An Applied Approach to Data Curation Training at the Inter-university Consortium for Political and Social Research (ICPSR)

Jared Lyle, Mary Vardigan
ICPSR, University of Michigan
Ann Arbor, MI, U.S.A.
{lyle, vardigan}@umich.edu

Jacob Carlson
Purdue University
West Lafayette, IN, U.S.A.
jakecarlson@purdue.edu

Ron Nakao
Stanford University
Stanford, CA, U.S.A.
ronbo@stanford.edu

***Abstract*—ICPSR recently developed two new training initiatives in digital curation: a week-long applied data curation workshop where participants learn the theories and methods of data curation using the ICPSR "processing pipeline" as framework, and an ongoing virtual working group of data librarians that discusses similar core data curation topics while giving participants independent access to curate their own data using ICPSR's processing environment and tools. This paper discusses the background, structure, and lessons learned from these new training initiatives.**

Keywords—Digital curation, data curation, training, curriculum.

I. Overview

The Inter-university Consortium for Political and Social Research (ICPSR), a research center in the Institute for Social Research at the University of Michigan and the world's largest archive of social science data, recently developed two new training initiatives in digital curation. The first initiative is a week-long applied data curation workshop offered as part of the ICPSR Summer Program in Quantitative Methods, where participants learn the theories and methods of data curation using the ICPSR "processing pipeline" as framework. The second initiative is an ongoing virtual working group of data librarians that discusses similar core data curation topics while giving participants independent access to curate their own data using ICPSR's processing environment and tools. This paper discusses the background, structure, and lessons learned from these new training initiatives.

II. Data Curation Workshop

As data multiply in sheer quantity and become increasingly important in the research process, the demand for data curation knowledge rises. What are the best practices for curating research data? How does one apply them to daily practice? What tools can assist in curation efforts? In 2011, ICPSR began planning a data curation workshop to address these questions.

A. Background

The workshop was intended for individuals interested or actively engaged in the management and curation of research data, particularly data scientists, data managers and analysts, librarians, archivists, and data stewards and curators. The initial goal of the workshop was to "raise awareness about the benefits of life cycle principles for data management, including how to create, comply with, and evaluate required data management plans, how to encourage and trace re-use, and how to manage data from its inception through archiving and beyond."

We believed, and continue to feel, that ICPSR is uniquely positioned to offer a course on data curation. First, ICPSR plays a central role in many social science data curation standards and activities, including serving as the home office for the Data Documentation Initiative (DDI) and as a founding member of the Data Preservation Alliance for the Social Sciences (Data-PASS). DDI has become an international standard for metadata in the social sciences. ICPSR and many other data archives use the DDI XML to document information about the data in our repositories; the ICPSR online catalog is also built on DDI metadata, allowing structured searching across the entire repository at the variable- and even the value-level. Data-PASS is a voluntary partnership of organizations created to archive, catalog, and preserve data used for social science research. The Data-PASS partners collaborate on best practices for data archiving and have a shared digital preservation strategy.

Second, ICPSR has established workflows for curating, preserving, and providing access to data. These workflows, described as the "ICPSR Pipeline Process" (Fig. 1), have been developed and refined over 50 years of archiving more than 8,000 research collections from across all social science

disciplines, and are informed by the Reference Model for an Open Archival Information System (OAIS) for the preservation of digital objects as well as other community-based best practices. The workflow segments, which are broken into digestible portions, make it easier for students to follow and learn curation processes.

