EDITOR'S SUMMARY

The mandate by the White House Office of Science and Technology Policy to improve access to federally funded research makes responsible data curation by academic libraries more important than ever. Libraries should provide curation services covering not only deposit and access but also appraisal, description, metadata creation, format transformation, archiving and preservation. A data curation pilot project by the University of Minnesota Libraries demonstrated an effective workflow model to curate research data and facilitate its reuse. Five research datasets were selected for the pilot, each analyzed by a separate team of library, information and technology professionals. The pilot provided groundwork for more robust data curation services and identified necessary support infrastructure.

KEYWORDS

data curation
research data sets
academic libraries
library and archival services
flow charting
information reuse

Developing a Data Curation Service: Step #1: Work With What You've Got

by Lisa R. Johnston

Academic libraries offer a wide-range of research data services. We consult on data management plans, educate students and faculty on the best practices for organizing, storing and maintaining long-term access to digital data (and in some cases, *provide* that long-term access), and we advocate for better, user-centered services and campus collaborations in support of all the disciplinary units that we serve. But now we must do more. Last year the Office of Science Technology Policy in the White House signaled a renewed interest in making the results of federally funded research more publicly accessible – no small feat. For example, at the University of Minnesota, federally funded projects account for over 68% of the several hundreds of millions of grant dollars received in 2012. Therefore, all of the digital research data generated from these grant dollars would need to be publically accessible for search, retrieval and analysis in the near future. Researchers already have options to fulfill these requirements such as figshare or data journals; however, it may be in the best interests of academic libraries to provide our own brand of support, lest the expensive digital data assets of our institutions be forgotten on unreliable publisher websites or in start-up disciplinary repositories with no plans for sustainability.

Lisa R. Johnston is associate librarian at the University of Minnesota - Twin Cities. She leads the libraries' research data management and curation initiative and is co-director of the University Digital Conservancy, the University of Minnesota's institutional repository. Her areas of research focus are scientific data curation, open access models for research publications and data, and educational approaches to training faculty, staff and students in data information literacy skills and best practices. She can be reached at ljohnsto<at>umn.edu.

This support is where data curation fits in. Curation services might go beyond deposit and access to the data. Here, a role for libraries might include appraisal, ingest, arrangement and description, metadata creation, format transformation, dissemination and access, archiving and preservation of digital research data. To better explore these roles, the University of Minnesota Libraries ran a data curation pilot in 2013. The results of the pilot include a workflow model for how the library might curate our researchers' digital data for public access and, more importantly, reuse. This project allowed us to test our current capacities in order to move forward on developing more robust data curation services as well as the technological infrastructure to support them (see the full report and the curated pilot data sets at http://purl.umn.edu/160292).

Our pilot began with three goals in mind:

1. Solicit, select and curate five pilot research datasets for discovery and reuse
2. Research and develop a data curation workflow utilizing existing university capacities and resources
3. Conclude with a summary report describing the successes and shortcomings of this approach.

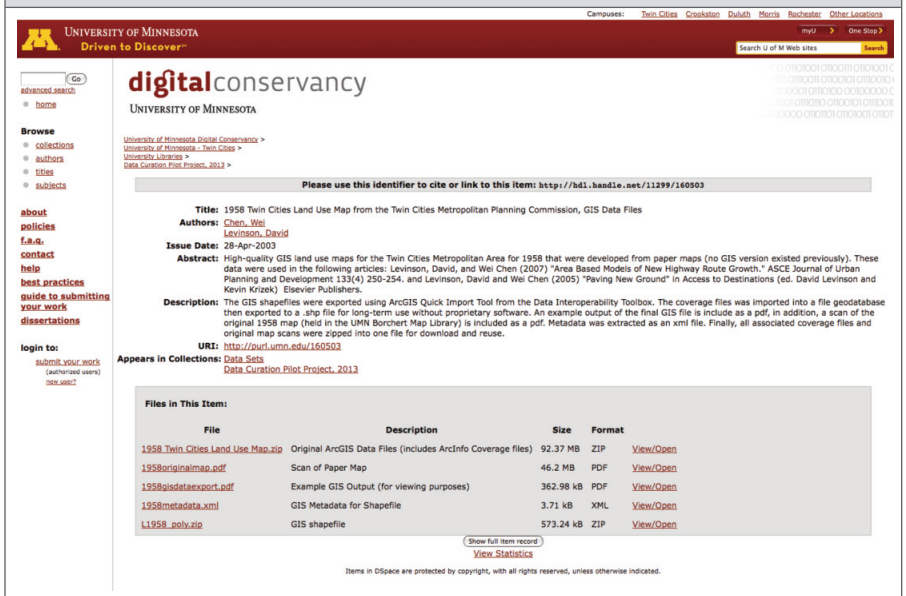
First, we needed participants. Our call for proposals went out to faculty, students and research staff at UMN. We advertised in a number of ways, and thanks in a large part to the help of subject liaison librarians and staff, our call received 457 web visits from July 2nd - November 27, 2013. Sixteen proposals were received from a useful cross-section of disciplines and data types that included at least one proposal from every major college on campus.

Working with our existing infrastructure, our DSpace-based institutional repository (conservancy.umn.edu), we knew that the pilot would have certain technical limitations that would affect what data we could curate. To narrow down the proposals, we evaluated the submissions against the criteria that were well publicized within the call. To further manage researcher expectations, we also conducted in-person interviews with the data authors in order to verify that their data would fit within our existing capacity. We selected five datasets that met our criteria and that were a good fit for the pilot.

The process of curating the data involved the help of many experts in the libraries and on campus. To enable a broad range of staff to engage with the pilot data and develop their knowledge base, we sponsored Digital Curation Sandbox, a half-day event that brought nearly 30 staff together from a cross-section of library science and information professionals. For each of the five datasets we assigned a digital technology expert, a cataloger/metadata expert, a subject librarian and an archivist/curator to take a deep-dive look at the particular curation needs of the data. In addition to these skills sets, each team had a facilitator who brought data-specific knowledge to the group. The group discussed existing curation workflows, analyzed the five datasets and drafted a treatment workflow for each. Following the event, we created a generalizable workflow model that included all of the questions that a data curator might consider when archiving digital research data in the library. Finally, the workflow and treatment actions were carried out, and the five datasets selected for the pilot were curated for reuse at http://purl.umn.edu/160292 (see Figure 1 for an example).

Feedback from the faculty was very positive and anticipated that this service might satisfy the upcoming requirements from federal funding agencies. For example, one participant commented, "With *data management* being critical in NSF proposals (and likely other funding agencies), it would be great to have a centralized service

FIGURE 1. An example screenshot of a curated dataset in the pilot.



like this to which data could be submitted for curation, storage and public access...this is a big need." Furthermore, our service includes persistent URLs that will continue to connect the research data to reuse through formal citations and download statistics.

The lessons learned were twofold: it worked, and we have more work to do. First, the success of the pilot is that we now have a model workflow for data curation. And although this project revealed a strong demand for domain-specific software and knowledge and that the researcher-provided documentation for data was not always sufficient, overall we found that our service is scalable to a variety of data types and disciplines. The next step will be to develop a data curation service that fits within the researcher workflow and expands our existing capacities. Using what we've learned from this pilot, Step #2 should be a piece of cake. ■

CONTENTS

TOP OF ARTICLE

< PREVIOUS PAGE