Making Student Research Data Discoverable: A Pilot Program Using Dataverse

Jennifer Doty, *Emory University*
Melanie Kowalski, *Emory University*
Bethany Nash, *Emory University*
Simon O'Riordan, *Emory University*

Journal Title: Journal of Librarianship and Scholarly Communication
Volume: Volume 3, Number 2
Publisher: Pacific University (Oregon) | 2015, Pages 1-25
Type of Work: Article | Final Publisher PDF
Publisher DOI: 10.7710/2162-3309.1234
Permanent URL: https://pid.emory.edu/ark:/25593/q4f1g

Final published version: http://dx.doi.org/10.7710/2162-3309.1234

JLSC

ISSN 2162-3309 | JLSC is published by the Pacific University Libraries | http://jlsc-pub.org

Volume 3, Issue 2 (2015)

Making Student Research Data Discoverable: A Pilot Program Using Dataverse

Jennifer Doty, Melanie T. Kowalski, Bethany C. Nash, Simon F. O'Riordan

Doty, J., Kowalski, M. T., Nash, B. C., & O'Riordan, S. F. (2015). Making Student Research Data Discoverable: A Pilot Program Using Dataverse. *Journal of Librarianship and Scholarly Communication,* 3(2), eP1234. http://dx.doi.org/10.7710/2162-3309.1234

ISSN 2162-3309 10.7710/2162-3309.1234

PRACTICE

Making Student Research Data Discoverable: A Pilot Program Using Dataverse

Jennifer Doty
Research Data Librarian, Emory University
Melanie T. Kowalski
Copyright & Scholarly Communications Librarian, Emory University
Bethany C. Nash
Scholarly Repository Librarian, Emory University
Simon F. O'Riordan
Metadata Analyst, Emory University

INTRODUCTION The support and curation of research data underlying theses and dissertations are an opportunity for institutions to enhance their ETD collections. This article describes a pilot data archiving service that leverages Emory University's existing Electronic Theses and Dissertations (ETDs) program. **DESCRIPTION OF PROGRAM** This pilot service tested the appropriateness of Dataverse, a data repository, as a data archiving and access solution for Emory University using research data identified in Emory University's ETD repository, developed the legal documents necessary for a full implementation of Dataverse on campus, and expanded outreach efforts to meet the research data needs of graduate students. This article also situates the pilot service within the context of Emory Libraries and explains how it relates to other library efforts currently underway. **NEXT STEPS** The pilot project team plans to seek permission from alumni whose data were included in the pilot to make them available publicly in Dataverse, and the team will revise the ETD license agreement to allow this type of use. The team will also automate the ingest of supplemental ETD research data into the data repository where possible and create a workshop series for students who are creating research data as part of their theses or dissertations.

Received: 02/28/2015 Accepted: 05/06/2015

Correspondence: Jennifer Doty, Robert W. Woodruff Library, 540 Asbury Circle, Atlanta, GA 30322-2870, jennifer.doty@emory.edu

INTRODUCTION

The research data landscape has shifted significantly in response to rising expectations to share data. Research data have become increasingly open artifacts of scholarly communication, independent of the publications they yield. Leveraging new technological advances, scholars can now store and share data easily and inexpensively. Disciplinary practices and funder expectations for data sharing are evolving from that of a private arrangement between two researchers agreeing to share data, to an open decision made by researchers to distribute their data publicly via deposit in an open data archive. In the interest of replication and research transparency, data sets are being included as supplements to electronic theses and dissertations submissions.

In support of this trend and changing expectations, many academic libraries now offer various data management and archiving support services to assist researchers in navigating the complexity of the research data landscape. However, academic libraries should also consider established programs that are natural partners for extending the reach of data management services. One such partner that is widespread across higher education is an Electronic Theses and Dissertations (ETD) program. In 2014, the Libraries at Emory University piloted the use of Dataverse, a data archiving and publication system, leveraging our existing ETD program as a source of content. During this pilot, we examined methods to assist students creating ETDs in archiving and making accessible their underlying research data through Dataverse. We also identified opportunities to provide better support to students as they prepare their research data for submission to an archival repository. By focusing our research data support services around the ETD process, we identified opportunities to teach students best practices in data management and archiving at the time of publication and better prepare the next generation of researchers.

Collie's (2015) recent panel session at the Research Data Access and Preservation (RDAP) Summit indicates that Emory University is not the only institution with an ETD system that does not make supplemental data files easily discoverable. Therefore, this case study describing our approach can better inform those practitioners interested in aligning their data management and archiving services with their ETD programs in order to enhance discoverability of supplemental research data files.

LITERATURE REVIEW

The literature on the implementation of ETD systems often describes the library or institution's role in preserving and distributing the intellectual output of graduate students in the context of the institutional repository (Yiotis, 2008). Fewer articles address the inclusion of supporting research data files with the submission of ETD documents. In fact,

ETDs are often one of the major collections hosted in institutional repositories (Alemneh et al., 2014; Schöpfel, 2013; Song, 2007) and can be considered the "low hanging fruit" for adding value to scholarly output by linking data sets with publications (Collie & Mitt, 2011; Schöpfel et al., 2014; Ubogu & Sayed, 2008). As Schöpfel et al. (2014) succinctly states,

> Linking data to documents is crucial for the interconnection of scientific knowledge…. While academic publishers make usage of new technologies to enrich the content and functionalities of their online products ('article of the future', enhanced multimedia content, etc.), universities have not so far really seized the opportunity of the supplementary files submitted together with electronic theses and dissertations (ETD). (p. 613)

Though institutional repositories may ingest a variety of formats and file types, including recognized research data formats, the submission process may not require the level of documentation necessary for the data to be useful to a future researcher. In order for data to be re-usable, subject matter experts and investigators familiar with the research should include context and methods used to collect or generate the data. In certain disciplines there is recognition that supplemental files can enhance the scholarly record by ensuring all information necessary to replicate research results is made accessible along with the publication (King, 1995). Data sets are archived in response to expectations and requirements to share the data produced from sponsored research (National Institutes of Health, 2003; National Science Foundation, 2010) and underlying journal articles (Alsheikh-Ali, Qureshi, Al-Mallah, & Ioannidis, 2011). However, publishing research data simultaneous to and independent from the research article is a trend observed in response to both these requirements and in recognition of the need to improve research reproducibility (Collins & Tabak, 2014). For data sets to be considered published as a form of scholarly output in their own right, they should meet similar criteria of established scholarly publications, such as being available, citable, and documented (Kratz & Strasser, 2014); questions and proposed solutions still surround the degree of review and validation required of archived data sets to ensure reproducibility and re-use (Kratz & Strasser, 2014; Costello, Michener, Gahegan, Zhang, & Bourne, 2013). In the case of our pilot, the data sets in question underwent a basic technical review to check the format of the data, ensure the files are not corrupted, and enter minimal descriptive information gleaned from the thesis or dissertation metadata. The current workflow for the ETD submission systems evaluated by Collie and Mitt (2011) identifies many limitations to effective data curation:

- data are not made automatically available to exam committee
- data are disjointed from the document, and co-linking between data and document is not possible

- data inherit the restrictions placed upon the ETD. (p. 169)

By fully incorporating data curation support into the ETD submission process, the dissertation as publication can be enhanced. (Collie & Mitt, 2011).

