University of North Florida

UNF Digital Commons

9-21-2020

Using the Harvesting Method to Submit ETDs into ProQuest: A Case Study of a Lesser-Known Approach

Marielle Veve

Follow this and additional works at: https://digitalcommons.unf.edu/library_facpub

Part of the Cataloging and Metadata Commons

Using the Harvesting Method to Submit ETDs into ProQuest

A Case Study of a Lesser-Known Approach

Marielle Veve

ABSTRACT

The following case study describes an academic library's recent experience implementing the harvesting method to submit electronic theses and dissertations (ETDs) into the ProQuest Dissertations & Theses Global database (PQDT). In this lesser-known approach, ETDs are deposited first in the institutional repository (IR), where they get processed, to be later harvested for free by ProQuest through the IR's Open Archives Initiative (OAI) feed. The method provides a series of advantages over some of the alternative methods, including students' choice to opt-in or out from ProQuest, better control over the embargo restrictions, and more customization power without having to rely on overly complicated workflows. Institutions interested in adopting a simple, automated, post-IR method to submit ETDs into ProQuest, while keeping the local workflow, should benefit from this method.

INTRODUCTION

The University of North Florida (UNF) is a midsize public institution established in 1972, with the first theses and dissertations (TDs) submitted in 1974. Since then, copies have been deposited in the library, where bibliographic records are created and entered in the library catalog and the Online Computer Library Center (OCLC). During the period of 1999 to 2012, some TDs were also deposited in ProQuest by the graduate school on behalf of students who decided to. This practice, however, was discontinued in the summer of 2012, when the institutional repository, Digital Commons, was established and submission to it became mandatory. Five years later, in the summer of 2017, interest in getting UNF TDs hosted in ProQuest resurfaced. This renewed interest grew out from a desire of some faculty and graduate students to see the institution's electronic theses and dissertations (ETDs) posted there, in addition to a recent library subscription to the ProQuest Dissertations & Theses Global database (PQDT).

A month later, conversations between the library and graduate school began on the possibility of resuming hosting UNF ETDs in ProQuest. Consensus was reached that the PQDT database would be a good exposure point for our ETDs, in addition to the institutional repository (IR), yet some concerns were raised. One of the concerns was cost of the service and who would be paying for it. Neither the library nor the graduate school had allocated funds for this. The next concern was the possibility of ProQuest imposing restrictions that could prevent students, or the university, from posting ETDs in other places. It was important to make sure there were no such restrictions. Another concern was expressed over students entering embargo dates in ProQuest that do not match the embargo dates selected for the IR. This is a common problem encountered by other libraries.[1] For that reason, we wanted to keep the local workflow. The last concern expressed during the conversations was preserving students' right to opt-in or out from distributing their theses in ProQuest. This is something both the graduate school and library have been adamant

Marielle Veve (m.veve@unf.edu) is Metadata Librarian, University of North Florida. © 2020.

about. In higher education, requiring students to submit to ProQuest is a controversial issue which has raised ethical concerns and has been highly debated over the years.[2]

Once conversations between the library and graduate school were held and concerns were gathered, the library moved ahead to investigate the available options to submit ETDs into ProQuest.

LITERATURE REVIEW

Currently, there are three options to submit ETDs into ProQuest: (1) submission through the ProQuest ETD Administrator tool, (2) submission via File Transfer Protocol (FTP), and (3) submission through harvests performed by ProQuest.[3]

