Learning to Curate

by Jennifer Doty, Joel Herndon, Jared Lyle and Libbie Stephenson

EDITOR'S SUMMARY

Three data specialists reviewed their experiences learning about and applying the Inter-university Consortium for Political and Social Research (ICPSR) processes and tools for curating research data. A small virtual community discussed curation theories on data acquisitions, review, processing, metadata and dissemination and shared progress implementing the ICPSR workflows and tools. Curators at Duke University, dealing with data on political donors, found processing obstacles from incomplete and mismatched data and faced confidentiality questions. At Emory University, gaps in coded data on home schooling practices revealed problems arising from lack of preplanning for long-term archiving and research and the need to clarify data needs for later re-use. By applying ICPSR processes, curators at UCLA's Social Sciences Data Archive were able to improve their workflow and understand the work necessary for open archiving. All participants gained from the opportunity to practice ICPSR curation practices, realized the resource demands and saw the value librarians can provide by consulting with faculty on data management and preservation.

KEYWORDS

data curation
research data sets
archival science
barriers
training

Jennifer Doty is research data librarian at Emory University. She can be reached at jennifer.doty<at>emory.edu.

Joel Herndon is head of data and GIS services at Duke University Libraries. He can be reached at joel.herndon<at>duke.edu.

Jared Lyle is director of curation services at ICPSR. He can be reached at lyle<at>umich.edu.

Libbie Stephenson is director of the UCLA Social Science Data Archive. She can be reached at libbie<at>ucla.edu.

While many 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. During the latter half of 2012, a working group of data specialists sought to address this need by meeting regularly (via online meeting) to discuss data curation topics and apply them to actual processing of data using the internal workflows and tools of the Inter-university Consortium for Political and Social Research (ICPSR). Working group members included Ron Nakao, Stanford University; Jared Lyle, ICPSR; Rob O'Reilly and Jennifer Doty, Emory University; Joel Herndon, Duke University; Libbie Stephenson, UCLA; and Jon Stiles, UC Berkeley.

Virtual sessions combined hands-on work and discussions of the theories behind the practice. Participants curated their own data and shared their own novel methods for improving the data curation experience. The following topics were included:

Acquisitions

- Gathering and collecting information from the data producer (how much should/will they contribute?)
- Legal agreements
- Appraisal
- What to keep, and for how long?

Review

- Quality review - are the data complete, accurate, and well documented?
- Disclosure review - is there sensitive or private information?
- Creation of a plan of attack

Processing
- Data cleaning
- Insuring data integrity
- Quality review - is the final package self-contained?

Metadata
- Standards overview
- Variable level metadata
- Study level metadata

Dissemination
- Final packaging and review
- Workflows
- Preservation policies
- Web delivery

Following are the perspectives from three of the working group participants.

Learning to Curate @ Duke - Presidential Donor Survey 2000-2004

At Duke, the ICPSR Learning to Curate project provided an opportunity to learn best practices in social science data curation from ICPSR while evaluating the feasibility of providing data curation as a service in Duke libraries. By participating in this project, we hoped to get a better sense of the amount of effort required to identify, process and publish data collections created by Duke researchers using ICPSR's curation standards. As a result of the curation project, we have created a section of our repository explicitly for datasets and now have a much better sense of the opportunities and challenges inherent in incorporating faculty created data collections into the library's institutional repository.

Data for this project came from a team of political scientists wishing to share their survey data on the characteristics of presidential donors during the 2000 and 2004 presidential elections (http://hdl.handle.net/10161/7882). The research team that conducted the study had always intended to share the data and had already documented the content of the associated data files. Additionally, we had direct access to one of the principal investigators, Alexandra Cooper, who frequently provided invaluable context for processing the study.

Despite the initial work by the research team to document their work, the curation team encountered a range of challenges in processing the files. First, the documentation occasionally omitted information necessary for secondary use and occasionally did not match the values that we found in the dataset. We expect that these alignment errors are probably present in most data projects produced by large research teams since it is difficult to document for secondary usage when your goal is to produce the primary research. Another challenge we encountered on the project was dealing with issues of confidentiality. The IRB requirements prohibited directly revealing the respondents of the survey, yet the dataset initially contained a large amount of demographic and geographic information on respondents. After some negotiation, the principal investigators agreed to provide some of the demographic information as long as it could be aggregated at a level that would make disclosure much less of a concern.

Overall, the Learning to Curate project benefited our data curation workflow in three ways. First, the experience allowed us to see the curation process at ICPSR firsthand, providing a much better appreciation of data curation at ICPSR and the efforts of their staff to produce quality datasets. Second, it provided us with a much better sense of the resource implications of providing quality data curation services to researchers on campus. At the end of the project, we realized that providing human mediated data curation proved extremely resource intensive. However, the project also raised many questions about the quality of data collections that do not receive a high level of processing/screening as they are archived. Finally, the project reaffirmed that the libraries could play a valuable role consulting with faculty on the best ways to manage and preserve data as a scholarly object.

