RDAP14: Learning to Curate Panel

Jennifer Doty, Emory University
Jared Lyle, University of Michigan
Joel Herndon, Duke University
Libbie Stephenson, University of California, Los Angeles

Presentation Place: San Diego, CA
Publication Date: 2014-03-27
Type of Work: Presentation
Permanent URL: https://pid.emory.edu/ark:/25593/rhsbj

Final published version:
http://www.slideshare.net/asist_org/rdap14-learning-to-curate-panel-32822019

Copyright information:
This is an Open Access work distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

Accessed August 4, 2017 12:19 PM EDT
Learning to Curate:
Lessons from an ICPSR Pilot

Jared Lyle
RDAP 2014
http://www.icpsr.umich.edu
Background
## Data Sharing (N=935)

<table>
<thead>
<tr>
<th>Federal Agency</th>
<th>Shared Formally, Archived (n=111)</th>
<th>Shared Informally, Not Archived (n=415)</th>
<th>Not Shared (n=409)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSF (27.3%)</td>
<td>22.4%</td>
<td>43.7%</td>
<td>33.9%</td>
</tr>
<tr>
<td>NIH (72.7%)</td>
<td>7.4%</td>
<td>45.0%</td>
<td>47.6%</td>
</tr>
<tr>
<td>Total</td>
<td>11.5%</td>
<td>44.6%</td>
<td>43.9%</td>
</tr>
</tbody>
</table>

Pienta, Alter, & Lyle (2010). “The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data”. [http://hdl.handle.net/2027.42/78307](http://hdl.handle.net/2027.42/78307)
Vines et al. Current Biology 24, 94–97, January 6, 2014
http://dx.doi.org/10.1016/j.cub.2013.11.014
What is Curation?
A well-prepared data collection “contains information intended to be complete and self-explanatory” for future users.
A corollary: Do no harm.

http://img.gawkerassets.com/img/17xbuy519gga2jpg/ku-xlarge.jpg
Collaborative Curation
Partnerships

“We propose that domain specific archives partner with institution based repositories to provide expertise, tools, guidelines, and best practices to the research communities they serve.”

Support:

INSTITUTE of Museum and Library SERVICES
Ron Nakao, Stanford

Jen Doty, Emory

Joel Herndon, Duke

Rob O’Reilly, Emory

Libbie Stephenson, UCLA

Jon Stiles, UC Berkeley
Pilot Goals
For participants:

- Apply curation theories to practice through actual data processing.
- Will have a fully curated data collection ready for archiving at the end of the session.
- Interact with and ask questions of other data specialists within a working environment.
- Gain first-hand experience using ICPSR’s internal tools and workflows for curation.
- Understand level of effort to work through collections and provide assistance to researchers.
- Learn about things not thought about (e.g., costing, standardized workflows).
For ICPSR:

• Engage with outside data curators to learn what others are doing and thinking.
• Polish internal procedures and tools by opening them to outside review and critique.
• More data will be curated and archived, benefiting the ICPSR membership and the entire social science community.
• Better utilize resources of the OR community, including personal relationships and, especially, their wide-ranging expertise.
• Train a data curation community of support
Schedule

Week 1 - Introductions & Data Sources

Week 2 – Acquisition

Week 3 - Review

Week 4 – Processing

Week 5 – Metadata

Week 6 – Dissemination
The Virtual Data Enclave (VDE) provides remote access to quantitative data in a secure environment.
Lessons Learned
Your ideas on collaborative curation?
Thank you!

lyle@umich.edu
LEARNING TO CURATE
@ EMORY

Jen Doty and Rob O'Reilly
Reasons to Participate

- well-timed with new RDM hires
- higher-up support for involvement in RDM projects
What's in it for us?