ProQuest ETD Administrator Submission Option
In this option, a proprietary submission tool called ProQuest ETD Administrator is used by students, or assigned administrators, to upload ETDs into ProQuest. Inside the tool, a fixed metadata form is completed with information on the degree, subject terms are selected from a proprietary list, and keywords are provided. The whole administrative and review process gets done inside the tool. Afterwards, zip packages with the ETDs and ProQuest's Extensible Markup Language (XML) files are sent to the institution via FTP transfers, or through direct deposits to the IR using the Simple Web-service Offering Repository Deposit (SWORD) protocol. The ETD Administrator submission method presents several shortcomings. First, the ProQuest XML metadata that is returned to the institutions must be transformed into IR metadata for ingest in the IR, a process that can be long and labor intensive.[4] Second, the subject terms supplied in the returned files come from a proprietary list of categories maintained by ProQuest, which does not match the Library of Congress Subject Headings (LCSH) used by libraries.[5] Third, control over the metadata provided is lost because the metadata form cannot be altered, plus customizations to other parts of the system can be difficult to integrate.[6] Fourth, there have been issues with students indicating different embargo periods in the ProQuest and IR publishing options, with instances of students choosing to embargo ETDs in the IR, while not in ProQuest.[7] Lastly, this method does not allow students' choice, unless the ETDs are submitted separately in two systems in a process that can be burdensome. Ultimately, for these reasons, we found the ETD Administrator not a suitable option for our institution.

FTP Submission Option
In this option, an administrator sends zip packages with the institution's ETD files and ProQuest XML metadata to ProQuest via FTP.[8] At the time of this investigation, there was a $25 charge per ETD submitted through this method.[9] We did not want to pursue this option because of the charge and the tedious metadata transformations that would be needed between IR and ProQuest XML schemas. Another way to go around this would have been to submit the ETDs through the VIREO application. VIREO is an open source, ETD management system used by libraries to freely submit ETDs into ProQuest via FTP.[10] This alternative, however, was not an option for us as our IR, Digital Commons, does not support the VIREO application.

Harvesting Submission Option
This is the latest method available to submit ETDs into ProQuest. In this option, ETDs are submitted first into an IR, or other internal system, where they get processed to be later harvested by ProQuest through the IR's existing Open Archives Initiative (OAI) feed.[11] At the time of this writing, we were not able to find a single study that documents the use of this method. This option

looked appealing and worth pursuing as it met most of our desired criteria. First, with this option, students' choice would not be compromised as ETDs would be submitted to ProQuest after being posted in the IR. Second, because the ETD Administrator would not be used, issues with conflicting embargo dates and unalterable metadata forms would be avoided. In addition, the local workflow would be retained, thus eliminating the need for tedious metadata transformations between ProQuest and IR schemas. From the available options, this one seemed the most feasible solution for our institution.

IMPLEMENTATION OF THE HARVESTING METHOD AT UNF

After research on the different submittal options was performed, the library approached ProQuest to express interest in depositing our future ETDs into their system by using a post-IR option.

In the first communications, ProQuest suggested we use the ETD Administrator to submit ETDs because is the most commonly used method. When we expressed interest in the harvesting option, they said "we have not been harvesting from BePress sites" (the company that makes Digital Commons) and suggested we use the FTP option instead.[12] Ten months later, they clarified the harvests could be performed from BePress sites and that the option is free, with the only requirement of a non-exclusive agreement between the university and ProQuest. The news appeased both the library's and the graduate school's previous concerns, as we would be able to adopt a free method that would not compromise on students' choice nor restrict students from posting in other places, while keeping the local workflow.

After agreement on the submittal method was established, planning and testing of the harvesting method began. The library worked with ProQuest and BePress to customize the harvesting process while the university's Office of the General Counsel worked with ProQuest on the negotiation process.

Negotiation Process
Before ProQuest could harvest UNF ETDs, two legal documents needed to be in place. The first document was the Theses and Dissertations Distribution Agreement, which specifies the conditions under which ETDs can be obtained, reproduced, and disseminated by ProQuest. The document had to be signed by the UNF's Board of Trustees and ProQuest. The agreement stipulated the following conditions:

- The agreement must be non-exclusive.
- The university must make the full-text Uniform Resource Locators (URLs) and abstracts of ETDs available to ProQuest.
- ProQuest must harvest the ETDs from the university's IR.
- The university and students have the option to elect not to submit individual works or to withdraw them.
- No fees are due from the university or students for the service.
- ProQuest must include the ETDs in the PQDT database.