Institutions with established repositories and ETD programs have an opportunity to provide access to more robust collections of research by incorporating support for curated data sets and establishing linkages between the published ETD and the underlying data that support the findings. The common practice of ingesting supplemental files into the institutional repository (IR) along with the ETD is a one size fits all approach to heterogeneous data. According to Bardi & Manghi (2014), academic journals take a similar approach in which "supplementary material is typically stored locally into the information system of the journal and it is not discoverable and not accessible outside the context of the related article." Schöpfel et al. (2014) argues that,

> because of the specific nature of data and supplementary files, it appears appropriate not to store text and data files in the same repository but to distinguish between document server and data repository and to deposit text and data files on different platforms, or at least to separate them on an early stage of the workflow and to handle them in different information system environments. (p. 618)

With our pilot program for ETDs at Emory University, we considered Schöpfel's concern and developed a hybrid approach that will allow the research data underlying theses and dissertations to remain in context within the ETD repository while encouraging discoverability and re-use.

INSTITUTIONAL CONTEXT

Emory ETD Program

The Emory ETD program launched in 2007 as a partnership between Emory Libraries and the Laney Graduate School. The program now also includes theses and dissertations from the Candler School of Theology, the Undergraduate Honors program, and the Rollins School of Public Health and is actively expanding as other schools and programs on campus develop graduate programs and express interest in making the work of their students openly accessible. The Emory ETD repository and application (http://etd.library.emory.edu), developed by the Libraries, accommodate a range of policies from schools that address embargoes and submission to ProQuest, a third-party information content vendor, for inclusion in its Dissertations & Theses Full Text Database. Librarians support

the Emory ETD program with a number of workshops covering topics such as copyright and submission to the repository.

The Emory ETD application manages and automates portions of the submission and publication process for schools that participate in the ETD program, ingests ETD content into Emory's Fedora repository, and serves as an avenue for researchers to locate Emory ETDs. The application, built in PHP, utilizes Dublin Core, Metadata Object Description Schema (MODS), PREservation Metadata: Implementation Strategies (PREMIS), and Relationships-External (RELS-EXT) datastreams to create a system able to accomplish these tasks.

The process begins when a student creates her record using the ETD application. The application connects with the central university database, and using a student's university-supplied login (NetID), the ETD application pulls data from the central database into the student's record. The NetID is also employed as an author disambiguation tool, as NetIDs are unique to each student and are not re-used after a student graduates. Before the student submits her record, she attaches a PDF of her thesis or dissertation, the same document in its original file format (usually Microsoft Word), and any supplemental files. Supplemental files can include video, audio, software, text, and data sets. Currently, we accept most file formats, but individual ETD records have a 4 GB maximum size limit.

After the student completes her submission, an administrator within her school reviews the submission for accuracy and completeness. The ETD system allows the administrator to either approve the record or return it to the student for corrections. Following approval by the student's school, it remains unpublished in the ETD repository until the student graduates.

Upon graduation, the central university database is updated to reflect the change in the student's status. Because Emory ETDs connects to the central database, this change in status also triggers the approved ETD record to publish in the ETD repository. Students are allowed to embargo their work; however, the embargo only applies to the uploaded files, abstract, and table of contents. The title, author name, degree, keywords, committee members, and other information are publicly accessible after publication unless the entire record is restricted due to patent, privacy, or security concerns.

Publication of an ETD record makes the record's metadata available to the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) feed. The library's discovery layer, a local implementation of Ex Libris' Primo, subscribes to this feed to enable discovery of ETD records through Primo. Emory ETD records are also discoverable directly via the ETD application and Google. The information pushed to Primo does not include metadata

about any supplemental files attached to an ETD record, and the search functionality in the Emory ETD application does not offer the ability to search for supplemental files or their respective metadata.

For students required by school policy to submit to ProQuest, the ETD application automates that process by bundling the PDF file of the dissertation with the associated metadata and submitting it directly to ProQuest. These students' publications are then discoverable in the ProQuest database. However, supplemental files are not included in the bundle sent to ProQuest. If a student wants to make her supplemental files available via the ProQuest database, she must submit them directly to ProQuest. There is minimal information available about the supplemental files, which may include research data, in the ProQuest database. These limitations of the Emory ETD repository and ProQuest's database decrease the discoverability of the research data associated with theses and dissertations. Therefore, a data repository, such as Dataverse, is an attractive tool to help bridge the gap in discoverability.

Emory's Research Data Management Support

Emory Libraries have a long history of supporting data users from across the academic community. In 1996, the library established the Electronic Data Center to serve as a resource for faculty and students in locating and acquiring data, and in preparing it for analysis. After participating in the Digital Library Federation (DLF) E-Science Institute in 2011, and in response to an increasing number of researchers seeking guidance on managing their data, Emory Libraries hired two additional professional positions in 2012 to focus on research data management. We conducted an institutional survey of all faculty researchers to gather information about existing perspectives and practices managing research data (Akers & Doty, 2013) and held follow up interviews with faculty and graduate student researchers to provide additional context to the environment at Emory. Concurrently, we acquired education and experience in data curation and management topics through participation in two separate external pilot projects in 2012-2013 (Doty, Herndon, Lyle, & Stephenson, 2014; Southeastern Universities Research Association, 2014). These pilots informed our goals and plans for data archiving support moving forward.

Our involvement with the Southeastern Universities Research Association (SURA) Dataverse Pilot Project allowed us to explore the capabilities of Dataverse as a data repository. The Dataverse Network software was developed as an open source application to facilitate the ability to publish, share, reference, extract, and analyze research data and is in use at several research universities and institutions (dataverse.org). When the SURA pilot ended, we elected to continue using the Dataverse Network hosted by the Odum

Institute at the University of North Carolina at Chapel Hill (Southeastern Universities Research Association, 2014). Dataverse provides a place for researchers to deposit data sets, receive a permanent identifier, include references to related publications, and be assured of a commitment to keeping the data accessible and preserved. Through an agreement established between Emory and Odum, we will continue to use Dataverse as an archive for data generated at our institution that do not have an appropriate disciplinary repository, particularly for researchers seeking to comply with funder requirements and journal policies. We are also exploring other options to identify or create a suitable long-term data archiving solution which meets the principles of data citation, accessibility, and preservation.

Dataverse Legal Requirements

When initially establishing the agreement with the Odum Institute to conduct the Dataverse Pilot Project, we wanted to ensure that the roles and responsibilities of the Odum Institute, Emory University, and the depositing Emory researcher were clearly articulated from the outset. We included not only the relationship between Emory and the Odum Institute, but the respective relationships between Emory and the researcher, as well as the researcher and the Odum Institute. To achieve this goal, we established a memorandum of understanding (MOU) with Odum, entitled "Data Deposit Agreement" (see Appendix A), and revised the standard Dataverse "Data Deposit Form" (see Appendix B) language to meet our institutional specifications. In constructing the language for both documents, we took care to ensure that the language appropriately reflected the roles of the three main parties involved with each data deposit: Odum as the owner of the service, Emory University as the middle man or access conduit for the researcher to the service, and the depositing researcher as the principal steward of the data.

