Current Issues and Approaches to Curating Student Research Data

by Andrew Creamer

Research Data Access and Preservation

EDITOR'S SUMMARY

The 2015 RDAP Summit hosted a panel discussion to address issues and challenges of curating students' research data and digital scholarship, bringing together librarians dealing with various aspects of digital curation and management. Few institutions have a policy on handling student research data, theses and dissertations. Obstacles include insufficient know-how, time, incentive and resources. Without guidance, data files vary widely in format and quality. Yet, as one panelist pointed out, without definitive standards, data librarians must contend with a messy reality and should be reasonable in their expectations for students. Planning for improvements must start by engaging students and advisors throughout the process of data creation. Panelists agreed on the need to further explore topics such as intellectual property, ownership and security, data quality, staff skill building and outreach to stakeholders.

KEYWORDS

data curation

dissertations

institutional policy

intellectual property

data security

graduate students

college students

Andrew Creamer is scientific data management specialist, Brown University Library. He can be reached at Andrew_Creamer<at>brown.edu.

While there is much consistency in the way that university libraries have collaborated with their undergraduate colleges and graduate schools to archive and/or publish their students' electronic theses and dissertations (ETDs) online, there is little consistency in the way that universities are handling the archiving, curation and publication of their students' research data and digital scholarship that underlies these ETDs. The Networked Digital Library of Theses and Dissertations (NDLTD) found that most of their member institutions had no policy on the stewardship of student data related to ETDs, and in the cases where students provided data with their ETDs, the libraries treated them as supplementary files [1]. Several years later, Collie and Witt [2] argued that libraries could seed their fledgling repositories with student research data: "Dissertation datasets represent 'low-hanging fruit' for universities who are developing institutional data collections." (p.166). Yet five years later, university libraries are still figuring out the best approaches for building collections of data related to ETDs. The 2015 Research Data Access and Preservation (RDAP) Meeting's panel "Current Issues and Approaches to Curating Student Research Data" explored these issues and challenges.

This panel was unique in that it was a hybrid made up of both invited panelists (information professionals who had conducted and presented recent research on the theme) and selected panelists (information professionals who had submitted proposals for papers that demonstrated current and innovative approaches.) The panelists included a digital curation librarian working within his organization to plan a path forward for curating students' digital scholarship; a research data management librarian and a subject librarian working together to explore a workflow with their campus partners to inform and educate students about submitting their ETD data

CREAMER, continued

and digital scholarship; a library administrator who conducted an assessment of the supplementary data files that were submitted by students with their ETDs; and a digital repository librarian who is sorting out the complexities of ingesting, modeling and describing student research data files.

Aaron Collie, head of digital curation at Michigan State University Libraries, opened the panel with his presentation, "Building Organizational Capacity for Data Collections Using Electronic Theses and Dissertations." He reflected on the article he published in 2011 with Michael Witt at Purdue Libraries that I referenced above. In addition to characterizing student ETD data as "low hanging fruit," Aaron and Michael saw the curation of students' ETD data as a scale model of the scholarly communication lifecycle, valuable collections that universities should pursue, archive and publish. So why did so many libraries fail to pursue student ETD data and use student research data to seed their fledgling data repositories? Aaron sees organizational capacity as the largest obstacle to building these ETD data collections. He described three organization-level challenges regarding digital curation that libraries need to address: people not knowing how to do the work, not enough time or incentive for people to learn and insufficient resources. According to Aaron, there is an unrealistic expectation held by academic library administrators that they can fit all digital responsibilities and expertise into just one fulltime employee (FTE), adding, "I think digital preservation is a strategic direction that is too often operationalized as an individual responsibility and skill set." He concluded his talk by sharing the progress he is making at his library to carry out a three-year strategic plan, laying the groundwork for a collaborative approach to digital curation that will put the policies, people and technologies in place to build organizational capacity.

While Aaron's paper highlighted the challenges that we may encounter inside the library, Dianne Dietrich, physics and astronomy librarian, and Wendy Kozlowski, data curation specialist, both from the Cornell University Library, described in their presentation, the challenges that university libraries face engaging and educating students as well as coordinating with their campus partners as they position themselves to curate and archive their students' ETD data. Like many institutions, the Cornell University Library collaborates with Cornell's Graduate School to ingest, archive and publish their students' ETDs. In this model, there are several nodes of communication along a student's path to writing his or her dissertation and graduating where students are informed about the guidelines for formatting and submitting their ETDs. For example, students commonly encounter theses and dissertation advisors, graduate program and administrative staff in their departments, and staff in the graduate school and library. Wendy and Dianne described the big challenges of making sure that the stakeholders at each of these nodes are informed about any changes to ETD submission guidelines to include submitting data with ETDs and for getting the word out to students and educating them about managing and sharing their research data. When and how many times should we engage students to prepare them to deposit their data as they move towards submitting their ETDs? In addition to communication and workflow, Dianne and Wendy also presented us with the policy challenges related to ingesting and describing data, such as often overlooked issues related to defining ownership, handling embargo situations and licensing and versioning.