The second document that needed to be in place was the Theses and Dissertations Availability Agreement, which grants the university the non-exclusive right to reproduce and distribute the ETDs. This agreement between students and UNF specifies the places where ETDs can be hosted and the embargo restrictions, if any. UNF already has been using this document as part of its ETD workflow, but the document needed to be modified to include the additional option to submit

ETDs into ProQuest. Beginning with the spring 2019 semester, the revised version of the agreement provided students with two hosting alternatives: posting in the IR only or in the IR and ProQuest.

Local Steps Performed Before the Harvesting
The workflow begins when students upload their ETDs and supplemental files (Certificate of Approval and Availability Agreements) directly into the Digital Commons IR. In there, students complete a metadata template with information on the degree and keywords related to the thesis are provided. After this, the graduate school reviews the submitted ETDs and approves them inside the IR platform.

Next, the Library Digital Projects' staff downloads the native PDF files of ETDs, processes them, and creates public and archival versions for each ETD. Availability Agreements are reviewed to determine which students chose to embargo their ETDs and which ones chose to host them in ProQuest, in addition to the IR. If students choose to embargo their ETDs, the embargo dates are entered in the metadata template. If students choose to publish their ETDs in ProQuest, a "ProQuest: Yes" option is checked in their metadata template, while students who choose not to host in ProQuest would get a "ProQuest: No" in their template. (The ProQuest field is a new administrative field that was added to the ETD metadata template, starting with the spring 2019 semester, to assist with the harvesting process. It was designed to alert ProQuest of the ETDs that were authorized for harvesting. More detail on its functionality will be provided in the next section.) The reason library staff enters the ProQuest and embargo fields on behalf of students is to avoid having students enter incorrect data on the template.

Following this review, the Metadata Librarian assigns Library of Congress Subject Headings to each ETD and creates authority files for the authors. These are also entered in the metadata template. Afterwards, the ETDs get posted in the Digital Commons' public display, with the full-text PDF files available only for the non-embargoed ETDs. Information that appears in the public display of Digital Commons will also appear immediately in the OAI feed for harvesting.

At this point, two separate processes take place:

1. Metadata Librarian harvests the ETDs' metadata from the OAI feed and converts it into MARC records that are sent to OCLC, with the IR's URL attached. The workflow is described at https://journal.code4lib.org/articles/11676.
2. On the seventh of each month, ProQuest harvests the full-text PDF files, with some metadata, of the non-embargoed ETDs that were authorized for harvesting from the OAI feed.

Harvesting Process (Customized for Our Institution)
To perform the harvests, ProQuest creates a customized robot for each institution that crawls OAI-PMH compliant repositories to harvest metadata and full-text PDF files of ETDs.[13] The robot performs a date-limited OAI request to pull everything that has been published or edited in an IR's publication set during a specific timeframe. Information to formulate the date limited request is provided to ProQuest by the institution for the first harvest only, subsequently, the process gets done automatically by the robot. The request contains the following elements:

- Base URL of the OAI repository
- Publication set
- Metadata prefix or type of metadata
- Date range of titles to be harvested

In the particular case of our institution, we needed to customize the robot to limit the harvests to authorized ETDs only. To achieve this, we worked with BePress to add a new, hidden field at the bottom of our Digital Commons' ETD metadata template. The field, called ProQuest, consisted of a dropdown menu with 2 alternatives: "ProQuest Yes" or "ProQuest No" (see figure 1). The field was mapped to an element in the OAI feed that displays the value of "ProQuest: Yes" or "ProQuest: No," thus alerting the robot of the ETDs that were authorized for harvesting and the ones that were not. The element used to map the ProQuest field in the OAI feed is the <dc:description.note>, which is a Qualified Dublin Core (QDC) element (figure 2). For that reason, the robot needs to perform the harvests from the QDC OAI feed in order to see this field.