Though it does not require an MOU with all participating institutions, the staff at the Odum Institute did provide us with an example. We then took this language and revised it to better reflect the specific instances of our Dataverse use. Since this was a pilot endeavor, we took care to ensure that the agreement clearly articulated how and when the data would be returned to the researcher should either Emory or the Odum Institute choose to end the service.

We revised the language for the Data Deposit Form signed by each researcher in accordance with Emory's specific use of Dataverse. Because the library's role in the implementation of Dataverse is one of a middle man, we ensured that the form defined the researcher's responsibilities in sharing the data. We included language that specified the researcher's responsibilities for maintaining confidentiality and alignment with human subjects research regulations (e.g. the Health Insurance Portability and Accountability Act (HIPAA) and

related policies) or other proprietary obligations. Additionally, we wanted to manage our future users' expectations by conveying to users how the data and metadata would be used following submission of the data. While the activities outlined above pre-dated the ETD pilot, developing appropriate data deposit language was a necessary step for implementation, and these types of activities could be beneficial to others piloting similar programs.

DATAVERSE ETD PILOT PROGRAM

Dataverse ETD Pilot Program Overview

Following the decision to continue with the Emory Dataverse hosted at the Odum Institute, we outlined possible next steps and defined an appropriate course of action that included conducting a local pilot using Emory University research data. We identified existing data within the Emory ETD repository as candidates for inclusion in this local Dataverse pilot. A member of the Emory Libraries' software engineering team compiled an initial report containing all ETD records with supplemental files published from 2007-2013, which the Scholarly Repository Librarian used to pinpoint ETD records that included research data as supplemental files. Shortly after the Dataverse pilot launched, Emory Libraries chartered a Metadata Working Group to promote best practices in metadata creation, and "to define a set of core, discovery-focused, schema-agnostic metadata elements supporting local content types" (Emory Libraries & Information Technology, 2015). This group's work resulted in the creation of a set of Core Metadata Elements (http://metadata.emory.edu/guidelines/descriptive/core-metadata.html) for digital collections at the institution. To develop the set of core elements, the Metadata Working Group collected required and recommended elements from Emory Libraries' applications and the metadata schema used, including the Emory ETD repository and Dataverse. The Working Group also analyzed fields from the Dataverse system and the Data Documentation Initiative (DDI) schema as part of the background research to develop the proposed elements. The Dataverse ETD pilot team recognized an opportunity to test the newly proposed Emory Libraries Core Metadata Elements by mapping them to corresponding metadata fields within Dataverse and creating a template of required and recommended elements for use in the project.

The addition of a Research Library Fellow assigned to work with Research Data Management Services allowed adequate staff time to fully embark on the pilot project with the ETD program in Fall 2014. A second ETD report was generated, this time for ETD records published in 2014. We created a template within Dataverse based on the proposed Core Metadata Elements, reviewed and prepared the ETD data files, and began ingesting the ETD records into Dataverse. After manually checking each ETD record containing supplemental research data, we selected records with identifiable research data (e.g. in recognized file

formats such as those listed in Table 1, page 10), downloaded the files from the ETD repository, and prepared them for archiving. Individual records were then created for each thesis or dissertation within Dataverse using the metadata template developed from the proposed Core Metadata Elements. The Dataverse ETD Pilot Workflow section following describes this process in greater detail.

Dataverse ETD Pilot Workflow

The local Dataverse pilot began with a report generated by library IT staff listing every record with supplemental files in the ETD repository. From 2007-2014 there were 183 records with supplemental files out of a total of 3,983 published ETD records.

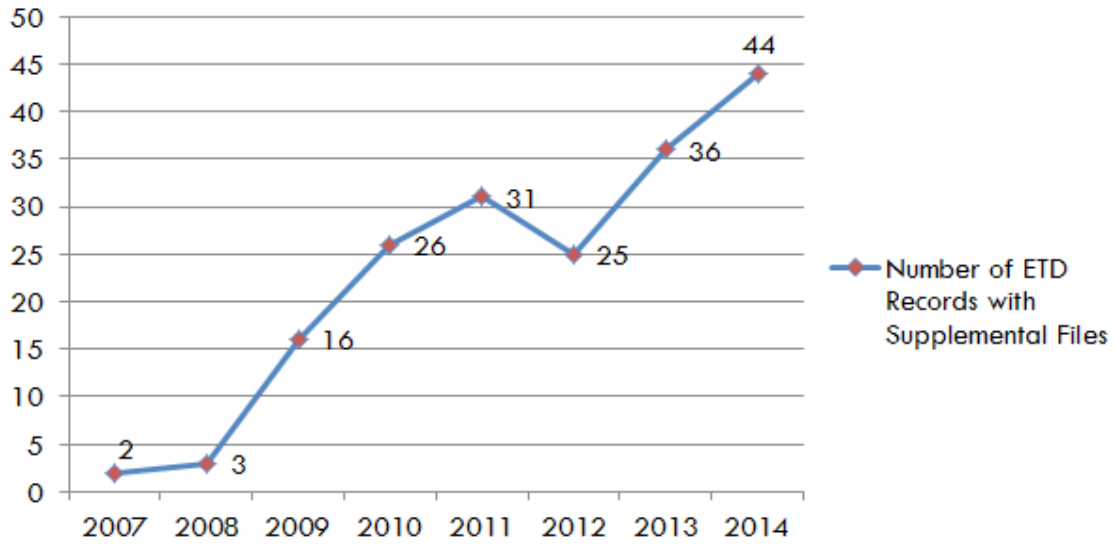


Figure 1. ETD Records with Supplemental Files by Year

Due to technical limitations, the report contained only the record's unique identifier, so the Scholarly Repository Librarian reviewed each record to determine the file formats of any supplemental data and created a more detailed version of the report for the Research Data Management Services team. After the Scholarly Repository Librarian forwarded the updated report, the Research Library Fellow manually checked whether the supplemental files contained identifiable research data. We defined "research data" as materials collected or generated in the course of conducting research that is necessary to validate research findings.

Example formats commonly used included tabular formats such as Comma Separated Values (.csv), Microsoft Excel (.xls, .xlsx), and files from recognizable statistical programs such as R and Stata, as well as software code written in C++ and Python (.cpp, .py), and geospatial file formats such as Keyhole Markup Language (.kml, .kmz) files, and map project files (.mxd) and shapefiles (.shp) used with Geographic Information Systems such as ArcGIS.