Sarah Shreeves, associate dean of digital strategies at University of Miami Libraries, presented "Supplemental Files for ETDs: Diversity, Documentation and Data," shifting the focus of the panel from preparing the library, campus partners and students for curating, archiving and publishing student ETD data to looking at lessons learned from libraries that have begun building student ETD data collections. She shared the results of her assessment of the data files that had been optionally submitted by students with their ETDs over several years at the University of Illinois at Urbana-Champaign (UIUC). Students at UIUC have the option to submit data and digital scholarship as supplementary files along with their ETDs; the files are considered appendix items and UIUC ETD submission guidelines inform students about their option to deposit and describe the process and rules for adding these digital objects to their appendices and depositing these as supplementary data files.

Between 2010-2014 she found there were 6,472 ETDs submitted by UIUC students to the library. Of these students, 129 or roughly 2% submitted supplementary files along with their ETDs. Of this sample, 94 were students in a program in the sciences, 19 were in a program in the arts and humanities

and 16 were students in a social sciences program. Eighty-eight students submitted between one and five files, 25 students submitted six to 20 files and 16 students submitted 21 or more files (she found a few instances among these 21 students with 1000+ files). The types of data files submitted included image, text, hypertext, tabular, sound and film files, among others. Roughly 64% could be characterized as data, 15% as code, 7% as data and code and the remainder as protocols and other digital material.

Sarah's assessment offers information professionals many insights on curating student ETD data. For example, her findings help us to get a rough idea of the percentage of students a library could expect to submit supplementary data with their ETDs if the library were to add this option to its ETD submission guidelines, and it helps us to predict the major file types and average number of files. Her research also helps us to ask important questions about the best ways to curate and describe student ETD data. Should there be more oversight over the documentation quality and quantity students provide with their datasets? Should these digital objects receive their own record and metadata, and if so, what are the best ways to express the relationships among these objects and the ETD? She also highlighted the challenges that we face in preserving student data. For example, can we make the same archival/preservation commitments to supplementary data files that we do for the pdf file of the ETD?

To conclude the panel, Steve Van Tuyl, data and digital repository librarian at Oregon State University, presented "Treating Data Like Data: Unifying Data Processing Workflows for Datasets in the IR," tempering the expectations for libraries transitioning to ingest students' ETD data files. He presented a sobering assessment of the student data submitted with ETDs in the ScholarsArchive@OSU. Of a sample of 93 ETDs with related data in OSU's repository, 45% were Excel files (30% of which had macros, charts and/or linked to other data), 22% were image files and 25% were document files. The remainder of the data included text, database and/or statistical software files, of which 23% were code (and 15% of these executable files), and 12% of the files were metadata. Of the 93 ETDs with data, 30% were unknown, un-operable and/or obsolete; and 3% of the ETDs were missing data files from what was listed among their manifests.

In addition to echoing Sarah's concerns about the archival and preservation commitments library repositories are making to depositors and whether we can keep these same commitments for their data, Steve also skillfully made the point that we cannot realistically expect data management perfection from students and, that given the great diversity and uniqueness of student research, it is difficult for repositories to have one-size-fits-all, definitive standards for description and curation quality. After presenting us with the messy reality of curating, ingesting and publishing student ETD data (including sharing with us a humorous document file submitted by a student along with his ETD containing a list of songs that he had listened to while collecting field data), Steve encouraged us to walk away from the panel ready to "treat data like data," treating the curation, ingest and publication of student research data related to ETDs in the same ways we would approach student research data not related to an ETD. He characterized this unified approach as one of iteration and encouragement, working with students to get them to try their best to meet the library's ingest and description standards and, if necessary, tattling and getting the student's advisor(s) involved in the process or communication chain.

The panel successfully identified several important issues and obstacles preventing libraries' curation of students' ETD data. Issues warranting further exploration and discussion, hopefully in future RDAP meetings, include intellectual property, ownership, copyright and licensing, data management and description quality, privacy and security, communication workflows with students and campus partners, building staff expertise and organizational capacity, embargo policies for student research data, and archival and preservation concerns. The panelists' consensus is that student data collections are worth pursuing and have much value for the public and research enterprise, but libraries interested in harvesting this so-called low-hanging fruit need to be prepared that this fruit may be a little higher on the vine than previously thought. ■

Resources on next page

Panelists Slides:

Aaron Collie
www.slideshare.net/aaroncollie1/building-organizational-capacity-for-data-collections-using-electronic-theses-dissertations

Dianne Dietrich and Wendy Kozlowski
www.slideshare.net/asist_org/cornell-etd-workflowrdap2015

Sarah Shreeves
www.slideshare.net/asist_org/rdap-15-47572605

Steve Van Tuyl
www.slideshare.net/asist_org/rdap-15-treating-data-like-data-unifying

Resources Mentioned in the Article

[1] Ubogu, F.N., & Sayed, Y. (2008, June). Management of research data in ETD Systems. Paper presented at ETD 2008: 11th International Symposium on Electronic Theses and Dissertations, The Robert Gordon University, Aberdeen, UK. Paper available at http://www4.rgu.ac.uk/etd/programme/page.cfm?pge=45695

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