**The International Journal of Digital Curation**

**Issue 2, Volume 6 | 2011**

# A Practice and Value Proposal for Doctoral Dissertation Data Curation

W. Aaron Collie,

Digital Curation Librarian,

Michigan State University

Michael Witt,

Assistant Professor of Library Science,

Purdue University

## Abstract

The preparation and publication of dissertations can be viewed as a subsystem of scholarly communication, and the treatment of data that support doctoral research can be mapped in a very controlled manner to the data curation lifecycle. Dissertation datasets represent "low-hanging fruit" for universities who are developing institutional data collections. The current workflow for processing electronic theses and dissertations (ETD) at a typical American university is presented, and a new practice is proposed that includes datasets in the process of formulating, awarding, and disseminating dissertations in a way that enables them to be linked and curated together. The value proposition and new roles for the university and its student-authors, faculty, graduate programs and librarians are explored.[1]

[1] This paper is based on the paper given by the authors at the 6th International Digital Curation Conference, December 2010; received December 2010, published July 2011.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. ISSN: 1746-8256 The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre.

# Introduction

The electronic thesis and dissertation (ETD) process is, in many ways, a scale model of the scholarly communication lifecycle. Each ETD cycle produces a dissertation document that is vetted by peers and disseminated when the degree is granted. Original data are produced that support the candidates' dissertation research; however, these data are rarely included and considered in their full form in the current practice. The Research Information Network (2008) recommends that services should be offered to meet the needs of users who are increasingly interested in gaining access to both data and documents. In this paper, a new practice and value proposal for ETD data curation are outlined and will be described in the context of a typical American university. New roles for the university and its student-authors, faculty, graduate programs, and librarians are explored.

ETDs present a tractable fragment of the research spectrum to address. At a typical university, doctoral candidates become accustomed to a regimented process for preparing and submitting their dissertations, defending them, and disseminating them on a common platform (e.g., ProQuest[2]) with other dissertations. As a "captive audience", they may be more motivated and inclined to self-submit their data and descriptive metadata than other possible content producers. As universities heed Swan and Brown's (2008) call for a strategic repositioning of the library to support data-intensive research, one "low hanging fruit" to consider for collection development are student research datasets. Implementing new practices, such as the ability to co-link data and documents; self-archive research data; and widen the access to data, can augment the value of ETD collections and improve their research impact.

Anecdotally, some doctoral candidates have expressed frustration over the limits imposed by the document format to properly substantiate their findings. For example, a rich, microscopic image may lose its informational value by being reduced to one-bit, black-and-white images on a printed page. Lippincott and Lynch (2010) express that while ETD technology has existed for years, students are still "advised to produce straightforward text dissertations that do not take advantage of new technologies." Granting students the option to self-select curation activities at the beginning of the dissertation writing process not only improves curation awareness by encouraging students to consider the inclusion of non-text formats, but as Borgman (2007) argues, combining scholarly documents and research data "enhance[s] the value chain of scholarship" from input to output. This means that students who wish to present research data should expect a greater return on their initial effort with a more valuable research output. Opportunities to disseminate data as citable, scholarly objects, alongside the document, increase exposure and potential impact at a critical point in the students' careers, when they are applying and interviewing for their first post-doctoral jobs.

From a university's perspective, collecting and preserving dissertation data are likely to be closely aligned with its mission and purpose. Faculty can have supporting data immediately in-hand to evaluate results and more effectively vet the student's research. Librarians who interpret and apply collection development policies to evaluate prospective datasets can make a strong argument for including ETD data in their collections as a part of the intellectual record of the institution. In the same way

[2] ProQuest: http://www.proquest.com/en-US/catalogs/databases/detail/pqdt.shtml.

that ETDs helped seed many early institutional e-print repositories, ETD datasets can help populate fledgling institutional data repositories. These practices align with Ubogu and Sayed's (2008) recommendation that the stewardship of dissertation data should be viewed in the broader context of the management of institutional research data. While their survey findings suggest a disconnect between ETD programs and data management centers, it also represents an opportunity for institutions to investigate data curation issues in a scaled environment.