File Format Type	# of ETD Records with this File Type
Tabular File Formats (.xls, .csv, .xlsx, .dta, .R, SAS Files)	13
Software Code (.hpp, .py, .rst, .cpp)	4
Geospatial Data (.mxd, .kmz, .kml, .gpx, etc…)	1

Table 1. ETD Records with Supplemental Research Data by File Type (2009-2014)

Next, the Fellow evaluated each supplemental data file to confirm the file could be opened and whether the thesis or dissertation record (including the supplemental files) was subject to an embargo. Once he evaluated all the ETD records, he downloaded any identified research data files from the ETD repository. If an ETD record had a large number of research data files, he batch downloaded and archived the files as a .zip file.

He cleaned the supplemental research data files by correcting misnamed file extensions (e.g. a .csv file labeled as a .doc file) and any obvious misspellings in the file name (e.g. "spreadsheet" for "spredsheet"). When a user downloads supplementary files from the ETD repository, the system automatically renames files with the convention "[ETDAuthor]_ supplement." The Fellow performed additional cleanup to rename the data file according to the naming convention "[ETDAuthor]_[original file name given by the student]".

After preparing the supplemental data files for ingest, the Fellow created a record in Dataverse for each corresponding ETD record. As discussed earlier, we based the metadata template used for creating the Dataverse records on the Emory Core Metadata Elements. The Fellow applied the same template to all ETD data files regardless of discipline, which echoes the treatment of ETD records in the ETD repository (see Appendix D).

The Fellow titled each Dataverse record as "Data for: (Title of ETD)" with the supplemental data set attached as individual files to the Dataverse record. He categorized each file as

a "data file" with the file name following the naming convention mentioned previously. When students provided documentation as a supplemental file to their ETD record, he included the file in the Dataverse record, naming it using the same naming convention, and categorized it as a "documentation file" (see Appendix D). In the case of the ETD records with a large number of data files, the Fellow created an additional .zip file comprising all of the data files made available alongside the individual files. Additionally, Dataverse automatically added a MD5 checksum for each file to verify data integrity during storage as well as a persistent identifier for a consistent citation. Each Dataverse record also includes the original thesis or dissertation citation in the "Publication" field with a link back to the associated record within the ETD repository. We kept the same keywords for each Dataverse record as those selected by students when submitting their theses or dissertations into the ETD repository. The contact for each record lists the Scholarly Communications Office, as the point of contact for Emory University's ETD program.

In total, seventeen ETD records contained supplemental files identified as research data: nine from the first sweep of ETDs published from 2009-2013 and eight from a second sweep of ETDs published in 2014. The majority of research data sets are from the life sciences, not a surprise considering Emory University's close connection to the Center for Disease Control and Prevention (CDC), and the Emory Healthcare system.

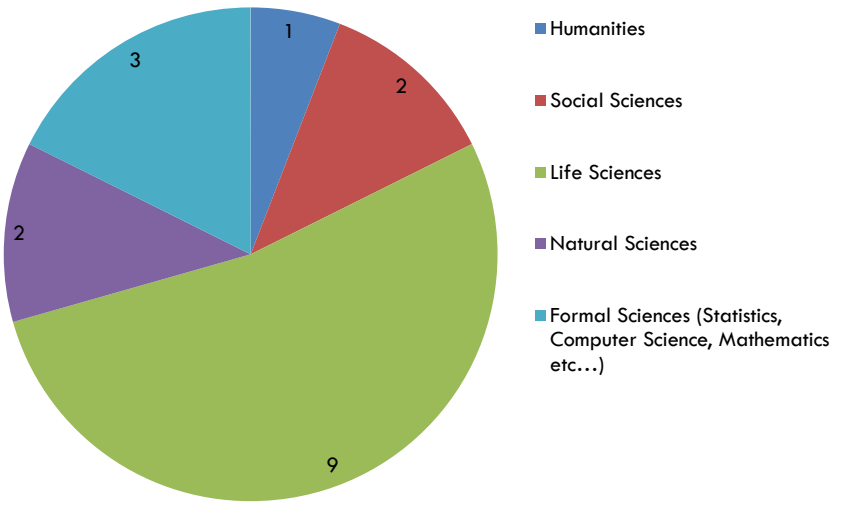


Figure 2. ETD Records with Supplemental Research Data by Discipline (2009-2014)

Another point of interest is that the number of ETD records with supplemental research data files doubled in one year alone (2014) compared with the previous ETDs identified with supplemental research data, which covered the period 2009 to 2013.

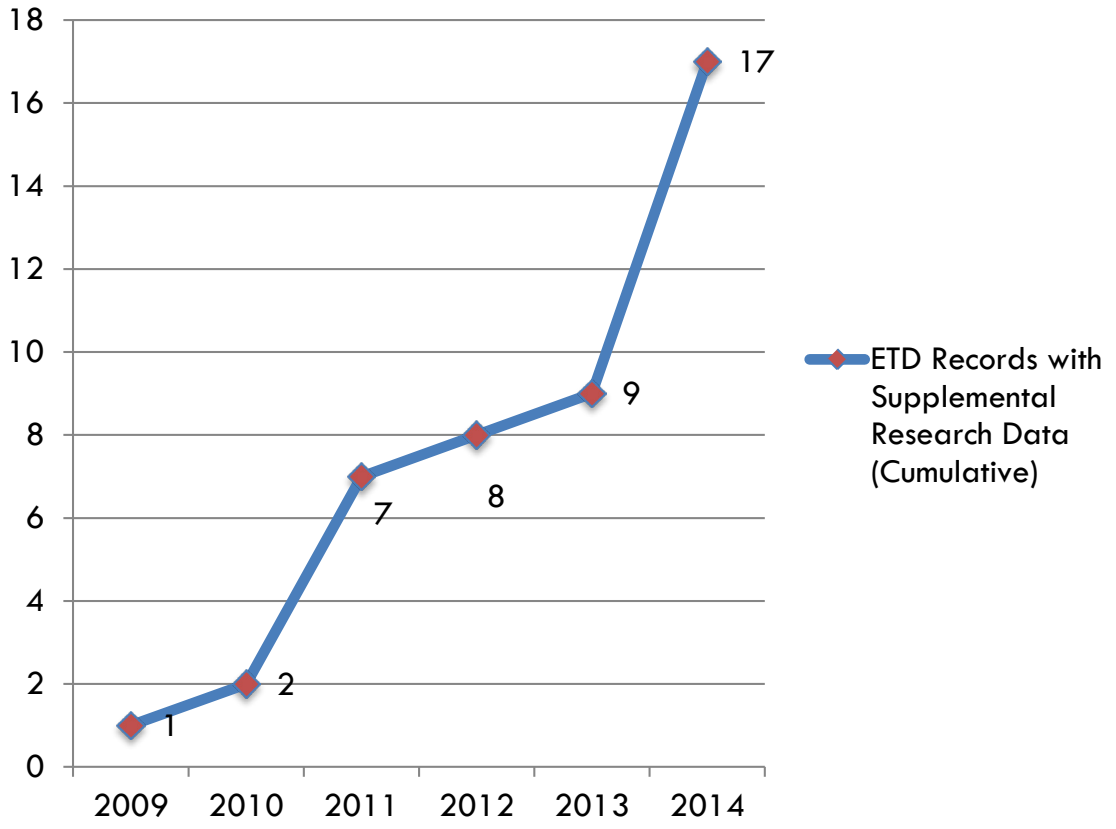


Figure 3. ETD Records with Supplemental Research Data (Cumulative)

Out of those seventeen identified ETD records, one had potential intellectual property issues and was deemed unsuitable for ingest. We deposited the remaining sixteen into the Dataverse. These Dataverse records are complete, but not yet released since the original ETD repository submission agreement (see Appendix C) did not grant the library permission for public dissemination of supplemental files through a separate data repository, and the students have not completed a Dataverse deposit form.