- learn from gold standard holders:
  - ICPSR processing pipeline and tools
  - implications of providing premium level service for staffing and resource allocation

Nobel Prize Illustration by Howdy, I’m H. Michael Karshis on Flickr / CC BY 2.0
The Data

- Panel Data - all states in the United States, 1972-2007, annual
- Coded Data - state-level data policies on home schooling, and relevant court cases
- Publicly-Available Data - a mix of demographic, economic, and social data from sources such as the BEA, the Census Bureau, the NCES
- No issues with regard to sensitivity of data or proprietary restrictions
## The Data

![Data Editor Screenshot](image)

The data table contains various state-level characteristics including:

- **state**: State Abbreviation (e.g., AL for Alabama)
- **stabb**: State Postal Abbreviation (e.g., AL for Alabama)
- **tips**: State FIPS Code
- **year**: YEAR
- **repun**: Percent Republican
- **democr**: Percent Democrats
- **unemp**: Annual unemployment rate
- **gdpcc**: GDP per capita, adjusted for inflation
- **nsstate**: Homeschool statute
- **citide**: Citizen Ideology
- **govid**: Government Ideology
- **nocase**: Various other categorical variables

Each row represents data for a different state and year combination, with columns providing detailed metrics such as unemployment rates, GDP per capita, and political affiliations.

The table is used for data analysis and visualization, allowing researchers and analysts to explore trends and patterns across different states over time.
Issues and Considerations

- Data assembled for particular project, not with long-term archiving and research in mind
- Discrepancies in documentation:
  - variable names
  - unclear citations
  - broken URLs
  - variables in data missing from codebook, and vice-versa
Issues and Considerations, Cont.

- Long history with the Principal Investigator for the project, which meant lots of context about the project and the data.
- Useful in clarifying ambiguities in the data, e.g. “it makes sense to us” citations.
- Even with that context, there was still much work and back-and-forth involved.
Absent that prior history, the climb would have been much more steep
Conclusions and Implications

- Overall: very impressive to “see how the sausage is made”
  - ICPSR processing pipeline
  - Hermes
  - SDE infrastructure
Realistically, providing premium level of data archiving service is not possible with existing staffing levels and resources.
Work in Progress

- Intent to archive dataset with ICPSR still holds, but delayed by:
  - necessity for further documentation from investigators
  - demands on our time from other projects
- Future plans for archiving datasets created by campus researchers informed by lessons learned from participating in pilot project
Contact

- Jen Doty – jennifer.doty@emory.edu
- Rob O’Reilly – roreill@emory.edu
Learning to Curate @Duke

Joel Herndon
Data and GIS Services
Duke Libraries
• Duke’s Institutional Repository
• Largely a home for scholarly publications and dissertations
• A few data collections attached to papers, but limited research data
Presidential Donor Survey
2000-2004

• Alexandra Cooper (Duke)
• Michael Munger (Duke)
• John Aldrich (Duke)
• Clyde Wilcox (Georgetown)
• John Green (University of Akron)
• Mark Rozell (George Mason)
Presidential Donor Survey
2000-2004

- FEC data on political donations
- Stratified by candidate
- Survey topics include:
  - political activities
  - political attitudes
  - political attributes
Initial Impressions

- Codebook included
- PI(s) available
- IRB protocol available

2000-2004 Presidential Donor Survey
*************Panel*************
Codebook and Study Information
*******Beginning Draft-Updated October 2000*******

Table of Contents
1. General Introduction
2. Study Description
3. Study Design
4. Notes on the Weights
5. File Structures
6. Codebook of Variables
7. A note on String Variables
8. A note on Removed Cases

2. Study Description

Following the 2000 United States presidential election, a team of conducted a mail survey of donors who had contributed to any of eight presidential candidates. These included two Democrats (Bill Bradley and Al Gore), five Republicans (Gary Bauer, George W. Bush, Bob Dole, John McCain, and Ralph Nader). Each candidate's sample included approximately 100 donors sampled from public lists of donors maintained by the Federal Election Commission. Envelopes for Dole, Nader, and Bush were cross-sampled for substantive reasons. No data was collected on modest (<$200) donors for the Perkes campaign, as his campaign did not participate in public funding. Note that these samples were ALL drawn from individuals who gave before July 31, 2000.