Figure 1. Display of the ProQuest Field's Dropdown Menu in the Metadata Template

Figure 2. Display of the ProQuest Field in the QDC OAI Feed

After the ETDs authorized for harvesting have been identified with help from the "ProQuest: Yes" field, the robot narrows down the ones that can be harvested at the present moment by using the <dc:date.available> element. This element, as the name implies, provides the date when the full-text file of an ETD becomes available. It also displays in the QDC OAI feed (see figure 3). If the date is on or before the monthly harvest day, the ETD is currently available for harvesting. If the date is in the future, the robot identifies that ETD as embargoed and adds its title to a log of embargoed ETDs with some basic metadata (including the ETD's author and the last time it was checked). The log of embargoed ETDs is then pulled out in the future to identify the ETDs that come out of embargo so the robot can retrieve them.

Figure 3. Display of the <dc:date.available> Element in the QDC OAI Feed
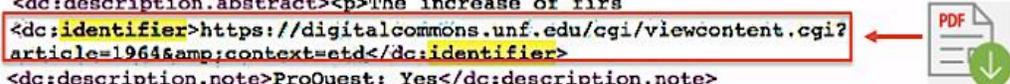
After the ETDs that are currently available for harvesting have been identified (because they have the "ProQuest: Yes" field and a present or past availability date), the robot performs a harvest of their full-text PDF files by using the third <dc:identifier> element, which displays at the bottom of records in the OAI feed (figure 4). The third <dc:identifier> element contains a URL with direct access to the complete PDF file of ETDs that are currently not embargoed. ETDs that are currently on embargo contain a URL that redirects the user to a webpage with the message: "The full-text of this ETD is currently under embargo. It will be available for download on [future date]" (see figure 5).

Figure 4. Display of the Third <dc:identifier> Element at the Bottom of Records in the QDC OAI Feed
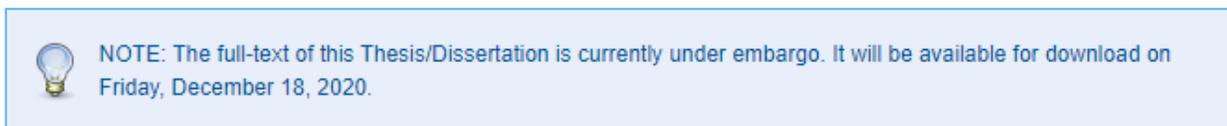
Figure 5. Message that Displays in the URL of Embargoed ETDs

Once the metadata and full-text PDF files of authorized, non-embargoed ETDs have been obtained by the robot, they get queued for processing by the ProQuest editorial team, who then assigns them International Standard Book Numbers (ISBNs) and ProQuest's proprietary terms. It takes an average of four to nine weeks for the ETDs to display in the PQDT database after been harvested. Records in the PQDT come with the institutional repository's original cover page and a copyright statement that leaves copyright to the author. Afterwards, the process gets repeated once a month. This frequency can be set to quarterly or semi-annually if desired.

ADDITIONAL POINTS ON THE HARVESTING METHOD

Handling of ETDs that come out of embargo.
When the embargo period of an ETD expires, the full-text PDF of it becomes automatically available in the IR's webpage, and consequently, in the third <dc:identifier> element that displays in the OAI record. Each month, when the robot prepares to crawl the OAI feed, it will first check for the titles in the log of embargoed ETDs to determine if any of them have become fully available through the third <dc:identifier> element. The ones that become available are then pulled by the robot through this element.

Handling of metadata edits performed after the ETDs have been harvested and published in PQDT.
Edits performed to metadata of ETDs will trigger a change of date in the <datestamp> element that displays in the OAI records. This change of date will alert the robot of an update that took place in a record, which is then manually edited or re-harvested, depending on the type of update that took place.