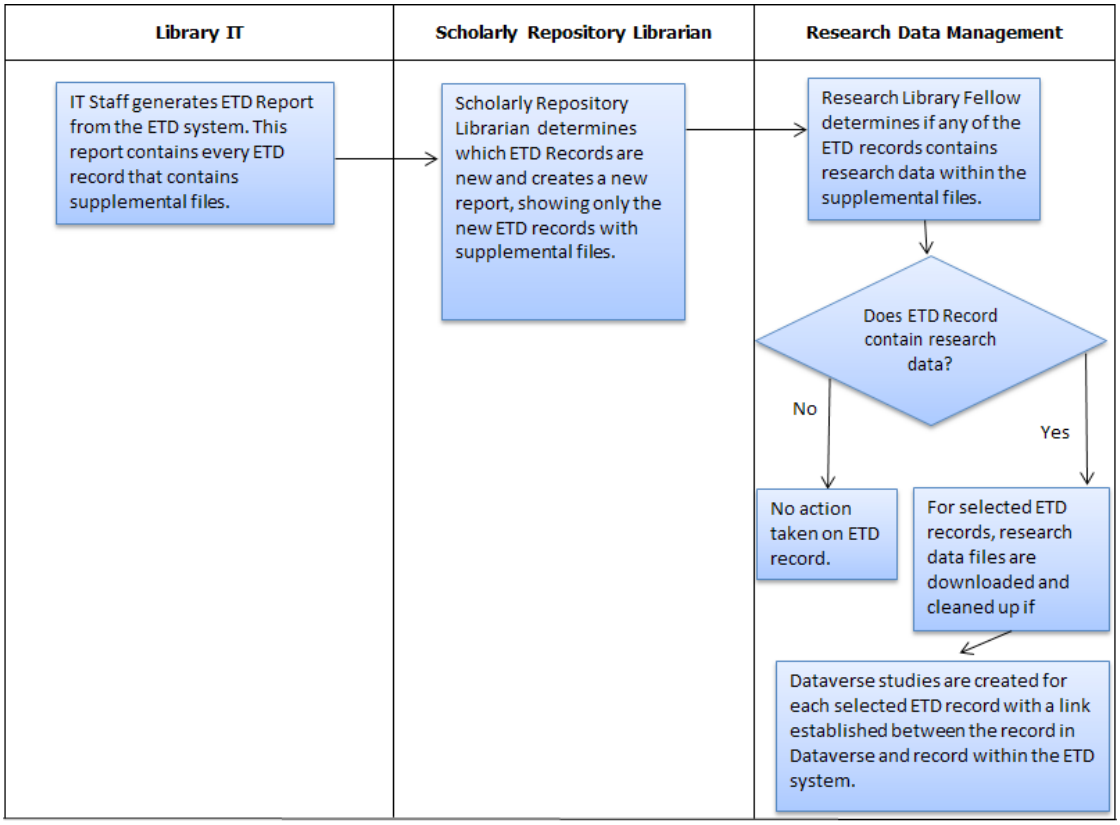


Figure 4. Workflow for Dataverse ETD Pilot Program

Research Data Services Outreach to Students

Beginning in the Fall 2014 semester, research data and related services were added to the topics covered during ETD submission workshops, including an introduction to the library staff providing these services. These workshops are held near the end of each semester and are designed to walk students through the submission process and address any last minute concerns before they submit their ETD to the repository. In addition to the submission process and research data, copyright is also addressed during the workshops. The submission workshops are well attended and consistently receive positive reviews in the survey distributed to attendees.

As another avenue to reach students as early career researchers, a session on "Archiving Dissertation Data to Support Responsible Research" was offered as part of the Jones Program in Ethics in the Laney Graduate School. All doctoral candidates in the graduate school

are required to complete the ethics program and can choose from a selection of sessions covering topics related to responsible conduct of research, including data management. The audience for the data archiving session consisted primarily of second and third year students in disciplines ranging from the humanities to the social sciences, health sciences, natural and physical sciences, and business. The objectives of the session were to raise awareness of best practices and trends in data management and consider issues surrounding responsible stewardship of research data. It also provided an opportunity to encourage students early in their PhD programs to think about the data they collect and generate during dissertation research, consider how they could make the data accessible upon publishing their findings, and prepare students to archive the underlying data associated with their dissertations.

LESSONS LEARNED

The initial phase of the Dataverse ETD Pilot wrapped at the end of the Spring 2015 semester, and we are currently evaluating our work. While we have not yet completed a full analysis, we have identified lessons learned that could be useful for other practitioners planning their own data archiving support service.

First, the small number of ETD records with supplemental research data files limited the size of our pilot. The submission of supplemental research data files is not required, which we believe results in fewer students providing these files with their ETD. Other institutions interested in leveraging their ETD program for research data files should consider this limitation before structuring a pilot. Additionally, we acknowledge that the pilot itself was not marketed directly to students. Rather than advertising our pilot program to leverage additional submissions, we chose to instead study the status quo. This strategy limited the number of supplemental research data files included in the pilot, but kept the total number of files manageable for a pilot. Were we to expand upon the pilot, we would need to consider the scalability of the service.

Second, the technical limitations of our ETD repository limited our options. We chose to explore Dataverse as an archiving and discovery solution rather than building out the existing Fedora repository because the latter option was not feasible for a pilot study. Any enhancements to the ETD application require development time from the Emory Libraries' software engineering team. Given the development needs of other library projects, we did not have access to software engineers for this pilot. Further, analyzing the contents of the ETD repository required a manual process because the ETD system lacks functionality to support analysis of supplemental files.

Third, the current policy of the ETD program placed restrictions on the pilot project. As stated above, the ETD program does not require students to submit research data

underlying their theses and dissertations as supplemental files. Additionally, the ETD repository submission agreement language does not explicitly give us the necessary rights to make supplemental research data available through a third party system (see Appendix C). Therefore, we did not make the identified data sets publicly available through Dataverse as part of our initial pilot.

Finally, we were constrained by the limitations of the Dataverse platform. Dataverse currently only allows users to select whether a particular record will be available to the public or restricted; there is no option to set an embargo for a specified amount of time. Any user who wishes to provide public access to their data must manually lift the access restriction. This lack of functionality is problematic, particularly for those users who need to place access restrictions on their data in order to comply with publisher or other mandated embargo periods. Because we were unable to provide public access to the records in Dataverse as part of our pilot, we were not hindered by the inability to embargo records for a set length of time. However, robust embargo functionality, including the ability to set automatic expiration dates, is desirable. Without it, the burden falls to the alumnus or repository staff.