## Typical Approach

### *Overview of Current ETD Workflows*

A typical workflow for the electronic deposition of a dissertation begins with a student either indicating or being assigned the status of a degree candidate and ends with the required deposition of a document and the optional "attachment" of supplementary data. The workflow concludes with a peer review process, a series of formatting specifications, and an official acceptance and deposition event.

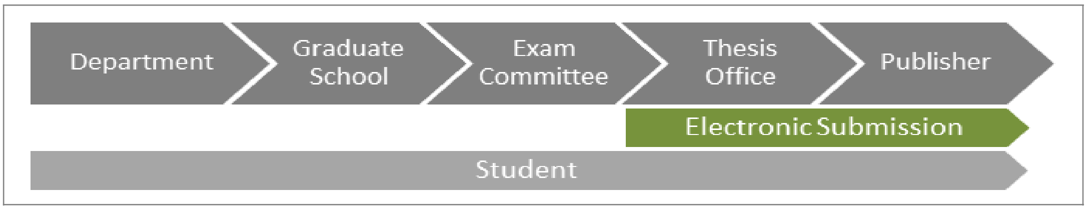

Figure 1. Overview of draft-to-dissertation in a typical system.

While various departments may provide assistance, the majority of processes are initiated and carried to completion by the student. These responsibilities may be mandatory (e.g., forms, signatures, approval letters) or optional (e.g., instruction sessions, pre-deposit format meetings, informal meetings with advisors). Some of the student's responsibilities are logistical, for example: managing deadlines; scheduling appointments; managing, submitting, forwarding and carbon copying appropriate forms; and acquiring signatures. While other responsibilities are academic and/or professional: communicating with advisors; drafting, redrafting, and revising dissertation documents; meeting the expectations of the exam committee; meeting the expectations of the departmental format check; and meeting the expectations of the university wide format check. The workflow of a typical American university is summarized in Figure 2.

