorporate those ETD resources to existing citation networks; it will also position them as source nodes within larger bibliographic networks such as OpenCitations and OpenAlex , which use the Crossref dataset as one of the data sources. Registering references with DOIs provides local benefits in addition to the data collected in the repository, such as better citation count reporting for formal publications (i.e., how often the ETDs are referenced by other registered scholarly works).

At this time Crossref generally does not include citation counts from ETDs because the references are not present in the dataset, representing a gap in available data. Crossref is a continuously-changing database of bibliographic metadata associated with the unique DOIs registered for different types of scholarly output. As of the writing of this paper, there are over 162,278,000 DOIs and associated metadata records registered with the system. These DOIs represent 30 different types including: Journal Articles (109,358,569), Book Chapters (22,074,321), Conference Paper (8,639,158), Datasets (2,995,681), Monographs (694,288), and Dissertations (688,324). It should be noted that the "Dissertation" type is used for both theses and dissertations.

Even among the most-represented types, only some of the records currently include associated references, e.g.: Conference Papers 5,214,697 (60.36%); Journal Articles 57,120,210 (52.23%); and Book Chapters 7,042,309 (31.90%). For the Dissertation type just 1,604 (0.23%) of registered records contain references and these are provided by just

23 of the 283 publishing institutions registering ETDs with Crossref (see Table 1) including one designated as "Test accounts."

Publishers of ETDs that have references registered in Crossref.

Publisher	Items
Riga Stradins University	630
University of St. Augustine for Health Sciences Library	247
Universidade Federal de Juiz de Fora	247
Vilnius Gediminas Technical University	92
Kryvyi Rih State Pedagogical University	89
University of Denver, University Libraries	73
Linnaeus University	72
Students Journal of Health Research Africa	38
Technische Hochschule Wildau	33
University of Szeged	25
Mississippi State University Libraries	15
Byte Systems - Solucoes Digitais	11
Universitatsbibliothek Kiel	6
Corvinus University of Budapest	5
SJC Publisher Company Limited	4
Test accounts	3
Banco de Mexico	3
Society of Psychoceramics	2
Cifra Ltd - Russian Agency for Digital Standardization (RADS)	2
Universitatsbibliothek Bamberg	1
Stavanger University Library	1
Elisava Barcelona School of Design and Engineering	1

To advance the work in addressing the scholarly citation gap, this paper presents efforts by the UNT Libraries to build reference lists for ETDs published by the university for two main purposes. First, these lists provide references to Crossref to support additional visibility and functionality. Second, they provide a building block for local collection analysis regarding high-level usage of library resources purchased for our students and faculty.

Methods

The overall process to enhance available citation data for UNT-published ETDs involves identifying references, extracting the reference text, and adding these references to the metadata for each ETD DOI registered with Crossref.

We decided to create a human-curated workflow to extract text references from ETDs. This approach was chosen instead of a computational approach that would require programming and software development due to its simplicity and potential replicability by other institutions. While there are examples of tools for reference identification and parsing mentioned earlier in this paper, many of these would still require manual interventions or review due to the complexity of the document formats for this project. We did not want this to be a technology-focused activity but instead focused on the end-product of usable reference lists. To that end, we implemented the following workflow for the manual extraction of these references.

The first step was to create a set of instructions with examples for a small group of student assistants. These instructions describe how to find and handle various reference styles used by different disciplines at UNT as well as how to format the resulting text-based citations for use later in the workflow. Once the instructions were developed, a group of four students were provided PDF versions of ETDs for the 2023 calendar year to start copying references from individual documents.

All of the ETDs for this project were already uploaded to the UNT Digital Library, which uses the Archival Resource Key (ARK) as a unique identifier. An ARK always includes an institutional prefix followed by an "Assigned Name" for an individual item. The Assigned Name (which is unique to an item) is also appended as a suffix in the DOI for the same ETD when it is registered with Crossref. For example, the ARK identifier for an ETD in the UNT Digital Library is `ark:/67531/metadc2137535`, (the Assigned Name is "metadc2137535"), so the URL to access the ETD's landing page is https://digital.library.unt.edu/ark:/67531/metadc2137535/ and the DOI for this ETD is `10.12794/metadc2137535`. In this workflow, the Assigned Name portion of the ARK was also used as the filename for the downloaded PDF files (e.g., metadc2137535.pdf). Using the same unique identifier in multiple ways provides continuity and prevents confusion in mapping or cross-identifying information related to the same item.