NEXT STEPS

The Dataverse pilot highlighted a number of ways we can improve our services. First, we intend to seek permission from the alumni whose data were included in the pilot project so that we can make the data publicly accessible through Dataverse. Currently, the data used in the pilot project are included in Dataverse in an archival sense. Due to what we have learned through the pilot, we plan to revise the ETD submission agreement to explicitly allow for this type of use in the future. The license agreement is covered in detail with graduating students during the submission workshops each semester, so the changes can be presented to students as they prepare to submit their ETD. In addition, we will cover applicable licensing and deposit terms for placing their research data in the Dataverse or an alternate disciplinary data repository.

A subcommittee of Emory's Research Data Management Task Force has been charged with researching and recommending a repository strategy for locally generated research data without an appropriate disciplinary archive. Based on the findings of that subcommittee, the data repository could take the form of an Emory-hosted instance of Dataverse or another data repository solution. While the subcommittee conducts its work, Emory will continue to utilize the Odum-hosted instance of Dataverse for this type of research data. The manual data entry and data migration of the pilot necessitated a high level of staff time. For each ETD record, an average of thirty minutes was required to clean and deposit the supplemental data into Dataverse. For records with a large number of data files, the

workflow was markedly longer. The lengthy, manual process is unsustainable as the service grows, particularly if the number of ETD records with supplemental research data continues to trend upward. Once the Task Force recommends a data repository solution, we will partner with our software engineering team to identify where the data ingest process can be automated in order to simplify the service for students and library staff. By automating the ingest of supplemental data into the data repository, we are in a better position to integrate this service into the ETD program.

Lastly, we plan to expand our workshop offerings to include a series focused on students who will be producing non-text files, including research data, as part of their theses or dissertations. These workshops will be scheduled early in the semester to reach students as they are preparing these files for inclusion in the repository. By reaching students before they are ready to submit their ETDs, we hope to discuss preferred file formats, considerations for archiving, and any requirements for their work in the repository. Additionally, we hope that these workshops will help the students minimize the mistakes made in creating and labeling their supplemental data. Teaching students how to better document and format their data sets will help with ingesting supplemental data into our future repository and also serve students well as they continue their careers as researchers by reinforcing the importance of making data available and accessible for re-use by providing them with that experience.

CONCLUSION

Providing support for students to archive the data underlying their ETDs is an ideal opportunity to both ensure the preservation of our institution's unique scholarly output and cultivate data management best practices with early career researchers. Preparation to document and disseminate data is currently lacking as part of the curriculum and research experience of the typical graduate student. As Carlson, Johnston, Westra, and Nichols (2013) reported from the Data Information Literacy Project,

> Faculty want their students to acquire a richer understanding and appreciation for good data management practices, but there are several barriers that restrain faculty from taking action. First, spending time on data management can be deemed detrimental if it is seen as distracting or delaying the research process. Second, faculty do not necessarily see themselves as having the knowledge or resources to impart these types of skills to their students themselves. One faculty member mentioned requirements by funding agencies for data management plans and journals accepting supplemental data files as positive steps, but researchers in his field were ill-prepared to respond. (p. 211)

Focusing our data archiving services at an identified point of need, when students are learning about the research publication life cycle as it pertains to the culminating product of their educational experience, provides an example for how to build system workflows and exploit available outreach opportunities to prepare this next generation of researchers to practice good data management.

ACKNOWLEDGEMENTS

We would like to thank Jonathan Crabtree and Thu-Mai Christian at The Howard W. Odum Institute for Research in Social Science at UNC-Chapel Hill for supporting Emory's continued use of the Odum Institute Dataverse Network. We would also like to thank Lisa Macklin and Robert O'Reilly for their assistance revising and finalizing the legal documents establishing the agreement between Emory University and the Odum Institute. Our appreciation extends to Alexander Thomas for generating the supplemental files report from the ETD repository, and to Frances Pici for her moral support in all things ETD. Finally, we thank Lisa Macklin for comments on the final draft of this article.

REFERENCES

Akers, K. G., & Doty, J. (2013). Disciplinary differences in faculty research data management practices and perspectives. *International Journal of Digital Curation*, *8*(2) 5-26. http://dx.doi.org/10.2218/ijdc.v8i2.263

Alemneh, D., Donovan, B., Halbert, M., Han, Y., Henry, G., Hswe, P., ... Wang, X. (2014). *Guidance documents for lifecycle management of ETDs (Version 1.0. ed.)* (M. Schultz, N. Krabbenhoeft, & K. Skinner, Eds.). Atlanta, GA: Educopia Institute. Retrieved from http://educopia.org/publications/gdlmetd

Alsheikh-Ali, A. A., Qureshi, W., Al-Mallah, M. H., & Ioannidis J. P. A. (2011). Public availability of published research data in high-impact journals. *PLoS ONE*, *6*(9), e24357. http://dx.doi.org/10.1371/journal.pone.0024357

Bardi, A., & Manghi, P. (2014). Enhanced publications: data models and information systems. *LIBER Quarterly*, *23*(4). Retrieved from http://liber.library.uu.nl/index.php/lq/article/view/8445/9825

Carlson, J., Johnston, L., Westra, B., & Nichols, M. (2013). Developing an approach for data management education: A report from the data information literacy project. *International Journal of Digital Curation*, *8*(1), 204-217. http://dx.doi.org/10.2218/ijdc.v8i1.254

Collie, A. (2015). Building organizational capacity for data collections using electronic theses & dissertations. *RDAP15 Summit*. Retrieved from http://www.slideshare.net/aaroncollie1/building-organizational-capacity-for-data-collections-using-electronic-theses-dissertations

Collie, W. A., & Witt, M. (2011). A Practice and value proposal for doctoral dissertation data curation. *International Journal of Digital Curation*, *6*(2), 165-175. http://dx.doi.org/10.2218/ijdc.v6i2.194

Collins, F. S., & Tabak, L. A. (2014). Policy: NIH plans to enhance reproducibility. *Nature*, *505*(7485), 612–613. http://dx.doi.org/10.1038/505612a

Costello, M. J., Michener, W. K., Gahegan, M., Zhang, Z-Q, & Bourne, P. E. (2013). Data should be published, cited and peer-reviewed. *Trends in Ecology and Evolution*, *28*(8), 454-461. http://dx.doi.org/10.1016/j.tree.2013.05.002

Doty, J., Herndon, J., Lyle, J., & Stephenson, L. (2014). Learning to curate. *Bulletin of the Association for Information Science and Technology*, *40*(6), 31-34. http://dx.doi.org/10.1002/bult.2014.1720400610

Emory Libraries & Information Technology. (2015). Metadata: About: Research and Process. Retrieved from http://metadata.emory.edu/about/research.html

King, G. (1995). Replication, replication. *PS: Political Science and Politics*, *28*(3), 444-452. http://dx.doi.org/10.1017/S1049096500057607