The ten-page survey instrument contained items designed to collect data on donor demographics, on political attributes, attitudes, and activities, on how the donors were solicited, and on how many times, how often, and to whom they gave. It was sent directly to sampled donors via first-class mail. Each envelope was addressed to a specific donor and included a personalized cover letter explaining the project. All letters were individually signed by hand. Business-reply return envelopes were provided with these materials. Approximately one month after the initial mailing, a follow-up request was sent to all sample members who had not returned a response. The text of the cover letter was changed to reflect a repeat request; otherwise, the contents were identical to the first mailing. One additional mailing was done for all remaining non-respondents, except those in Nader's strata. Participants were given the option of responding anonymously to the survey by erasing the unique identifying number written in pencil on their surveys. Eight percent of the total sample chose to respond anonymously, with the highest rates of anonymous respondents from the Bauer (11%) and Perkes (16%) strata, and the lowest rates of anonymous responses from the Gore (6%), McCain (4%), and Bush (4%) strata. The three waves were sent over the period from August to November of 2001.

Following the 2004 United States presidential election the team sent...
(Initial) Challenges

- Codebook alignment
- Documentation Issues
- Confidentiality
- Missing Data
Explorations
Data Sets

Browse by

- Authors
- Titles
- Subjects
- Duke-affiliated Authors
- Duke Departments
- Issue Date
- Submit Date

These data sets were created or used by members of the Duke University community and curated by the Duke University Libraries. This program is intended to facilitate dissemination of research data. Preference will be given to data that may be distributed under the CC-BY license.

If you would like to submit data, please contact askdata@duke.edu.

Data is made available under the Creative Commons Attribution 3.0 Unported License unless otherwise specified.

Collections

- Data Sets
  Data sets independently archived.

http://dukespace.lib.duke.edu/dspace/handle/10161/8356
2000-2004 Presidential Donor Survey

Aldrich, John ; Munger, Michael C. ; Cooper, Alexandra ; Rozell, Mark ; Wilcox, Clyde ; Green, John
2013-10-01

Abstract:
Following the 2000 United States presidential election a team of faculty conducted a mail survey of donors who had contributed to any of eight presidential candidates. These included two Democrats (Bill Bradley and Al Gore), five Republicans (Gary Bauer, George Bush, Bob Dole, Steve Forbes, and John McCain), and Green candidate Ralph Nader. The sample was stratified by candidate; each candidate’s sample included approximately 300 donors sampled from public lists of donors maintained by the Federal Election Commission. Strata for Dole, Nader, and Bush were over-sampled for substantive reasons. No data was collected on modest (<$200) donors for the Forbes campaign as his campaign did not participate in public funding. Note that these samples were ALL drawn from individuals who gave before July 31, 2000. The ten-page survey instrument contained items designed to collect data on donor demographics, on political attributes, attitudes, and activities, on how the donors were solicited, and on how much, how often, and to whom they gave. It was sent directly to sampled donors via first class mail. Each envelope was addressed to a specific donor and included a personalized cover letter explaining the project. All letters were
Statistics

Total Visits

<table>
<thead>
<tr>
<th></th>
<th>Views</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000-2004 Presidential Donor Survey</td>
<td>229</td>
</tr>
</tbody>
</table>

Total Visits Per Month

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2000-2004 Presidential Donor Survey</td>
<td>0</td>
<td>0</td>
<td>109</td>
<td>41</td>
<td>32</td>
<td>24</td>
<td>23</td>
</tr>
</tbody>
</table>

File Visits

<table>
<thead>
<tr>
<th>File Name</th>
<th>Views</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presidential Donor Survey Codebook.doc</td>
<td>67</td>
</tr>
<tr>
<td>donors2000to2004.dta</td>
<td>45</td>
</tr>
<tr>
<td>Presidential Donor Survey Codebook.doc.txt</td>
<td>7</td>
</tr>
<tr>
<td>license.txt</td>
<td>1</td>
</tr>
</tbody>
</table>

Top country views

<table>
<thead>
<tr>
<th>Country</th>
<th>Views</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States of America</td>
<td>174</td>
</tr>
<tr>
<td>China</td>
<td>12</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>5</td>
</tr>
<tr>
<td>Spain</td>
<td>4</td>
</tr>
<tr>
<td>Germany</td>
<td>3</td>
</tr>
</tbody>
</table>

dukespace.lib.duke.edu › ... › Data Sets › Data Sets › Duke University

by J Aldrich - 2013

Oct 1, 2013 - Following the 2000 United States presidential election a team of faculty conducted a mail survey of donors who had contributed to any of eight ...