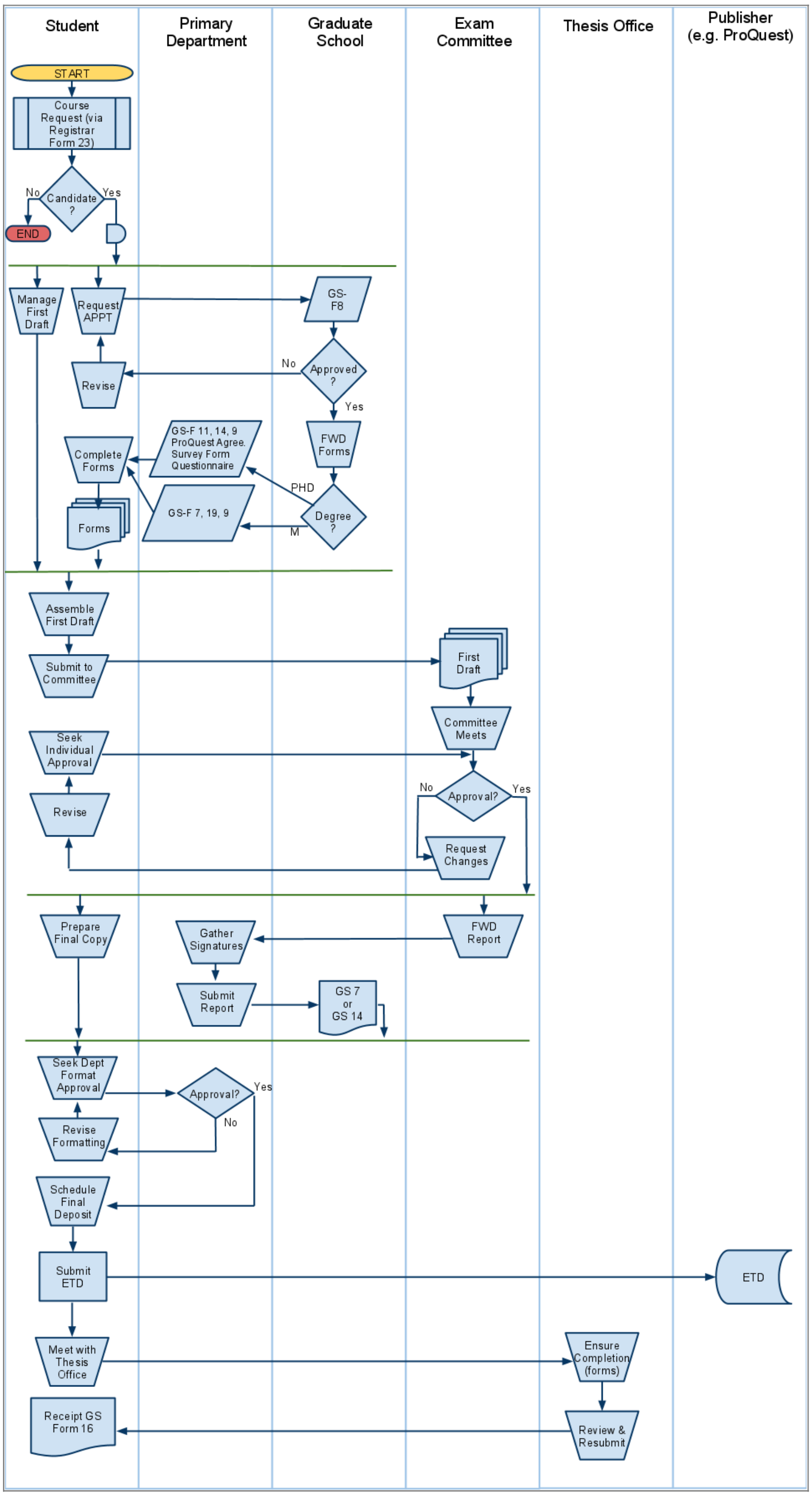

Figure 2. An ETD workflow at a typical, American university.

### *Current Workflow Narrative*

In the current practice, a student initiates the workflow by requesting an exam committee. The graduate school must approve the exam committee and schedule its meeting(s). The graduate school then forwards appropriate forms to the primary department. The student acquires, completes, and attaches forms to the dissertation draft. The student then submits a completed draft to exam committee. If the student meets all expectations of the exam committee, he or she will acquire their signatures. The exam committee will then forward an approval report to the student's primary academic department. The primary department ensures all signatures have been acquired and submits a report of approval to the graduate school. The student is then required to meet all expectations for department- and university-wide format specifications. Once the student has acquired all forms and met all expectations, he or she schedules an appointment with the thesis office. The student then deposits the dissertation prior to final appointment. The final deposit requires that student meets with thesis office staff to verify completion of all forms, all format checks and all final corrections. During the final deposit, data may be attached to the ETD package as supplementary files. The thesis office provides receipt of deposit.

### *Limitations of Current ETD Workflows*

Current workflows that support ETDs are not built to support the curation of research data. ETD workflows are typically extensions of print-based dissertation workflows, and are developed out of necessity, as graduate schools shift from print to digital deposition. With the gradual implementation of ETD workflows, thesis offices issued revisions or supplements to dissertation manuals. It is possible to see then how the traditional method of attaching research data to a print-based dissertation (attaching a CD-ROM supplement in a back pocket on the inside of the binding) has been revised to attaching data as a "last step" in the deposition process. Data submitted as a supplemental attachment during the final electronic deposition of a dissertation or thesis exposes the following limitations:

- Data are not made automatically available to the exam committee and require extra steps to an already complex approval process;
- Data are disjointed from the document, and co-linking between data and document is not possible;
- Data inherit the restrictions placed upon the ETD package.