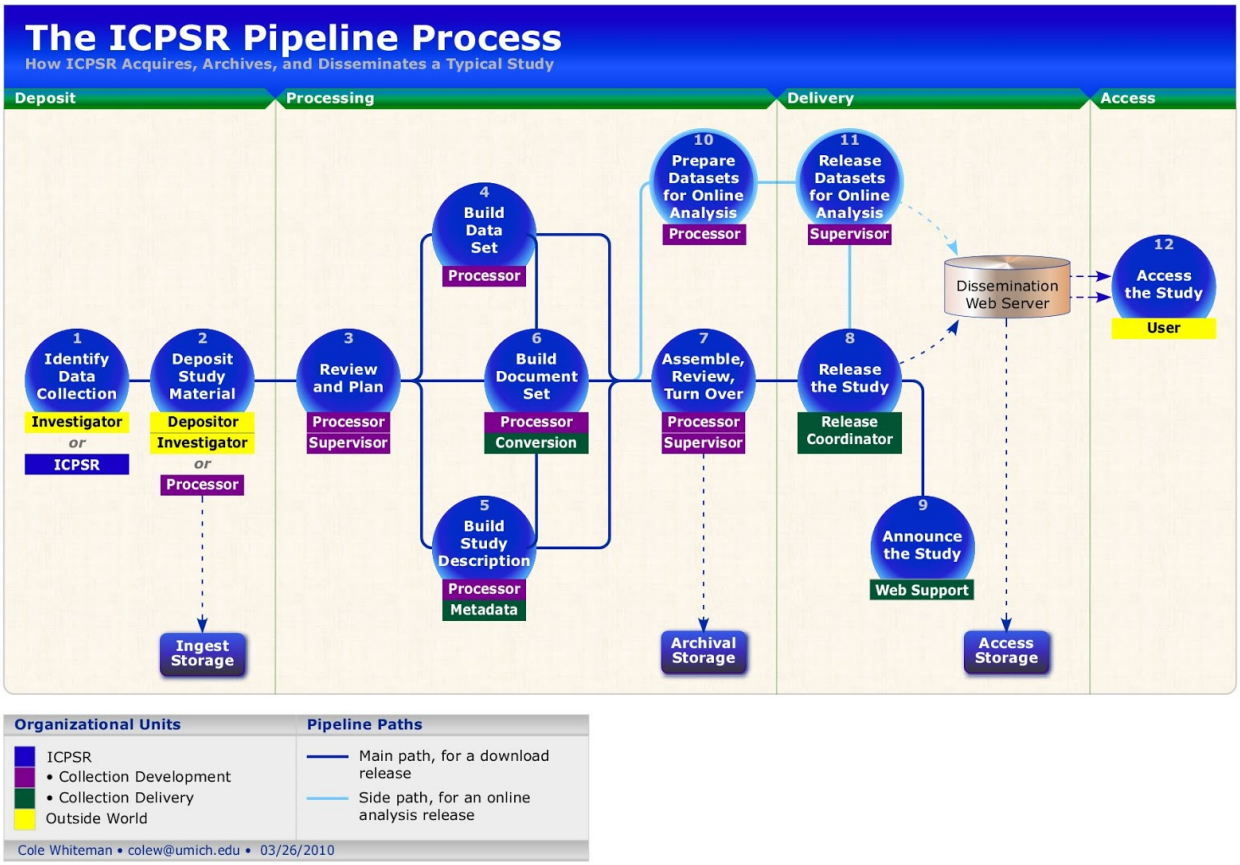

Figure. 1. ICPSR Pipeline Process.

Third, ICPSR has an established Summer Program in Quantitative Methods that offers more than 70 courses every summer. The program provides an instructional infrastructure readily accessible for curation instruction. For the past several years, for instance, we have offered a course for data librarians called "Providing Social Science Data Services: Strategies for Design and Operation." More recently, a course on confidential data, "Assessment and Mitigation of Disclosure Risk in Data: Essentials for Social Science," was offered.

Finally, ICPSR is committed to global leadership in the area of digital curation, especially through instruction. Direction 1 of the ICPSR Strategic Plan reads: "Through global leadership and strong partnerships, set standards for excellence in data curation and in the ethics of data access and protection for the social sciences and related disciplines." The ICPSR Council, which is elected by the Consortium membership and provides overall guidance, strongly encourages our participation in initiatives to promote digital curation. We are eager to share our experience and knowledge. We also recognize and appreciate the benefits from the course: increased connection with front-line curators, improved understanding of the needs and workflows of the community, and new opportunities to influence the curation of data further upstream in the data lifecycle (i.e., closer to the original production of the data).

B. Structure

The workshop, titled "Applied Data Science: Managing Research Data for Re-Use," was held July 23-27, 2012 in Ann Arbor, Michigan. ICPSR teamed with the University of Michigan School of Information to host the workshop. The core instructors were Mary Vardigan and Jared Lyle from ICPSR,

Kathleen Fear from the UM School of Information, and Jake Carlson from Purdue University.

Twenty-five participants attended, representing diverse institutions from the United States and Canada, as well as a range of disciplines, including engineering, chemistry, physics, the physical sciences, and the social sciences. Participants came to the workshop with a wide variety of interests. Many participants were interested in broad-based training. Others were establishing or expanding their own repositories and needed "shovel ready" plans for curating data. Still others came with very specific questions in mind, such as how to manage confidential data or how to address copyright questions.