Kratz, J., & Strasser, C. (2014). Data publication consensus and controversies [v3; ref status: indexed, http://f1000r.es/4ja]. *F1000Research*, *3*(94). http://dx.doi.org/10.12688/f1000research.3979.3

National Institutes of Health. (2003). *Final NIH statement on sharing research data.* Retrieved from https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html

National Science Foundation. (2010). *NSF data sharing policy*. Retrieved from http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/aag_6.jsp#VID4

Schöpfel, J. (2013). Adding value to electronic theses and dissertations in institutional repositories. *D-Lib Magazine*, *19*(3/4). http://dx.doi.org/10.1045/march2013-schopfel

Schöpfel, J., Chaudiron, S., Jacquemin, B., Prost, H., Severo, M., & Thiault, F. (2014). Open access to research data in electronic theses and dissertations: an overview. *Library Hi Tech*, *32*(4), 612–627. http://dx.doi.org/10.1108/LHT-06-2014-0058

Song, I. (2007). Promoting open access to scholarly data: A case study of the electronic thesis and dissertation (ETD) project at the Simon Fraser University Library. *Data Science Journal*, *6*, S70-S78. http://dx.doi.org/10.2481/dsj.6.S70

Southeastern Universities Research Association. (2014). *SURA DVN pilot project report of findings*. Retrieved from http://www.sura.org/news/docs/DVNPilot.pdf

Ubogu, F. N., & Sayed, Y. (2008). Management of research data in ETD systems. In *ETD 2008 11th International Symposium on Electronic Theses and Dissertations,* June 4-7, 2008, The Robert Gordon University, Aberdeen, UK. Retrieved from http://www4.rgu.ac.uk/etd/programme/page.cfm?pge=45695

Yiotis, K. (2008). 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*, *24*(2), 101-115. http://dx.doi.org/10.1108/10650750810875458

APPENDIX A
DATA DEPOSIT AGREEMENT BETWEEN ODUM AND EMORY

DATA DEPOSIT AGREEMENT

This Data Deposit Agreement ("Agreement") is between the University of North Carolina at Chapel Hill on behalf of The Howard W. Odum Institute for Social Sciences ("Odum Institute") and the institution that signs below ("Emory University").
WHEREAS, the Odum Institute is a research institute with a world-renowned data archive and Emory University is a research university with a community of researchers interested in sharing and preserving research data; and
WHEREAS, the Odum Institute and Emory University seek to collaborate in the archiving of research data to increase access to research and promote inter-institutional cooperation; and
WHEREAS, Emory University seeks to provide access to the Odum Institute's data repository for Emory University affiliated faculty, students, and staff ("Emory Users") planning to deposit their research data for sharing and preservation purposes.
NOW THEREFORE, the Odum Institute and Emory University hereby agree as follows:

1. **LICENSE.** Emory University, on behalf of Emory Users, grants Odum Institute permission to use any data deposited by Emory Users pursuant to this agreement ("Data") as specified in the applicable Data Deposit Form, insofar as Emory University holds any rights in the Data. Emory University, on behalf of Emory Users, further grants Odum Institute the right to reproduce and make derivatives of the Data solely to the extent necessary in connection with its services pursuant to this Agreement and for preservation purposes, insofar as Emory University holds any rights in the Data. Emory Users reserve all rights to Data not specifically granted herein.
2. **RIGHTS.** Emory University, on behalf of Emory Users, represents to Odum Institute that it has all rights necessary to deposit Data in the Odum Institute repository, insofar as Emory University holds any rights in the Data, including without limitation any permission required to include personal information relating to research subjects.
3. **LIABILITY.** Emory University agrees that Odum Institute and the University of North Carolina at Chapel Hill assume no liability pursuant to this Agreement for any claims arising out of any legal action concerning identification of research subjects, breaches of confidentiality, or invasions of privacy by or on behalf of research subjects, or for any loss or damage to deposited Data. Emory Users assume

responsibility for ensuring that Data deposited do not include sensitive, confidential, or proprietary information.

4. **WITHDRAWAL.** Emory Users may voluntarily withdraw their Data from Odum Institute repository at any time, provided they give written notification to Emory University, and in such event Odum Institute shall provide the Emory User with a copy of such Data and remove such Data from its repository, provided however Odum Institute shall have the right to retain one copy of such Data solely for preservation purposes unless the Emory User informs Odum Institute otherwise.
5. **FEES.** No fee shall be associated with hosting of Data by Odum Institute, provided however that Odum Institute reserves the right to withdraw its services upon sixty (60) days written notice in the event Odum Institute elects to charge a fee for hosting Data and the parties fail to execute a written agreement setting forth mutually agreeable fees. In the event Odum Institute elects to cease hosting Data, it shall provide all such data to Emory University prior to removing the Data from its repository.
6. **ODUM INSTITUTE RESPONSIBILITIES.** Upon receipt of Data, Odum Institute agrees that it shall host Data on secure servers that are backed up daily, and that it shall, at the election of Emory Users as indicated on the Data Deposit Form, make such Data available to users of the Odum Institute website. Odum Institute further agrees to (i) include the Data in its catalog of holdings; (ii) archive the Data consistent with other data in the Odum Institute repository; (iii) make Data available to Emory University, Emory Users, and any other third parties as specified in the applicable Data Deposit Form; and (iv) implement any restrictions to the Data as indicated by Emory Users in the applicable Data Deposit Form.
7. **EMORY UNIVERSITY RESPONSIBILITIES.** Emory University will ensure that every Emory User depositing Data in the Odum Institute repository signs the applicable Data Deposit Form.
8. **TERMINATION.** This Agreement may be terminated by either Party upon thirty (30) days prior written notice to the other party. In the event either Party elects to terminate the Agreement, Odum Institute shall provide all Data deposited by Emory Users to Emory University prior to removing the Data from its repository.
9. **USE OF NAMES.** Neither party shall use the name or marks of the other in any promotional material or other publicity without the prior written consent of that party.
10. **FORCE MAJEURE.** Neither party shall be liable to the other for failure to perform any of its respective obligations imposed by this Agreement provided such failure shall be occasioned by fire, flood, explosion, lightning, windstorm, earthquake,

subsidence of soil, governmental interference, civil commotion, riot, war, terrorism, strikes, labor disturbance, or any other cause beyond its reasonable control.

11. **ENTIRE AGREEMENT.** Unless otherwise specified, this Agreement and the Data Deposit Form embody the entire understanding between Emory University and Odum Institute with respect to the Data, and any prior or contemporaneous representations, either oral or written, are hereby superseded. No amendments or changes to this Agreement shall be effective unless made in writing and signed by authorized representatives of both Odum Institute and Emory University.

IN WITNESS WHEREOF, Odum Institute and Emory University, intending to be legally bound, have executed this Agreement as of the date of last signature below by their respective duly authorized representatives.