A typical method of mitigating the new burdens of electronic deposition is to outsource collection and dissemination of student's dissertations from the library to commercial vendors. Unfortunately, many vendor platforms impose further limitations on the ability to curate data. Students who are restricted to using the deposition package provided by the vendor are also limited to using only the file formats supported by the vendor's database. This means that format and file-size restrictions limit what the student can attach as data. Similarly, because the primary means of accessing dissertations is often through the vendors interface, which is typically a series of forms on a website that are oriented towards describing and uploading a single document. While functionality exists for attaching supplementary data to a dissertation in systems such as ProQuest's, this functionality is often poorly implemented; in our example case it was used by fewer than 1% of student-authors in the last five years. A vendor's license and access restriction on the dissertation usually imposes the same restrictions on the supplementary data. Dissertations that are outsourced to commercial vendors are submitted in batch loads and can take up to 6-9 months to appear in print and online.

## Data-Augmented Approach

### *Overview of Proposed ETD Workflow*

A practical approach to ETD management utilizes a content management system that is specialized to manage the submission and tracking of digital dissertations. Examples of ETD management systems (ETDMS) include openETD, VIREO, VALET, HYDRA, and FEZ. ETDMS are capable of the total management of dissertations from authentication of candidacy to deposition of ETD. They can streamline workflow, improve status tracking, and increase the operational efficiency of the dissertation process. Open source ETDMS can be extended by software developers to provide new functionality to meet local needs. An example of such an extension may be to introduce functionality that includes research data with the early drafts of a document and a parallel workflow to support data curation, such as minting persistent identifiers for data objects.

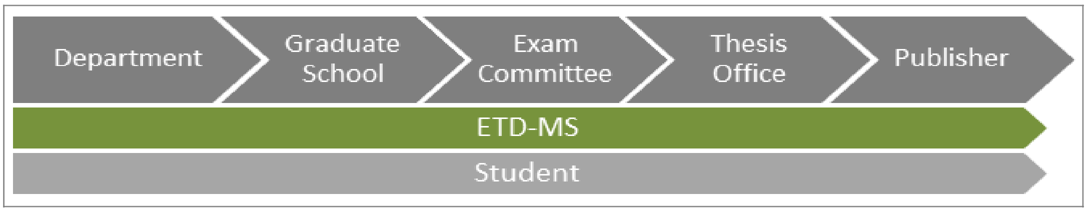

Figure 3. Overview of draft-to-dissertation in proposed system.

In our theoretical case, an ETDMS improved efficiency by nearly 50% by decreasing the amount of required manual processes from 21 to 11. A considerable portion of the stress is alleviated from the student's role by decreasing manual processes from 13 to 6 (see Figure 4). These decreases are a result of automating and re-engineering nine processes, while at the same time modifying the workflows to support data. An example workflow is summarized in the next section.

### *Example Workflow*

The data-augmented approach begins by an ETDMS notification alerting the student to their obligations as a degree candidate. This notification triggers subsequent exam committee scheduling and deadline management notifications, as well as the forwarding of appropriate forms. The student is required to negotiate these deadlines and to submit a draft dissertation to the ETDMS prior to the exam committee meeting. Upon submission of the dissertation draft, the student views the option to include any research data in the review. Should the student wish to include data, the ETDMS assigns persistent identifiers (e.g., a DOI) to the student's data objects. The student is presented with a citation to the data objects with instructions for citing the data in their dissertation document.

Faculty advisors and the major professor receive notification of the status update once the student has submitted their dissertation draft and complementary data. The exam committee may choose to view and respond to the dissertation draft and data within the context of the ETDMS, or in a parallel print process. Upon approval of the draft, the ETDMS forwards the appropriate reports and prepares the student (via notifications) for the final submission process. During the final submission, the student has the option to self-archive the document and data in a local pre-print repository. Lastly, the final version of the dissertation is exported to the publisher.