A total of 412 ETDs published in 2023 were downloaded as PDFs and named according to this process so that students could refer back to the online version as needed. Once the files were available, students would select an individual ETD PDF and begin to identify the reference sections within the document. These sections are copied from the PDF by highlighting the references and using the copy/paste commands to transfer references into a text document via a common text editor such as Notepad++ or Visual Studio Code. The text document is also named using the Assigned Name, so an ETD PDF named `metadc2137535.pdf` will have an accompanying text file named `metadc2137535.txt` containing the references from the ETD. Once the references are copied into the text file, they are reformatted so that there are no line breaks within a single reference and each reference is followed by a single blank line, to assist in visually separating the references. During this process, students perform simple proofreading to ensure that there weren't any unexpected formatting errors introduced when copying the references (such as missing spaces between characters). Students are instructed not to modify or correct the references when they notice that the authors did not follow the citation format closely enough.

Once this basic reformatting has been completed, staff performed a normalization step on each of the references in the file. A first pass converts characters that may be represented in different ways into a standardized format by using the Unicode's Normalization Form KC (NFKC) – i.e., Compatibility Decomposition, followed by Canonical Composition on all characters in the reference . Next, we replace different forms of single and double quotes, sometimes called "curly quotes," with straight quote characters and convert different types of dashes (em-dashes, en-dashes, and 3-em-dashes) into hyphens. We then perform a series of simple string formatting operations to correct common typographic errors we encounter, such as missing spaces or extra spaces between components that do not need those spaces. Finally, we try to normalize several of the variant formats of URLs and DOIs that occur in the references. This normalized text file is added to a public Git repository on Github.com (https://github.com/unt-libraries/digital-collections-item-references/) where it can be used as a starting point for other analysis and data mining.

The next step in the process is to convert the text-based reference list into the XML format defined by Crossref for registering references. This conversion was accomplished with a locally-developed python script that used the formatted reference list as the input and exported the required XML.

References can be added at the time that a DOI is created or they can be added and updated after the DOI is already registered. We chose to keep these steps separate and register the DOIs first as it provided different options for workflows. Each reference submitted to Crossref is required to have a unique identifier or key; our identifier keys use the Assigned Name plus a three-digit, zero-padded number based on the sequence that the reference occurred in the ETD. For example, the tenth reference in the previous example ETD would have the identifier key of `metadc2137535-010`.

There are also several options for communicating references when registering them using XML tags to clarify how the information is organized. When the reference includes a known DOI, you can submit that identifier by itself to establish the reference linkage. For other references, the choice is to parse and format individual references by dividing them into component elements – e.g., title, authors, journal, volume, issue, pages – or to submit each reference in an unstructured, unformatted text string. During the pilot work described in this paper, we chose not to parse citations into component elements and made use of the unstructured text strings for loading references. We made use of the web-based depositor form for submitting the final XML metadata documents needed to register references within the Crossref system. In future work we plan to make use of their POST API to pragmatically register the references for each of our ETDs once they are submitted and processed.

Results

In 2023, 412 ETDs were added to the UNT Digital Library; however, two documents submitted as Master's Theses in the field of creative writing did not contain any references. Of the remaining 410, there were a 49,150 total references in all of the documents combined. Across all 2023 ETDs, the minimum number of references is 0 and the maximum number of references in a single ETD is 653 (*M* = 119.296, *SD* = 99.5803).

During the process of extracting and formatting references from the ETDs, a number of unexpected situations required additional review. Originally, the process was planned around two standard reference section formats: alphabetically-ordered references and numerically-ordered references. For alphabetical lists, references have no additional formatting changes aside from blank lines between citations and the normalizations described previously. For numerically-ordered references, there were too many variations in numbering formats to clearly identify necessary information, e.g.: a number in brackets, a number with a trailing period, a number with no punctuation, a number with a trailing period within brackets, etc. To clearly delineate these different types of reference list styles, we decided to enclose all enumeration in square brackets, followed by a space and then the remainder of the original reference string. This allowed us to keep most of the reference as it appears in the document but also have consistency for later parsing of these references.