Learning to Curate @ Emory – Home Schooling Policies in the US 1972-2007

Emory's participation in the ICPSR Data Curation Working Group coincided with the library having completed the ARL/DLF E-Science Institute in 2012 and subsequently having hired a new specialist for research

data management. We anticipated that we would learn quite a bit from the experience of working with ICPSR's processing pipeline and tools for data archiving. Long recognized for being the gold standard among social science data archives, ICPSR has set an example of best practices in data curation across many disciplines, with references to their data management guidelines appearing in multiple federal agency data sharing plans. As we were considering the development of appropriate services at Emory to assist researchers in preparing and depositing data for long-term access and preservation, we also felt it could prove useful to determine the implications of providing a premium level of service for both staffing and resource allocation.

The dataset we used for the ICPSR project consists of coded data on home-schooling policies in the United States from 1972-2007. Data was taken from a variety of publicly accessible and/or freely available sources (National Conference of State Legislatures, Census Bureau, National Center for Education Statistics), so we had no issues with proprietary data or sensitive, human-subject research to consider while preparing the data for archiving and sharing. However, the data had been assembled for a particular project and was not documented with long-term archiving and research reuse in mind.

Concurrent with the virtual group meetings, we worked through the dataset and identified issues requiring further clarification and adequate documentation for potential reuse. There was further back-and-forth with both the principal investigator (PI) and the graduate student researcher who had done much of the actual documentation and assembling of the data. It was to our advantage that we already had a long-standing relationship with the PI and that one of us had actually been involved in helping her locate relevant data sources for the project. Even with that prior knowledge, there was still a lot of clarification involved to make the data fit for long-term archiving.

This clarification did create a challenge in terms of using the ICPSR's Secure Data Enclave (SDE). Much of our processing work took place outside the SDE, since we were getting follow-up datasets and further documentation from the grad student who was our primary contact on the project (the PI had left Emory for another school). The amount of work involved just for this dataset raised valid questions of how to handle potential issues with datasets from other investigators (for example, sensitive information or lack of prior knowledge). What is required to provide such levels of service to researchers, and what are the resulting implications in terms of personnel, budgets and similar factors?

Our participation in the project was worth the time and effort involved. It was both impressive and informative to peer behind the curtain of the ICPSR data archiving process and see how the sausage is made, so to speak. And it aided us in answering some of those questions about how to provide assistance to researchers in preparing data for deposit. With existing staffing levels and resources, we could not realistically allot this degree of our time to curating individual datasets. But with the knowledge we have acquired about data archiving best practices, we feel comfortable providing consultations and guidance to PIs, especially if they have resources to bring to the table, such as graduate research assistants and funding support. And last but not least, we have the satisfaction of assisting this particular dataset in finding a good home for the benefit of both the original investigators and future researchers.

Learning to Curate @ UCLA – The Los Angeles County Social Survey

At the UCLA Social Sciences Data Archive (SSDA), we identified three goals we hoped to achieve through our participation in the project and use of the ICPSR curation tools:

1) develop a better understanding of the use of tools in data curation processes;
2) compare our local workflow with the ICPSR pipeline process; and
3) provide enhanced processing for legacy files.

Established in the mid-1960s, the SSDA facility at UCLA is a small domain-specific archive of data used in quantitative research, and the collection consists primarily of surveys, enumerations, public opinion polls and administrative records, many of which have been deposited by UCLA researchers.

SSDA carries out a data quality review of all files, creates detailed DDI-compliant metadata using tools from Colectica, carries out format migration as needed and processes data for online analysis with Survey Documentation and Analysis (SDA). SSDA is a member of the Data Preservation Alliance for Social Sciences (DataPASS) and the SSDA holdings are shared with DataPASS partners. For the ICPSR project, the Los Angeles County Social Survey (LACSS) was chosen because it is a uniquely held study, is one of the SSDA legacy files and would benefit from enhanced processing.

The learning curve created some initial barriers to effectively using the ICPSR curation tools. However it was immediately obvious that the ICPSR tools helped to improve the overall quality of the study for long-term preservation. By studying the processing pipeline at ICPSR, we were able to refine and streamline local workflows at the SSDA. The data quality review aspects showed how essential these steps are if the goal is to carry out the OAIS (Open Archival Information System) defined process for providing that data will be independently understandable for informed reuse. In contrast to the hands-on aspects of gold standard curation, we still have some qualms about the use of self-deposit tools where there is no data quality review, because it has already been shown that relying on researchers to do this deposit will result in significant data loss.

Some final thoughts in conclusion: there are new needs for institutions to work collaboratively with units external to libraries and archives, develop institution-wide policies for what to preserve, determine how to sustainably fund this work and develop and hire a trained workforce. We reiterate the need to recognize the kind and scope of commitment the preservation of usable research data will require. Institutions will need to evolve in terms of having a workforce with new responsibilities, in building and using data management tools beyond the institutional repository and in reallocating funding to support these endeavors.

Acknowledgements

We would like to thank our fellow participants in the ICPSR data curation working group: Ron Nakao, Rob O'Reilly and Jon Stiles. ■