## Value Proposition

The data-augmented approach illustrates one workflow that would support the kind of early inclusion of data that would make curation possible. The primary modification made in our example case is to avoid a supplementary, “data as attachment” treatment. Data that are attached at the end of the ETD process bypass the possibilities to add value, whereas data-augmented ETDs allow data and document to pass through all activities of the ETD workflow and be treated together. The following value proposition highlights some of the value that can be added by including research data earlier in the ETD workflow:

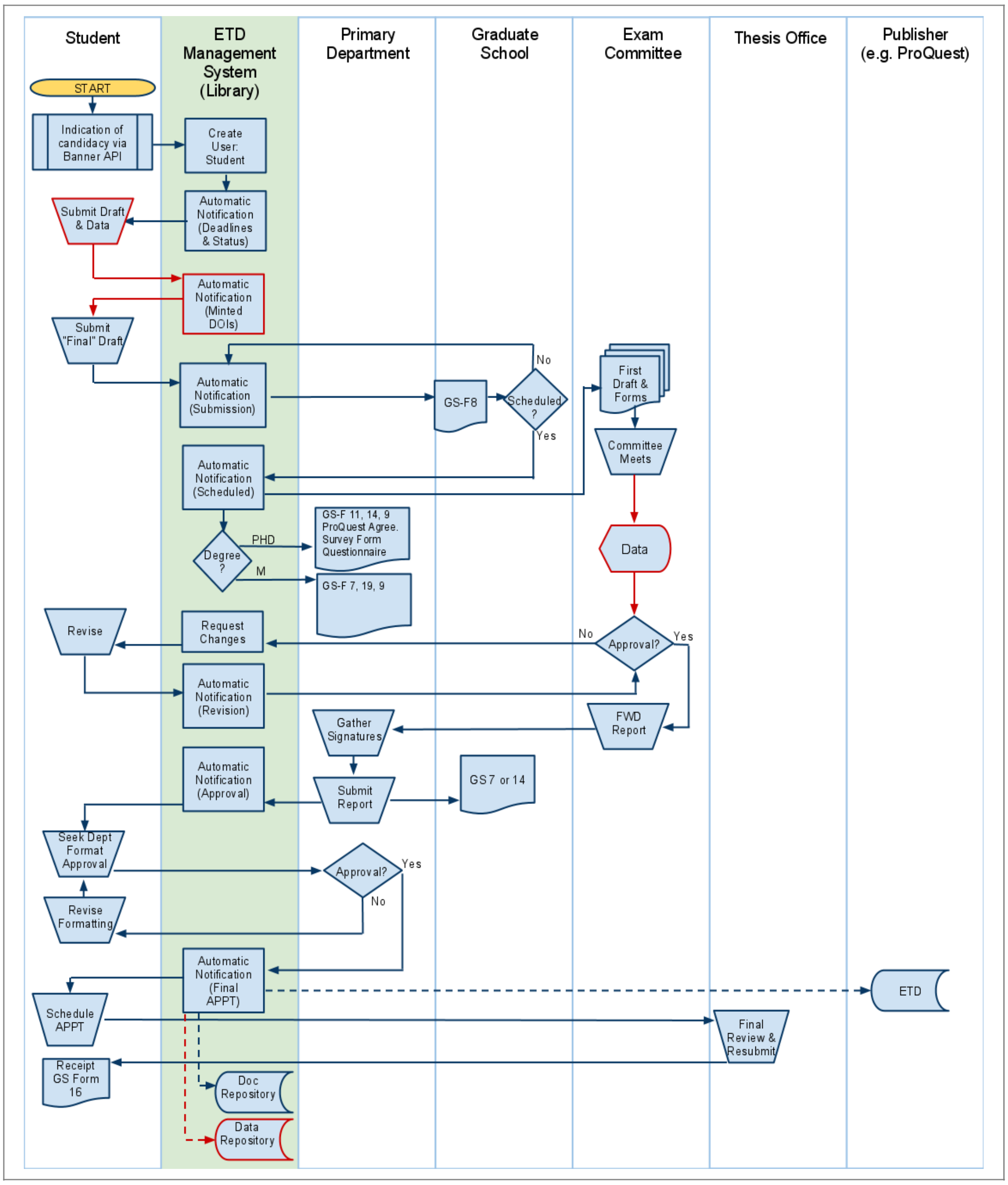
Student
ETD Management System (Library)
Primary Department
Graduate School
Exam Committee
Thesis Office
Publisher (e.g. ProQuest)
START
Indication of candidacy via Banner API
Create User: Student
Automatic Notification (Deadlines & Status)
Submit Draft & Data
Automatic Notification (Minted DOIs)
Submit "Final" Draft
Automatic Notification (Submission)
GS-F8
Scheduled ?
No
Yes
First Draft & Forms
Committee Meets
Automatic Notification (Scheduled)
GS-F 11, 14, 9 ProQuest Agree. Survey Form Questionnaire
PHD
Degree ?
M
GS-F 7, 19, 9
Data
Revise
Request Changes
No
Approval?
Yes
Automatic Notification (Revision)
Gather Signatures
FWD Report
Automatic Notification (Approval)
Submit Report
GS7 or 14
Seek Dept Format Approval
Approval?
Yes
No
Revise Formatting
Automatic Notification (Final APPT)
ETD
Schedule APPT
Final Review & Resubmit
Receipt GS Form 16
Doc Repository
Data Repository

