8 | Systematically Populating an IR With ETDs: Launching a Retrospective Digitization Project and Collecting Current ETDs

Meghan Banach Bergin and Charlotte Roh

The University of Massachusetts Amherst Libraries established their institutional repository (IR), ScholarWorks@UMass Amherst, in 2006, and we began by systematically populating it with electronic theses and dissertations (ETDs). We currently have a little over 4,500 dissertations and theses in our IR, and they are some of the most highly used content in our repository. Through a partnership with the Graduate School, we collect and disseminate all of our current master's theses and doctoral dissertations through ScholarWorks. We recently launched an ambitious project to scan all 24,000 of our print dissertations and theses and upload them to our IR.

In this chapter we will outline the details of our retrospective digitization project as well as our policies and procedures for collecting current ETD submissions. We will also discuss our recent decision to stop requiring our graduate students to submit their dissertations to ProQuest and the reasons we decided to make this change. At a glance this timeline shows the development of our ETD program:

1997: Began accepting electronic submissions of doctoral dissertations through the ProQuest online ETD submission system.

2006: Began a Digital Commons repository, called ScholarWorks@UMass Amherst, to showcase the research and scholarly output of our students, faculty, and researchers.

2007: Started collecting electronic submissions of master's theses for the first time. Students submit their theses via an online deposit to ScholarWorks.

2009: UMass Amherst Graduate Council institutes a new policy allowing students to choose open access, campus access, and embargoes for their theses and dissertations.

2010: Library decides to go completely e-only for dissertations and theses; print copies are no longer accepted.

2013: Began retrospective digitization project for our print theses and dissertations.

2014: Revised access options for current ETD submissions. We eliminated the permanent campus-only restriction option and replaced it with a temporary campus-only restriction for one year or five years, after which it becomes open access (except for the MFA theses).

2014: Stopped submitting dissertations to ProQuest through their online ETD submission system. All dissertation submissions are now deposited directly into our IR and submission to ProQuest is optional.

BACKGROUND

At the University of Massachusetts Amherst we started our Digital Commons institutional repository, called ScholarWorks@UMass Amherst, to showcase the research and scholarly output of our students, faculty, and researchers. At that time Digital Commons was sold and supported by ProQuest, and one of the selling points of the Digital Commons platform was that it would come prepopulated with a metadata feed linking to all of our digital dissertations in ProQuest's Dissertations and Theses database. This way we did not have to start with a completely "empty box." We knew that for the IR to be successful and to attract our faculty to deposit their research, it had to contain high-quality scholarly content. One of the easiest types of content to collect was our dissertations and theses, since graduate students were already accustomed to submitting their print theses and dissertations to the library. So in 2007, shortly after implementing our IR, we approached the Graduate School about having students submit their master's theses to the IR. The doctoral dissertations were already being submitted electronically to ProQuest for inclusion in their Dissertations and Theses database, but the master's theses were still being submitted on paper, bound, and added to the libraries' print collection. The Graduate School wanted to move to electronic submission for master's theses, and the IR

proved to be just the right solution at just the right time. With the metadata feed linking to our dissertations at ProQuest and the current master's theses being submitted to our IR, we began to think about digitizing all of our older print dissertations and theses in order to build a comprehensive collection.

RETROSPECTIVE DIGITIZATION

It was a long and winding road to launching our retrospective thesis and dissertation digitization (RTD) project. The project had been under consideration since the establishment of our institutional repository. Though we were unable to dedicate time and resources to the RTD project, we did not forget about it. By digitizing our print collection of theses and dissertations and disseminating them through ScholarWorks, we knew they would receive much more use than they do in print format, since print versions are only available to those outside of our university through an ILL request. There were approximately 12,000 print theses and 15,500 print dissertations in our libraries' stacks. Looking back at our circulation statistics, we found that most of them had not circulated since 2006. Only about 3,000 out of 15,500 dissertation titles had circulated since 2006, and the highest circulation amount for any title was 14. Only 1,500 out of 12,000 thesis titles had circulated since 2006, and the highest circulation amount for any thesis was 21. Primarily to make our print theses and dissertations more accessible and increase their chances of being used, we wanted to start digitizing them as soon as we had the resources available to undertake such a large and complex project.

