ICPSR, Data Producers, and Institutional Repositories: Successes and Challenges

Jared Lyle
IASSIST 2013
About ICPSR

• Founded in 1962 as a consortium of 21 universities to share the National Election Survey

• Today: 700+ members around the world

• Data dissemination for more than 20 federal and non-government sponsors

• 600,000+ visitors per year
What we do

• Acquire and archive social science data
• Distribute data to researchers
• Preserve data for future generations
• Provide training in quantitative methods

Archive size

• 8,000 data collections, over 60,000 data sets
• Grows by 300+ collections a year
• Soon to be 40+ Terabytes
Background
“It saves funding and avoids repeated data collecting efforts, allows the verification and replication of research findings, facilitates scientific openness, deters scientific misconduct, and supports communication and progress.”

ABOUT DATA-PASS

The Data Preservation Alliance for the Social Sciences (Data-PASS) is a voluntary partnership of organizations created to archive, catalog, and preserve data used for social science research. Examples of social science data include: opinion polls, voting records, surveys on family growth and income, social network data, government statistics and indices, and GIS data measuring human activity.

As a National Digital Stewardship Alliance Founding Member, the Data-PASS partnership works to:

- Archive social science data collections at risk of being lost.
- Catalog and promote access to archived collections in the Data-PASS shared catalog.
- Develop preservation of archived collections.
- Advocate best practices in digital preservation.

Current partnership activities include:

- Expanding the partnership to involve others who are interested in preserving social science research.
- Promoting creation standards for research data.
- Developing a self-contained life-cycle system to manage and audit replication of digital holdings.

Read more about Data-PASS.

SHARED CATALOG

The Data-PASS shared catalog enables anyone to search and browse the entire holdings of most Data-PASS partners, and to download and analyze many publicly accessible studies.

SEARCH My Datasource  Search

http://www.data-pass.org/
Most PIs indicated that they wanted to be “Good Citizens” and help:

“This sounds like an exciting project.”

“I hope your project is successful because I think that it is important.”
“Good Citizens” = high willingness

...but no time, money, or resources to submit data to us.
http://ori.hhs.gov/content/research-research-integrity-rri-conference-2009

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
• Data management planning requirements
• Federal mandates
• OSTP memorandum
A well-prepared data collection “contains information intended to be complete and self-explanatory” for future users.
<table>
<thead>
<tr>
<th>Action</th>
<th>ICPSR Id</th>
<th>Title</th>
<th>Principal Investigator</th>
<th>Date Of Future Contact</th>
<th>Last Modified Date</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1U01DA010375-01</td>
<td>DRUG ABUSE TREATMENT OUTCOMES--A TX CAREERS PERSPE...</td>
<td>M. Anglin</td>
<td>3/1/2011</td>
<td>9/24/2010</td>
</tr>
<tr>
<td></td>
<td>1U01DA010375-01</td>
<td>Modeling HIV Diffusion through Drug Using Networks</td>
<td>William Zule</td>
<td>2/28/2009</td>
<td>9/10/2010</td>
</tr>
<tr>
<td></td>
<td>1U01HL06865-01</td>
<td>TRIAL OF ACTIVITY FOR ADOLESCENT GIRLS (TAAG)</td>
<td>John Elder</td>
<td>2/21/2011</td>
<td>10/30/2010</td>
</tr>
<tr>
<td></td>
<td>1U01NH066177-01A1</td>
<td>Depression in Alzheimer's Disease Study 2 (DIAD-2)</td>
<td>Anton Porstensson</td>
<td>9/30/2010</td>
<td>10/10/2010</td>
</tr>
<tr>
<td></td>
<td>1U22AA011089-010001</td>
<td>SOCIOEMOTIONAL DEVELOPMENT OF CHILDREN WITH PRENATAL...</td>
<td>Daphne Calmes</td>
<td>3/21/2011</td>
<td>12/20/2010</td>
</tr>
<tr>
<td></td>
<td>1Z01AG000839-06</td>
<td>INDIVIDUAL CHANGES IN FUNCTIONING WITH AGE AND TARGET C...</td>
<td>L. Verbrugge</td>
<td>6/8/2011</td>
<td>7/6/2011</td>
</tr>
<tr>
<td></td>
<td>1Z01AG000839-06</td>
<td>INDIVIDUAL CHANGES IN FUNCTIONING WITH AGE AND TARGET C...</td>
<td>L. Verbrugge</td>
<td>6/8/2011</td>
<td>7/6/2011</td>
</tr>
<tr>
<td></td>
<td>6323118</td>
<td>Social and Cultural Aspects of Modernization</td>
<td>Alex Ingle</td>
<td>8/22/2011</td>
<td>10/26/2010</td>
</tr>
<tr>
<td></td>
<td>6406289</td>
<td>Kinship and Social Structure</td>
<td>Eugene Hammel</td>
<td>3/21/2011</td>
<td>12/2/2010</td>
</tr>
<tr>
<td></td>
<td>6608268</td>
<td>Social Networks and Mobility in Urbanization</td>
<td>Eugene Hammel</td>
<td>3/21/2011</td>
<td>12/2/2010</td>
</tr>
<tr>
<td></td>
<td>6751195</td>
<td>Members' Response to Organizational Hierarchy</td>
<td>Arnold S. Tannenbaum</td>
<td>6/6/2011</td>
<td>10/26/2010</td>
</tr>
<tr>
<td></td>
<td>6851195</td>
<td>Member's Responses to Organizational Hierarchy</td>
<td>Arnold S. Tannenbaum</td>
<td>6/6/2011</td>
<td>10/26/2010</td>
</tr>
<tr>
<td></td>
<td>7132354</td>
<td>SCIENCE DEVELOPMENT IN SOCIOLOGY</td>
<td>Glen Terrell</td>
<td>11/24/2010</td>
<td>1/24/2010</td>
</tr>
<tr>
<td></td>
<td>7205326</td>
<td>The Sociology of Science</td>
<td>Robert K. Merton</td>
<td>10/12/2010</td>
<td>10/12/2010</td>
</tr>
</tbody>
</table>
Emerging sources and types of data