Figure 4. Augmenting the ETD workflow at a typical American university to include datasets with dissertations.

### *Collection of Digital Assets*

Ensuring the acquisition of data along with the dissertation document during the deposition process further builds data collections and grows institutional assets. Serving these assets to the university and disciplinary communities via local repositories re-establishes the library as a primary service point for the university's intellectual output.

> "Institutional repositories – digital collections that capture and preserve the intellectual output of university communities – respond to two strategic issues facing academic institutions: 1) they provide a central component in reforming scholarly communication by stimulating innovation in a disaggregated publishing structure; and 2) they serve as tangible of an institution's quality, thus increasing its visibility, prestige, and public value."
> (Crow, 2002)

Finally, presenting the library as a data aggregator could encourage departments to develop data management plans for their students, and enables curators to discuss data curation activities with academic departments.

### *Availability of Data Citations*

Creating persistent identifiers upon early submission of data allows for in-document citation of data. Such citations to data can be a first step towards reframing datasets as scholarly objects. Data citations may help decrease the amount of data unnecessarily duplicated by helping to improve the discovery and referencing of existing data.

> "Dataset identification is a key element for allowing citation and long term integration of datasets into text as well as supporting a variety of data management activities. Also, to foster a culture of data integration, scientists need to be convinced that preparing their data for online publication is a worthwhile effort. It would be an incentive to the author if a data publication had the rank of a citable publication, adding to his reputation and ranking among his peers. To achieve the rank of a publication, a data publication needs to meet the two main criteria, persistence and quality. Whereas the latter is a very difficult concept that should be made part of the workflow of data integration in the data producers, data persistency is a rather simple problem."
> (Brase, 2009)

### *Open Access Opt-in*

Providing an option to self-archive dissertations and data in a local open-access (OA) repository can improve usage and reduce the time it takes for a document to become available online during a critical time in the student's career. A parallel deposition process could allow for deposition of document and data in an OA repository as well as a simultaneous export to a commercial database or embargo.

### *Ensuring Research Integrity*

The availability of data to be viewed alongside the dissertation document during the review and approval processes allows reviewers to validate claims and verify findings. The integrity of research is correlated with the transparency and reproducibility of applied methodologies and the data thus generated.

### *Improved Research Impact*

Capitalizing on improved accessibility, integrity, and citability of both the dissertation document and its supporting data synergistically improves students' research impact and provides the greatest return on a capstone achievement.

### *Preparation for Grant Writing*

The process of writing the dissertation is modeled, in part, to prepare students for the rigors of scholarly publishing and participating in the research enterprise. Mandates by large grant funding bodies, such as NSF and NIH, are beginning require data management and sharing plans. Including data in the dissertation process presents an opportunity to introduce or reinforce data management concepts to the student and prepare them to meet such funder mandates in their future, sponsored research.

## Discussion

In the same way that the university prepares new scholars by gradually scaling the scope of their achievement from the graduate thesis to multi-institutional research reports, scaling the unwieldy "data deluge" down to a tightly controlled and localized environment of graduate and doctoral students may help universities grapple with changes created by data-centric science. The opportunity to work with students and their data may lead to increased interest from faculty, and the resulting conversations may inform and support broader participation in data curation and the development of related policies and solutions for the campus.