After several years of focusing on scanning books through our scanning contract with the Internet Archive, we had digitized most of the out-of-copyright unique books in our collection and were thinking about what materials to digitize next. An obvious body of unique material was our print dissertations and theses collections. In December 2011, our associate director for Library Services convened a working group to draft a project proposal for our Senior Management Group (SMG) to consider. The working group included the associate director for Library Services, the head of the Information Resources Management (IRM) Department, the Bibliographic Access and Metadata coordinator, the Materials Management Unit coordinator, our Copyright and Information Policy librarian, and our director of Library Development and Communication. The proposal outlined some of

the major benefits of the project, which included showcasing our university's research, making the theses and dissertations openly accessible to a worldwide audience of users, providing access for the graduate students who authored the works, and preserving fragile paper copies. We proposed that the project use an "opt-out" model to digitize these materials. We would make reasonable efforts to contact the authors and let them know about the project to digitize their thesis or dissertation. If the author or copyright holder didn't object, we would make the work openly available through our institutional repository, ScholarWorks@UMass Amherst. If they opted out, their dissertation or thesis would still be digitized but the digital copy would be restricted to campus-only access and ILL lending. We also proposed withdrawing the circulating copies of UMass print dissertations and theses after they are digitized. However, we would be careful to make sure that there was an archival print copy available at the Five College Libraries Depository first. If there was no print copy at the depository, the circulating copy would be transferred to that facility instead of being discarded.

PROJECT IMPLEMENTATION AND WORKFLOW

The retrospective digitization project proposal was approved by the Senior Management Group and a team was formed. The team was headed by the assistant to the associate director for Library Services as project manager and representatives from the IRM Department, Library Development and Communication, and the Scholarly Communication Office. The plan was to first digitize all of the pre-1923 dissertation and theses titles that were in the public domain, and then to start digitizing the W. E. B. Du Bois Department of Afro-American Studies dissertations in the fall of 2013. This department seemed appropriate as the main building of the UMass Amherst Libraries is the W. E. B. Du Bois Library, and there is a strong connection between the libraries and the Afro-American Studies Department. From there we would go department by department to digitize all of the theses and dissertations. In 2015, we are digitizing all of the theses and dissertations from the Astronomy, Chinese, History, Psychology, and Polymer Science Departments, which will total about 2,400 titles. At this rate, we estimate that it will take about 10 years to complete the project.

Initially it took quite a bit of planning and preparation to get the project up and running. Our database analyst/programmer pulled a list of all

of our dissertations and theses and created an Excel spreadsheet with columns for author, title, year of publication, call number, department, and other information from the bibliographic records in our Aleph library catalog. We then added a number of other columns to the spreadsheet to aid us in tracking and organizing the project. These included fields such as scanning status, permissions response, link requests, and author contact information, among others. We call this spreadsheet the Master List.

We also drafted detailed workflow documentation for the project. Our director of Library Development and Communication worked with the university's Alumni Office to obtain contact information for our graduate alumni and worked on drafting a letter to use when contacting authors about the project. The letter informs the authors that UMass Amherst Libraries are undertaking a project to digitize all of our print theses and dissertations and that our goal is to preserve the documents and provide public access to them. We convey to them that we intend to include their thesis or dissertation in the project. We include a form with the letter and tell the authors that if they wish to receive a link to their dissertation after it has been digitized and made available through ScholarWorks@UMass Amherst, they should return the form to us along with their current contact information. We also let them know that if they do not want their dissertation made available for public access, they should select "Opt-Out" on the form. If they do not return the form with the opt-out option checked off, we will digitize the dissertation and make it publicly available through the ScholarWorks IR. We are also placing a list of authors and dissertation titles on our libraries' Web site that we hope allows authors to contact us to either opt-out or request updates. Staff in the Scholarly Communication Office collect the responses from the paper forms and track the information in the Master List. Our library director also writes a letter to the department head to inform him or her about the project each time a new department's theses and dissertations are scheduled to be digitized.

So far the response to the project has been very encouraging. As we notify alumni of the project, we have been asking them to consider a gift to the library in support of the digitization effort, and we're happy to see our graduate alumni giving back. To date we have sent 1,517 letters to our alumni and we have only received 52 opt-out requests. We received another 456 replies from alumni offering their support of the project and requesting

a link to their dissertation. So far only 3% of our authors have chosen to opt out of having their work digitized and made openly available online, and already we have had many positive communications and interactions with our graduate alumni.