A second unexpected obstacle related to the variation in reference sections and their locations. Recently, the University of North Texas has expanded allowable formatting to include ETDs based on a combination of several previously-published articles, often presented as a collection of articles. ETDs using this format will have separate reference sections at the end of each chapter (i.e., previously-published article or revised version of a previously-published article) rather than a single list of references at the end of the document. Additionally, since the individual articles in these compiled documents may have been published in different outlets, each chapter may have references formatted according to a different publisher's requirements, including a combination of both alphabetical and numerical references within a single ETD. In at least one case the ETD author also included a full reference section comprising all chapter reference sections at the end of the document; in this instance we decided to only include the chapter reference sections.

Aside from identifying the location of sections and individual citations, some disciplines that rely heavily on archival research and primary resources will include multiple types of source listings. Generally this may include the repositories consulted (for primary sources) as well as a more traditional reference section of "Secondary Sources" that align with other disciplines' reference patterns. In these situations we documented only references that constituted clearly-identified items instead of including references to broader resources or institutional sources.

A final anomaly discovered during the process was that several ETDs from the sciences had a single numerical reference in a reference list that contained multiple individual citations. These individual references were often preceded by an alphabetical ordering like a), b), c), and d), e.g.: "[4] a) K. M. Kadish, Prog. Inorg. Chem. 1986, 34, 435-605. b) Electrochemistry of N4 Macrocyclic Metal Complexes, Eds. J. S. Zagal and F. Bedioui, Springer, 2016." For these references, we separated the combined references into individual references – such as [4a] K. M. Kadish...[4b] Electrochemistry... etc. – so that there would be individual, numerical references for each publication.

It is still too early to report on the effect that the registering of the references for the 2023 ETDs has on the overall use of these document. Additionally, there hasn't been enough time to allow these ETD references to make their way into the larger citation networks that leverage Crossref data. However, we hope that over time we will be able to show the value of this work with an increase of impact as measured by citation and reuse. Moving forward with this project we will continue to register DOIs for each of the ETDs submitted to the UNT Digital Library and also register the references for these documents.

Conclusion

This paper presents the workflow that has been piloted at the UNT Libraries to extract and normalize references for the 412 ETDs submitted in 2023, with a description of processes used to prepare these reference lists. The goal is to provide a clear set of steps and methods that other institutions could replicate without complicated technologies or processes to increase reference registration in Crossref.

During this effort a number of things became apparent. First, manually extracting and formatting references into simple text lists provides a straightforward starting point this work. Second, given an average annual submission of 400 ETDs at UNT, the amount of work required by this process is reasonable. For other organizations, this does assume that student workers are available who can be trained to do these tasks and that an appropriate amount of time is allocated – particularly to address any unexpected complications that arise at the start of implementation. In the future, we will follow a parallel approach of locally extracting references for new submissions while also working with a vendor that has the capacity to work at greater scale (as of the summer of 2024) who is able to assist with the reference extraction to the documents from 1999-2022. Our goal is to register the references for all of the ETDs published by UNT students held in our digital repository (around 21,000 documents).

Aside from continued citation extraction, the collection of nearly 50,000 references for the 2023 ETDs provides a glimpse into the citation patterns of students across the university. While this paper does not delve into that dataset, a cursory look at the references makes it clear that pre-submission activities may be useful to communicate with students in Master's and doctoral programs about the importance of properly-formatted references before ETDs are finalized. As citation networks grow, this process could also provide incentives for students to understand the benefits of paying attention to reference formatting, as well as the utility of including the DOI in the references whenever possible (which makes automated reference matching algorithms more successful).

One final thing that we noticed during this process is that extracting references and registering them with Crossref also highlights connections for restricted research before full data is available. For example, students at UNT have the ability to restrict access to their ETDs for periods of time as outlined by the Toulouse Graduate School (e.g., if they plan to formally publish research elsewhere or to revise an ETD into a book). Because of this, each semester many ETDs are embargoed upon submission and are not accessible to the public until the embargo period elapses. These ETDs were still included in the project because they have been uploaded into the UNT Digital Library with these restrictions. The process of extracting and registering references provides a way of connecting these publications to the greater network of citations without infringing on the students' wishes to keep their ETDs private.

In closing, we have found that the registering of references for ETDs with Crossref is a straightforward way that we can increase the impact of the scholarship of our graduate students at the University of North Texas. By proactively extracting, reformatting, and submitting these references, we are able to contribute to the greater efforts of citation linking and building a more complete and open citation network. These extracted references also offer new and exciting opportunities to better understand what resources our students are directly using for their scholarship. We hope that the model described in this paper will be something other institutions are able to apply and adapt for their local needs so that they too can begin to register the references for the ETDs they manage.