New practices that support the curation of doctoral research data in the ETD environment will challenge existing roles. Students may feel that including data complicates the process, and they will require additional outreach and training. Exam committee members may not be accustomed to including data in their review and validation activities, and they may lack the time required for the additional work.

Similarly, graduate schools that feel pressure from an increasingly electronic and data-aware publishing environment will need to update thesis and dissertation guidelines in reaction to changes in philosophy and technology. Thesis and dissertation committees must survey faculty to discover if the current ETD workflow will adequately prepare students to publish in their respective discipline. This will mean that guidance committees and faculty must consider deposition mandates, university infrastructure, and data management requirements placed upon doctoral researchers now or in the future.

Thesis offices and libraries must align instruction opportunities and emerge with better-acclimated services for students publishing dissertations and data in an entirely digital environment. Because these organizations are typically service-driven, an efficient response must adapt existing librarian expertise to address issues with data collections, such as intellectual property, evolving citation standards, metadata, data literacy, digital preservation, data management practices, data standards, and disciplinary repository options. Thesis offices will also need to proactively adjust outreach and instruction to better administer a more efficient workflow. This might mean that the thesis office will spend less time corralling submissions and more time tracking and acting on changes in student status.

Libraries are uniquely positioned in that they possess an organizational capacity that is sufficient to collaborate and support the scale of effort required to re-engineer thesis and dissertation deposition workflows. Libraries should promote the expertise of faculty and staff who are familiar with the management of dissertations, and should join local conversations in an effort to align interest, resources and energy to leverage into new practices. For instance, libraries have become particularly adept at negotiating with vendors over complex publishing and intellectual property issues. Librarians are also familiar with multipart workflows, which surround the management of electronic documents, and have traditionally supported such workflows with localized and customized service models. Libraries will be required to engage with their host institution at differing levels and therefore must be equally prepared to take on roles of support or leadership. The role and practice changes described above will capitalize on existing personnel and proficiencies, and should require very little restructuring.

Student-authors are the primary beneficiaries of curating dissertation data. Students profit from changes to an outdated workflow that better reflect changes to scholarly communication and digital media. Student-authors enter the workforce touting a demonstrable application of a data management strategy along with credible, citable scholarly data.

In conclusion, the possible benefits of new practices of including research data in the dissertation process can holistically add value to both the process and its resulting product. As Lippincott and Lynch (2010) advise, empirical data is needed to substantiate that students are willing to share data and that demand for such services exists. The emerging cultural shift towards increased data-sharing is complemented and enabled by new cyberinfrastructure. What is less clear, from an institutional perspective, is who will share what data and how data collections will effectively take shape. Student-authors who wish to deposit ETD datasets represent a tractable portion of this challenge. Encouraging institutions to acquire and steward ETD datasets can be supported by a strong value proposition and new practices that are well-aligned with the mission of the university and the academic library.

## References

Borgman, C.L. (2007). *Scholarship in the Digital age: Information, Infrastructure, and the Internet*. Cambridge, MA: MIT Press.

Brase, J. (2009). DataCite: A global registration agency for research data. *In the proceedings of the Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology*. Beijing, China.

Crow, R. (2002). The case for institutional repositories: A SPARC position paper. [Discussion Paper] *Scholarly Publication and Academic Resources Coalition.* Washington, D.C.

Lippincott, J.K. & Lynch, C.A. (2010). ETDs and graduate education: programs and prospects. *Research Library Issues, 270*.

Research Information Network. (2008). *Stewardship of Digital Research Data: A Framework of Principles and Guidelines*.

Swan, A. & Brown, S. (2008). *Skills, role and career structure of data scientists and curators: Assessment of current practice and future needs.* JISC: UK.

Ubogu, F.N. & Sayed, Y. (2008). Management of research data in ETD systems. *Proceedings of the 11th International Symposium on Electronic Theses and Dissertations*. Aberdeen, Scotland.