One of the most interesting of these exchanges happened when our Book Repair coordinator found several handwritten notes and seven one-dollar bills tucked inside a bound psychology thesis from the 1970s by the author, Rod Kessler, class of ’78. Kessler explained that when he returned to campus with his son, a sports reporter, for athletic events or for Undergraduate Research conferences, he would leave a note and another dollar in the pages of his thesis, each time upping the ante for a potential finder and reader. After graduating from UMass, Kessler eventually went on to become an English professor at Salem State, teaching writing, coordinating the Creating Writing program, managing the campus literary magazine, and serving as head of the magazine before retiring last year. Our director of Development invited Professor Kessler to visit the library, and he accepted the invitation. While at UMass, Professor Kessler expressed his approval of the project, saying, “People spend a lot of time and energy to write these things, and then many of them are never read. I’m glad to have the work out there.” Another author wrote to us saying, “I wrote my dissertation in 1980. I bought one of those IBM typing balls to give to various typists who typed my dissertation. I wanted to be sure that every page looked like it was typed on the same typewriter. I had a few graphs to describe my data. I went to the art supply store and bought some press-on letters and some very thin black tape for the axes and data line. I was very proud of the finished result. Little did I know that one day I would be writing via e-mail to UMass about something called ‘digitizing’ and that I would get a link to my dissertation. Things have changed a lot in 35 years.” While contacting each author has been a lot of work, it has been encouraging to hear the positive responses from people who are glad their work is available to both them and the public.

After the letters were sent out to the authors, the basic workflow of the project was divided into prescanning work and postscanning work. The print copies of the dissertations to be digitized are pulled from the stacks by the Materials Management unit in the Information Resources Management (IRM) Department. Our project includes a detailed prescanning quality

control check to inventory the material, inspect its condition, make repairs, dis-bind if appropriate, note if a copyright symbol is present, and page the archive copy to be sent for scanning if the circulating copy is in poor condition. An Excel spreadsheet that lists all of the titles being sent out for scanning is generated and sent to the Internet Archive. (The Internet Archive calls this spreadsheet a picklist.) Materials are checked out for scanning in the library catalog so we know where they are and so they do not show as available to patrons. The materials are then packed and shipped to the Internet Archive to be scanned.

After the dissertations are scanned by the Internet Archive, the returned shipment is unpacked and the preservation specialist inventories the items and updates the titles in the Master File with the date of digitization. The circulating print copies are then withdrawn from the library catalog and discarded. The completed picklist is sent to our Bibliographic Access and Metadata Unit so that the digital versions of the theses and dissertations can be cataloged and uploaded to our institutional repository. The digital versions of the theses and dissertations are cataloged with an automated cataloging process. We use the completed picklist to identify the Aleph bib numbers of the catalog records for the print versions and then derive new catalog records for the digital versions from the print version records. Those MARC records are then transformed to the bepress XML schema, and the PDFs and their associated metadata are batch uploaded to ScholarWorks. Once the dissertations and theses files have been uploaded, a list of their ScholarWorks URLs is generated and those URLs are inserted into the MARC records with another automated process.

MOVING AWAY FROM PROQUEST

In 2014, we ceased making it a requirement for graduate students to submit their dissertations to ProQuest and instead made it a requirement that they be submitted to our IR. When we initially started working with ProQuest, it was a clear solution because it was the only solution available not only for us but for most academic libraries. ProQuest was, quite frankly, the only game in town, and it was in the common interest for everyone to use the same system so that ETDs would be discoverable in that same database. However, as IRs came into use and as more and more people were using Google and other search engines to find ETDs, it became unnecessary to

have them disseminated by ProQuest. In fact, the statistics show that our dissertations are receiving much more use in our IR than they are in the ProQuest database. In 2011, we contacted ProQuest to ask how much use our dissertations received on ProQuest's website, and ProQuest reported that they were only downloaded seven times on average. This is compared to 360 downloads on average for a dissertation in ScholarWorks.

Another reason we made this change had to do with the fact that the ProQuest and UMass systems did not "talk" to each other, so there was no way to automatically get the dissertation files into our IR. ProQuest would FTP our dissertation files to us and then student library workers had to manually upload them to the IR. This took a lot of time and cost quite a bit in terms of student salaries. Graduate students would also ask ProQuest to embargo things without permission from the Graduate School, which governed policy regarding embargoes. In several instances students embargoed items with ProQuest so that UMass actually did not have access to the dissertations! Through some work, we set up ScholarWorks so that it was capable of handling our ETD submissions with our particular embargoes and access restrictions. We also found that the search engine optimization was much better through the bepress Digital Commons system that ran ScholarWorks, so that search results to a particular title through Google led directly to the ScholarWorks version, which was open and accessible, rather than the entry in the ProQuest database, which was limited to paid subscribers.