APPENDIX B
DATA DEPOSIT FORM

Data Deposit Form for Emory University

PART I – DATA DEPOSIT AGREEMENT

All data provided pursuant to this Data Deposit Form are subject to the Data Deposit Agreement ("Agreement") executed between the Odum Institute for Research in Social Science and Emory University on behalf of Emory affiliated faculty, students, and staff ("Emory Users").

Please sign below. By signing, you agree to the following:

- You own the data collection and/or you have secured permissions to make it publicly available through the Odum Institute Dataverse Network, and you agree to comply with the Dataverse Network Account Terms of Use, as included in the network account creation process.

- In preparing this data collection for archiving and public distribution, you have removed all information directly identifying the research subjects in these data, and have used due diligence in preventing information in the collection from being used to disclose the identity of research subjects. You affirm that these data do not contain any sensitive, confidential or proprietary information that you desire or are required to keep confidential.

- The Odum Institute and the University of North Carolina at Chapel Hill assume no liability from the Agreement for claims arising out of any legal action concerning identification of research subjects, breaches of confidentiality, or invasions of privacy by or on behalf of said subjects, or for any loss of or damage to deposited data collections.

- The Odum Institute has the right to use the data collection for the following purposes, without limitation:
  - To disseminate this data collection under Odum's standard terms of use, including the sharing of data with the partners of the Data-PASS project for the purpose of preservation and future access.
  - To promote and advertise the data collection in any publicity and form
  - To describe, catalog, validate and document the data collection
  - To incorporate metadata or documentation in the data collection into public access catalogues
  - To store, translate, copy, or re-format the data collection in any way to ensure its future preservation and accessibility
  - You may voluntarily withdraw your data at any time, provided you give Emory University written notification. Odum Institute shall provide you with a copy of your data and remove such data from its public repository. Odum Institute reserves the right to retain one copy of such data after removal solely for preservation purposes unless you inform Odum Institute otherwise.

Printed Name and Title

Click here to enter text.

APPENDIX C
ETD REPOSITORY SUBMISSION AGREEMENT

Electronic Thesis and Dissertation (ETD) Repository Submission Agreement Form

For Masters Theses and Doctoral Dissertations

Student Name:	
Student ID#:	
Graduate Program:	
Document Title:	

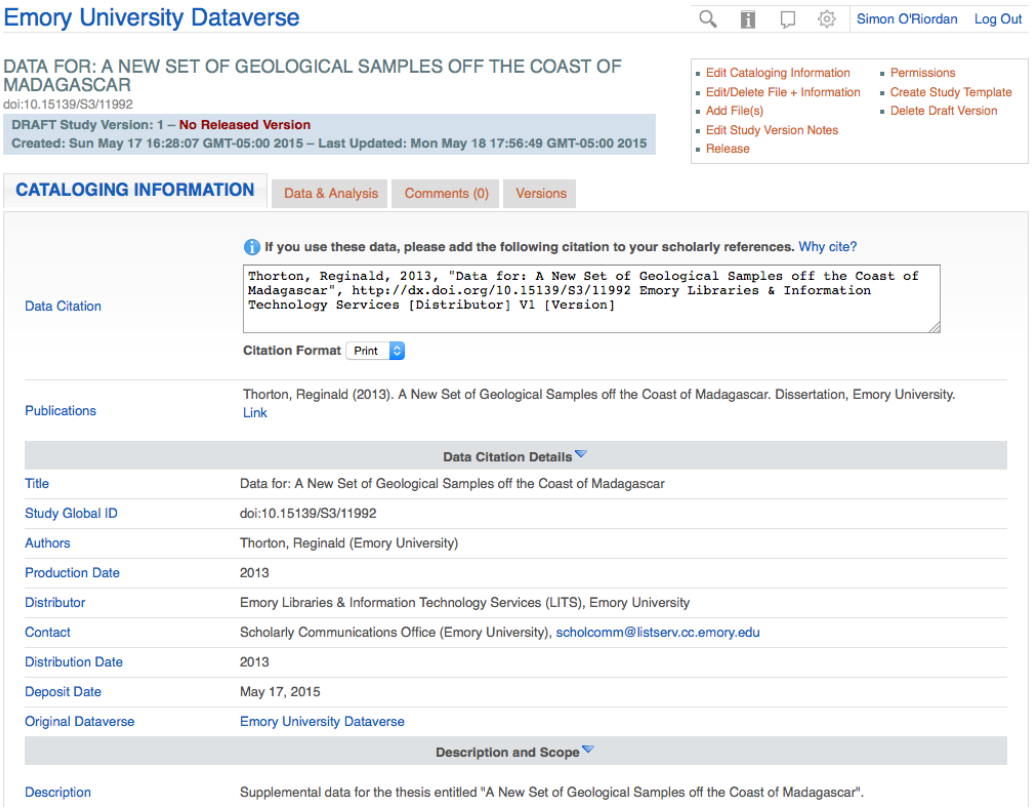
(Check One:) Thesis ☐ or Dissertation ☐

Please Note: You are the owner of the copyright in your thesis/dissertation. By executing this document you are granting permission to Emory University to publish this document on the world wide web (immediately upon graduation unless otherwise specified). For dissertations, you will be completing an additional set of forms granting permissions to ProQuest/UMI.

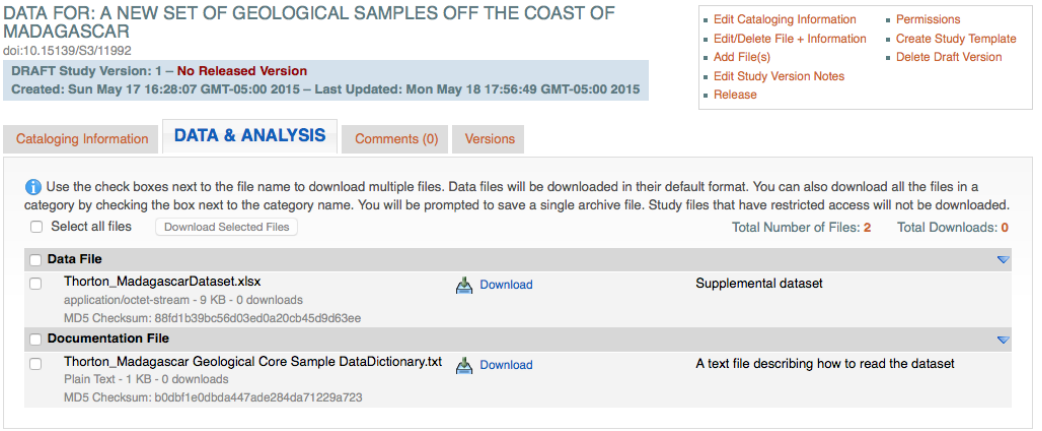
Part 1 - Author Agreement:

I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display, subject to the conditions specified below in Part 3, my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including the display of the thesis or dissertation on the world wide web. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation. I certify that my electronic submission is the version of my thesis/dissertation that was approved by my committee.

APPENDIX D
SCREENSHOTS OF A DATAVERSE RECORD



Screenshot of a section of a completed Dataverse record



Screenshot of where the supplemental research files are located within the Dataverse record.