[doc] Codebook and Study Information - DukeSpace
dukespace.lib.duke.edu/.../Presidential%20Donor%20Su...

by J Aldrich - 2013

Study Description. 3. Study Design. Following the 2000 United States presidential election a team of faculty conducted a mail survey of donors who had ...
Concerns

• Resource Implications
• Defining library policies for curation
• Timely engagement with projects
Conclusions

- Greater appreciation of ICPSR’s curation role
- Resource implications for “curation as a service”
- Helps clarify our role for consulting for the full data life cycle
Contact

Joel Herndon, askdata@duke.edu
Learning to Curate: Lessons from an ICPSR Pilot RDAP Conference, March 26, 2014

Libbie Stephenson
UCLA Social Science Data Archive
libbie@ucla.edu
Social Science Data Archive

- Established mid-1960’s
- Small domain-specific archive of data for use in quantitative research
  - Surveys, enumerations, public opinion polls, administrative records
- Two full time staff; part time student interns
- Holdings are partly files deposited by faculty
Goals in project

• Learn new skills in curation process
• Compare local workflow with ICPSR Pipeline process
• Focus on legacy files; enhanced processing
• Improve condition of data deposited to ICPSR
• Consider how researchers would benefit
• Advise other local professionals
Current curation practices

• Follow OAIS to appraise and ingest files
• Data Quality Review: Compare codebook with data; compare to system files and/or create; run freqs; minimal disclosure checks
• Metadata from data deposit form
• Data and metadata processed in Colectica
• Carry out media format migration when necessary
• Process for use with SDA
• DataPASS deposit
Operational schematic
Los Angeles County Social Survey

- Annual survey of about 1000 respondents; oversample of Blacks and Asian-Americans
- Topics: attitudes and views of living in Los Angeles, neighborhoods, public services, and political views
- Used computer-assisted telephone interview (CATI); Spanish and English with the CASES tools.
- Geography by zip code within county
Thoughts on the project

• Learning curve; since not using the tools daily difficult to remember steps
• ICPSR tools help to quickly improve data quality
  – Disclosure, naming conventions, missing data, etc.
• Following parts of the ICPSR pipeline process would streamline local work flow
• Project demonstrated that data quality review aspects are essential to preservation.
Where we are now

• Explore cooperative arrangement with Library
  – Archive to curate; use of IR for bit-level maintenance

• Re-evaluate acquisitions/collection policies
  – Find better ways to estimate resource needs
  – Increase advisory role; make use of ICPSR and

• Redesigned workflow for appraisal and ingest
  – More focus on data quality review
  – Increase use of tools to create metadata
  – Write training manual – shorten learning curve

For more on Data Quality Review: [http://www.dcc.ac.uk/sites/default/files/documents/IDCC14/Parallels/Committing%20to%20data%20quality%20review%20parallel%20C1.pdf](http://www.dcc.ac.uk/sites/default/files/documents/IDCC14/Parallels/Committing%20to%20data%20quality%20review%20parallel%20C1.pdf)
Committing to Curation - Conclusions

- Goal is to ensure long term usability of scholarly output
  - It is ALL digital
  - Cannot preserve, store, identify, curate it ALL
  - Have to set priorities, establish policies, develop criteria for what to preserve for long term usability.
  - Bit-level processes are NOT enough
  - DIY tools for self-deposit vs DQR for preservation

- Curation is a NEW aspect of librarianship

- It IS a commitment
  - Develop expertise
  - Acquire, license or build tools
  - Staff – numbers and expertise required
  - Financial impact is not trivial