The workshop was grouped into five themed days that followed an ICPSR dataset across the data life cycle through creation, deposit, data processing, dissemination, preservation, and reuse [1]. Day 1 provided an overview of the research life cycle stages and data curation. Day 2 covered data management planning and acquisitions. Day 3 highlighted metadata. Day 4 covered data processing, confidential data management, and repository requirements. Day 5 addressed dissemination, preservation, and tracking reuse.

Throughout the workshop, guest speakers provided insight on a wide variety of curation topics, such as managing video data, geospatial data, provenance, and repository assessment. Case studies and hands-on curation activities designed to help participants apply the material presented were woven throughout the workshop. Examples of hands-on activities included creating study- and variable-level metadata, reviewing unprocessed data within Google Refine, and checking a dataset for confidentiality issues.

C. Lessons Learned

Overall, the participants had very positive comments about the workshop. Most rated it as "exceptional" or "above average" when compared to other graduate level courses they have taken.

Expertise, breadth of subject material, and applicability were main strong points mentioned in the course evaluations. "This workshop provided an insider's view of the data curation process," wrote one participant, adding that "having presenters that specialize in key parts of the process was very valuable." Another participant noted, "The 'pipeline' served as an excellent framework." Yet another appreciated "the hands-on aspects of the course and the various print-based handouts."

As this was the first time this workshop was offered, we were particularly active in gathering feedback. We surveyed the participants at the end of each day of the course and applied the feedback we received to adjust the course pace and content for the subsequent days. At the end of the course, the Summer Program also conducted an official, proctored evaluation. This feedback now informs our future development. Some of the shortcomings of the workshop that were identified, along with plans to address them, include:

1) *Covering Too Much Content:* While many participants enjoyed the broad range of curation topics discussed, we also heard comments like "Almost too much material...difficult to digest in short space of time" and "Too many briefings that tried to cover too much material in a short presentation." We intend to remedy this by discussing fewer topics but diving more deeply. Instead of discussing, for instance, the many possible data types in detail, leaving small chunks of time to each, we intend to provide a quick but broad overview of the subject and then spend quite a bit of time discussing the specifics of one or two examples with hands-on activities.

2) *More Discussion and Collaboration:* A few of our days were especially long on lectures and short on discussion. We wanted to impart as much of our knowledge as possible, along with that of our invited experts. What the participants really wanted was a mixture of learning from experts and discussion among their peers. "Would have liked more opportunity to share challenges/solutions with participants," wrote one attendee. Another said, "A forum for discussing individual situations, problem-solving suggestions for next steps, etc. would be helpful." As a solution, we are building more discussion time into the schedule, including structured thirty-minute blocks each morning and afternoon and a longer lunch break. We are exploring building peer-to-peer collaboration into the exercises as well. We intend to better capitalize on the expertise and knowledge that many workshop participants bring with them.

3) *Applied, Applied, Applied:* Though we tried to pair applied examples and exercises with each lecture, workshop participants wanted more. Many participants mentioned there are quite a few opportunities to learn about curation, but few chances for hands-on active learning and interaction. While we feel applied interaction is one of the strengths of our workshop, we are looking to fine-tune the exercises that worked well and add others.

4) *More Science in the Curriculum:* As a social science data archive, the curation material that we discussed naturally emphasized methods and content from just one slice of the research data spectrum. Our participants recognized the applicability of social science data curation to all types and formats of data, and we did include some examples from the 'hard sciences.' That said, the participants wanted to "cover a wider array of data types and the unique management issues for each." While we will continue to highlight our own data and methods from the social sciences, we can attempt to better

diversify the types of data covered in the exercises and discussions. One option, for example, would be to offer participants a choice of the types of data to work with during exercises.

III. Data Curation Working Group

Shortly before the start of the summer data curation workshop, ICPSR discussed with Ron Nakao, Stanford University, some possible mechanisms to provide more hands-on, localized data curation training to librarians, especially the Official Representatives at member institutions who assist faculty, staff, and students with ICPSR resources. Many librarians have limited experience with data management and curation. In addition, as budgets are increasingly tightening, librarians may not have the chance to travel for week-long training. Even the more experienced data librarians do not have the tools or resources that ICPSR can provide. Although multiple venues exist to meet and discuss data curation topics -- from listservs to conferences -- few opportunities arise for data curators to engage in personalized but collaborative hands-on work using the tools of an established domain repository.

A. Background

We proposed a virtual data curation working group where participants would apply curation theories to practice through actual data processing, interact with and ask questions of other data specialists within a working environment, and gain first-hand experience using ICPSR's internal tools and procedures for curation. The course would last approximately four months, with one virtual meeting of 1 ½ hours approximately every other week.