Another issue that led to our departure from ProQuest was that our graduates began to find their theses and dissertations for sale on Amazon.com and Barnes & Noble. Legally, ProQuest was within their rights, since students had agreed to third-party sales. However, this check box was not fully explained and was assumed to be for the sake of third-party sales in the form of library databases, not as published books and articles. Students were dismayed to find their work for sale, and there was a real fear that publishers would not contract a book that was already on the market. On the one hand, it behooves all of us to be more careful when reading the fine print. On the other hand, since tenure and promotion is directly tied to publication with established venues, it was difficult to understand why ProQuest did not more thoughtfully consider the impact of its sales program. In November 2014, ProQuest announced that it would no longer sell

theses and dissertations through third-party retailers like Amazon.com, and that it would remove all items currently for sale. This announcement came two years too late, as UMass Amherst, like many libraries, felt that trust had been broken and had already moved away from ProQuest as an ETD solution.

COPYRIGHT AND POLICY

Many of our policies for theses and dissertations were created with the Graduate School. Graduating students retain the copyright to their work, and they can make their work accessible by choosing:

- Complete open access through ScholarWorks (and ProQuest, if they so desire)
- One-year or five-year campus-only access, which moves to open access after the one year or five years is up
- Six-month or one-year embargo, which is a complete restriction to both campus and noncampus users (can be extended)

The embargo can be extended for any number of reasons, whether because a patent is pending, because of issues of research subject privacy, and even for national security. One exception to note is the Master of Fine Arts program here at UMass, which has the option of a permanent campus-only restriction, due to the unique circumstances of the students who are concerned about future publication and sales of their original work.

Students who previously had restrictions will still have those restrictions honored as applicable. For example, James Foley, a journalist who perished in Syria, graduated from UMass Amherst and had chosen to make his thesis available through campus access only. This is a request we continue to honor here at UMass Amherst.

We sometimes receive requests from alumni or recently graduated students asking if they can go back and edit or delete parts of their dissertation or thesis. In situations like this we let the author know that unfortunately we can't make edits to their dissertation or thesis. We explain to them that the libraries are the custodian of the dissertations, but the Graduate School is the approving authority and that requests to alter the works have to go through the Graduate School.

As previously discussed, we work hard to contact all our authors and respect their wishes. However, like many repositories, we find that sometimes that communication is not returned or the rights holder cannot be found. Our policy is to digitize and make public the work and include a responsible take-down policy if the creator contacts us (unless, of course, the work is in the public domain). This policy was formulated with our copyright lawyer/librarian and is based on the legal rights that go along with an implied license. By submitting their work to the library, the authors have given UMass Amherst license to disseminate their work through the library circulating system. Previously this was done in print, but as so many resources have moved online, it is implied that the library has license to disseminate the works through electronic discovery and access.

Every year there are some students who ask if they should register the copyright for their work. It is an additional fee to register a copyright, and typically we advise students that, unless they plan to benefit commercially from their work, registration does not provide additional rights. In fact, making one's work available publicly through the IR does the work of establishing copyright, since there is a record of creation.

CONCLUSION

Our retrospective digitization project is a large, costly, and labor-intensive project, but by spreading the scanning costs and labor out over a 10-year period, rather than trying to digitize everything all at once, we are able to manage it. Since this is still a fairly new project for us, we are continually working to refine and improve our processes. This project requires a great deal of tracking and organization between many different staff members in various departments in the libraries as well as coordination with the authors of the dissertations and theses. We would like to develop better tools and more efficient methods for keeping track of things like permissions, correspondence with authors, whether a title has been digitized or not, if it has been cataloged, and if it has been uploaded to ScholarWorks.

However, there is no denying that there have been huge benefits to students, faculty, and alumni by having work available through ScholarWorks. The usage numbers are dramatic. As previously mentioned, prior to digitization, only 3,000 out of 15,500 print dissertations were checked out. The most highly used print dissertation was checked out 14 times. Only

1,500 out of 12,000 print theses were checked out, and the most highly used print theses had been checked out 21 times. Since digitization, we can see that every single electronic thesis has been downloaded at least once on ScholarWorks. Even if this is just by the author, it is good that the author has easier access to his or her own dissertation or thesis. The average number of full-text downloads for an electronic thesis is 994, and the most highly used thesis on ScholarWorks has been downloaded 231,000 times. As the numbers show, having ETDs available through the IR has been an excellent way to showcase the work of UMass Amherst graduate students and provide worldwide access to their unique and important research.