Sending MARC records to OCLC.
As part of the harvesting process, ProQuest provides free MARC records for the ETDs hosted in their PQDT database. These can be delivered to OCLC on behalf of the institution on an irregular basis. Records are machine-generated "K" level and come with URLs that link to the PQDT database and with ProQuest's proprietary subject terms. We requested to be excluded from these deliveries and continue our local practice of sending MARC records to OCLC with LCSH, authority file headings, and the IR's URLs.

Notifications of harvests performed by ProQuest and imports to the PQDT database.
When harvests or imports to the PQDT have been performed by ProQuest, institutions do not get automatically notified. Still, they can request to receive scheduled monthly reports of the titles that have been added to the PQDT. UNF requested to receive these monthly reports.

Usage statistics of ETDs hosted in PQDT.
Usage statistics of an institution's ETDs hosted in the PQDT can be retrieved from a tool called Dissertation Dashboard. This tool is available to the institution's ETD administrators and provides the number of times some aspect of an ETD (e.g., citations, abstract viewings, page previews, and downloads) has been accessed through the PQDT database.

Royalty payments to authors.
Students who submit ETDs through this method are also eligible to receive royalties from ProQuest.

OBSTACLES FACED

During the planning phase, we encountered some obstacles that hindered progress on the implementation. These were:

- Amount of time it took to get the ball rolling. Initially, we were misled by the assumption we would not be able to use the harvesting method to submit ETDs into ProQuest because we were BePress users, as we were originally told, but that ended up not being the case. Ten months later, we were notified by the same source that the harvesting option for BePress sites would be possible and doable by ProQuest. These were ten months that delayed the implementation process.

- Amount of time it took to get the paperwork finalized and signed before the harvesting. From the moment first contact was initiated with ProQuest, to the moment the last agreement was finalized and signed by both parties, 21 months went by. There was a lot of back and forth in the negotiation process and paperwork between the University and ProQuest.
- Inconsistent lines of communication. There were multiple parties involved in the communication process and some of the emails began with one person only to be later transferred to someone else. This lack of consistency in the communication lines made it difficult to determine who was in charge of particular tasks at certain stages of the process.

CONCLUSION AND RECOMMENDATIONS

Although problems were encountered at the beginning, implementation of the harvesting process at UNF was a complete success. Once the process started, it ran smoothly without complications. Harvests were performed on schedule and no issues with unauthorized content been pulled from the OAI were faced. Fields used to alert the robot in the OAI of the ETDs authorized for harvesting worked as planned, and so did the embargo log used to identify and pull the out of embargo ETDs. It should be noted that Digital Commons users who want to exclude embargoed ETDs from displaying in the OAI can do so by setting up an optional yes/no button in their submission form. This button prevents metadata of particular records from displaying in the OAI feed. We did not pursue this option because we have been using the ETD metadata that displays in the OAI to generate the MARC records we send to OCLC. In addition, we took the necessary precautions to avoid exposing full content of the embargoed ETDs in the OAI feed.

Institutions planning to use this method should be very careful with the content they display in the OAI as to avoid embargoed ETDs from been mistakenly pulled by ProQuest. Access restrictions can be set by either suppressing the metadata of embargoed ETDs from displaying in the OAI or by suppressing the URLs with full access to the embargoed ETDs. The same precaution should be taken if planning to provide students with the choice to opt-in or out from ProQuest.

Altogether, the harvesting option proved to be a reliable solution to submit ETDs into ProQuest without having to compromise on students' choice nor rely on complicated workflows with metadata transformations between IR and ProQuest schemas. Institutions interested in adopting a simple, automated, post-IR method, while keeping the local workflow, should benefit from this method.