ICPSR would benefit from the group as well. By opening our processing environment and tools to outsiders, we would learn more about the tools and services data librarians want and need, and the suitability of expanding the use of ICPSR's own curation tools to a broader community. This interest coincides with our work in an IMLS National Leadership Grant (LG-05-09-0084-09) to investigate tools and services to assist librarians with specialized tasks in the archiving and dissemination of social science data. Another benefit of the working group would be that more data would be curated and archived, benefiting the ICPSR membership and the entire social science community.

B. Structure

The working group first met -- virtually -- in September 2012. Participants hailed from Emory, Duke, UCLA, and UC Berkeley, along with Jared Lyle from ICPSR as facilitator and Ron Nakao as the chair. Participants received access to the ICPSR secure processing environment and brought their own data to curate. Bi-weekly discussions focused on topics similar to those found in the summer data curation workshop: acquisition (gathering information from the data producer, legal agreements, and appraisal), review (quality and disclosure review), processing (data cleaning, insuring data integrity, and quality checking), metadata (standards, and variable- and study-level metadata), dissemination (final packaging, delivery mechanisms), and preservation (policies and actions).

At this time, the working group is still active. Participants have access to the ICPSR secure data processing environment through September 2013.

C. Lessons Learned

As in the workshop, participants were generally excited to be learning about and practicing data curation. "This was a fantastic opportunity," wrote one participant. "The most useful/informative aspect has been applying the ICPSR's workflows and practices to an actual data collection and seeing what's involved in getting the data in sync with those workflows and practices."

Since the group is ongoing, and since group members are still processing and curating their data, we anticipate learning more about the successes and challenges of this training format. In the meantime, we offer a few in-progress lessons learned.

5) *Bring Your Own Data:* All working group participants brought their own data to process and curate. As a result, the participants were highly invested and motivated; the questions and discussions raised were timely and relevant rather than purely theoretical.

6) *Hands-on Activities Were Key:* Similar to bringing their own data, hands-on activities using ICPSR's processing environment and tools helped the group members understand and experience the core work of curation instead of just talking through what can seem like generalized concepts. As one participant mentioned, "...The real work was with going through the data and documentation and seeing things like discrepancies in variable names and the need to flesh out citations to make them more informative. That was both interesting in its own right and illuminating to provide a sense of what data curation actually consists of in practice."

7) *Scheduling Issues:* Virtual meetings have distinct benefits, including saving time and money, and allowing participants to practice methods and tools in between group discussions. However, many in our group experienced one big drawback: scheduling conflicts. As on member lamented, "I guess the only real 'problem' with the group was that scheduling/timing issues were such that we had to do a lot of the work during the semester, when other demands on my time made it hard to focus on the project in a sustained manner." Another member

expressed similar frustration. “Unfortunately, my schedule shifted pretty dramatically this semester, and it was often difficult to fit in the call and prep work needed to make the call most useful.” By not leaving their physical job work environments, it was increasingly challenging for participants to carve curation time away from the everyday job demands and expectations.

IV. Summary

As part of ICPSR’s commitment to global leadership in the area of digital curation, especially through instruction, we will offer the data curation summer workshop again in July 2013. Likewise, the data curation working group is running through September 2013.

We see continued demand by professionals to learn about curation, especially through applied learning, and feel we can play a role in helping educate the research and digital curation community through teaching and discussing the curation experiences and processes that have shaped our 50 years as a data archive. As we do this, we recognize and appreciate the benefits: increased connection with front-line curators, improved understanding of the needs and workflows of the community, and new opportunities to influence the curation of data further upstream in the data lifecycle.

Acknowledgment

We wish to acknowledge Nancy McGovern, ICPSR Digital Preservation Officer from 2007-2012, who played a leading role in developing the initial draft and goals of the workshop. We thank the workshop and working group participants for their feedback and participation. We also thank Dan Meisler for his edits. IMLS National Leadership Grant (LG-05-09-0084-09) supported data curation working group activities that identified services to assist with archiving and disseminating social science data.

References

[1] J. Carlson, K. Fear, J. Lyle, and M. Vardigan. “Applied Data Science: Managing Research Data for Reuse,” Workshop Syllabus. ICPSR Summer Program in Quantitative Methods, July 2012. http://www.icpsr.umich.edu/files/sumprog/biblio/2012/Applied%20Data%20Science%20Managing%20Research%20Data%20for%20Reuse.pdf.