- Geo-spatial
- Video
- Administrative data
- Online text
- Transactions
- Clicks
- Sensors
Distributed 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
This project is about building partnerships between institutional repositories and social science data archives. Improvements in data processing and storage technology are resulting in an increase in research data on a variety of social, economic, and political subjects. Many datasets could be profitably reanalyzed, but they are at danger of being lost since they are never properly archived. Institutional repositories are charged by their institutions to preserve the scholarly products of their faculty or institution, but not all feel prepared to curate and archive social science data. How can data archives and institutional repositories improve coordination to successfully work together to archive data?

http://www.icpsr.umich.edu/icpsrweb/IR/
5 Pilot Data Collections

http://www.flickr.com/photos/smithsonian/2551170386/
Selection & Appraisal
Recovery

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Width</th>
<th>Decimals</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>V16</td>
<td>Numeric</td>
<td>1</td>
<td>0</td>
<td>WIDOW/WIDOWER</td>
</tr>
<tr>
<td>V17</td>
<td>Numeric</td>
<td>1</td>
<td>0</td>
<td>SOCIAL CLASS</td>
</tr>
<tr>
<td>V18</td>
<td>Numeric</td>
<td>1</td>
<td>0</td>
<td>RACE</td>
</tr>
<tr>
<td>V19</td>
<td>Numeric</td>
<td>1</td>
<td>0</td>
<td>ELDERLY CHAR HEALTHY?</td>
</tr>
</tbody>
</table>
Finding interested partners

http://www.flickr.com/photos/usnationalarchives/4726917373/
Survey of Repositories’ Data Needs

Repository Suggested Solutions:

- Media recovery, format migration, data recovery
- Cost estimating and policy review
- Metadata tools, documentation, and catalog linkages
- Support networks and training
- Confidential data dissemination and confidentiality review
1. Community Wayfinder
2. Confidentiality Review & Treatment
• Suppressing unique cases
• Grouping values (e.g., 13-29=1, 30-49=2)
• Top-coding (e.g., >1,000=1,000)
• Aggregating geographic areas
• Swapping values
• Sampling within a larger data collection
• Adding “noise”
• Replacing real data with synthetic data
QualAnon

DSDR Qualitative Data Anonymizer

View QualAnon Instructions (PDF)

Transcript file to upload:  Choose File  No file chosen
Name Key file to upload:  Choose File  No file chosen

- Annotate: anonymized output with original names
- Allow duplicates: in name key -- use with caution!

Upload Files  Clear Form

http://www.icpsr.umich.edu/icpsrweb/content/DSDR/tools/qualanon.html
3. Access to Processing Tools
The Virtual Data Enclave (VDE) provides remote access to quantitative data in a secure environment.
Hermes Outputs

• ASCII data files
  – Column- and tab-delimited

• Stat package setup files
  – SAS, SPSS, Stata (.do and .dct)

• “Ready-to-go” data files
  – SAS transport (CPORT engine)
  – SPSS system (.sav)
  – Stata system (.dta)
  – R (.rda)
Your ideas on partnerships?

Useful categories for discussion?

• Media recovery, format migration, data recovery
• Cost estimating and policy review
• Metadata tools, documentation, and catalog linkages
• Support networks and training
• Confidential data dissemination and confidentiality review
Thank you!

lyle@umich.edu