ENDNOTES

1 Dan Tam Do and Laura Gewissler, "Managing ETDs: The Good, the Bad, and the Ugly," in *What's Past Is Prologue: Charleston Conference Proceedings,* eds*.* Beth R. Bernhardt et al. (West Lafayette, IN: Purdue University Press, 2017), 200-04, https://doi.org/10.5703/1288284316661; Emily Symonds Stenberg, September 7, 2016, reply to Wendy Robertson, "Anything to watch out for with etd embargoes?," *Digital Commons Google Users Group* (blog), https://groups.google.com/forum/#!searchin/digitalcommons/embargo$20dates%7Csort:date/digitalcommons/RNInGtRarNY/6byzT9apAQAJ.

2 Gail P. Clement, "American ETD Dissemination in the Age of Open Access: ProQuest, NoQuest, or Allowing Student Choice," *College & Research Libraries News* 74, no. 11 (December 2013): 562–66, https://doi.org/10.5860/crln.74.11.9039; FUSE, 2012-2013, Graduate Students Re-FUSE!, https://oaktrust.library.tamu.edu/bitstream/handle/1969.1/152270/Graduate%20Students%20Re-FUSE.pdf?sequence=25&isAllowed=y.

3 "PQDT Submissions Options for Universities," ProQuest, http://contentz.mkt5049.com/lp/43888/382619/PQDTsubmissionsguide_0.pdf.

4 Meghan Banach Bergin and Charlotte Roh, "Systematically Populating an IR With ETDs: Launching a Retrospective Digitization Project and Collecting Current ETDs," in *Making Institutional Repositories Work*, eds. Burton B. Callicott, David Scherer, and Andrew Wesolek (West Lafayette, IN: Purdue University Press, 2016), 127–37, https://docs.lib.purdue.edu/purduepress_ebooks/41/.

5 Cedar C. Middleton, Jason W. Dean, and Mary A. Gilbertson, "A Process for the Original Cataloging of Theses and Dissertations," *Cataloging and Classification Quarterly* 53, no. 2 (February 2015): 234–46, https://doi.org/10.1080/01639374.2014.971997.

6 Wendy Robertson and Rebecca Routh, "Light on ETD's: Out from the Shadows" (presentation, Annual Meeting for the ILA/ACRL Spring Conference, Cedar Rapids, IA, April 23, 2010), http://ir.uiowa.edu/lib_pubs/52/; Yuan Li, Sarah H. Theimer, and Suzanne M. Preate, "Campus Partnerships Advance both ETD Implementation and IR Development: A Win-win Strategy at Syracuse University," *Library Management* 35, no. 4/5 (2014): 398–404, https://doi.org/10.1108/LM-09-2013-0093.

7 Do and Gewissler, "Managing ETDs," 202; Banach Bergin and Roh, "Systematically Populating," 134; Donna O'Malley, June 27, 2017, reply to Andrew Wesolek, "ETD Embargoes through ProQuest," *Digital Commons Google Users Group* (blog), https://groups.google.com/forum/#!searchin/digitalcommons/embargo$20proquest%7Csort:date/digitalcommons/Gadwi8INfgA/sg7de7SdCAAJ.

8 Gail P. Clement and Fred Rascoe, "ETD Management & Publishing in the ProQuest System and the University Repository: A Comparative Analysis," *Journal of Librarianship and Scholarly Communication* 1, no. 4 (August 2013): 8, http://doi.org/10.7710/2162-3309.1074.

9 "U.S. Dissertations Publishing Services: 2017-2018 Fee Schedule," ProQuest.

[10] "Support: ProQuest Export Documentation," Vireo Users Group,
https://vireoetd.org/vireo/support/ProQuest-export-documentation/.

[11] "PQDT Global Submission Options, Institutional Repository + Harvesting," ProQuest,
https://media2.proquest.com/documents/dissertations-submissionsguide.pdf.

[12] Marlene Coles, email message to author, January 19, 2018.

[13] "ProQuest Dissertations & Theses Global Harvesting Process," ProQuest.