Discovery Acceleration Workshop

Secondary Data Series
Enhancing Your Research with Secondary Data

Daria Orlowska
Data Librarian and Assistant Professor
Waldo Library
Workshop Goals

1. List ways that secondary data can be incorporate into research and the classroom

2. Learn about funding opportunities for secondary data

3. Gain a basic understanding of ICPSR and All of Us
1. Why Secondary Data?
Let’s Talk About Secondary Data

Secondary data is collected by someone else than the user, or used for a different purpose than originally intended.
Benefits to Using Secondary Data

• Less resource intensive
• Potentially more representative
• Advances science through replicability
• Untapped potential
• Excellent teaching tool
When Not to Use Secondary Data

- Project requiring specific data
  - Control over methodology
  - Accuracy and reliability
2. Picking the *Right* Secondary Data
Secondary Data Strategies

Two Approaches

• Formulating questions, then finding relevant dataset(s)
  
  **PROS:** Creativity
  
  **CONS:** Compromise

• Locating dataset, then formulating questions
  
  **PROS:** Faster process
  
  **CONS:** Less research freedom
Secondary Data Strategies
Definitions

• Data vs Statistics (Microdata vs Aggregate)
• Survey vs Census vs Administrative
• Cross-sectional vs Longitudinal vs Time-series
Figuring Out Your Data Needs

Develop a Research Question

How many minority students are currently enrolled in public vs private universities?

How does this compare between places that are conservative vs liberal?
Figuring Out Your Data Needs

Data Timeline

- Has this research been conducted before?
- Single source or multiple?
- Readily available or requires aggregation?
- Downloadable or requestable?
- De-identified or restricted?
- Free or behind paywall?
- Open format or proprietary?
- Requires cleaning/transformation?

Figuring Out Your Data Needs
Assess Data Credibility and Usability

• Authorship
• Connection to publication or institute
• Documentation
• Methodological issues
Figuring Out Your Data Needs
Inspect the Methodology

• What was the data’s original purpose?

• How was the population sampled?
  • Who was excluded?

• What is the sample size?

• Is the data longitudinal?

• What is the smallest data unit?

• How was missing data treated?
## Figuring Out Your Data Needs

### Determine Reusability Limitations

<table>
<thead>
<tr>
<th>License</th>
<th>Copy and redistribute?</th>
<th>Attribution required?</th>
<th>Commercial Reuse?</th>
<th>Adaptation allowed?</th>
<th>License change allowed?</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC0</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>PDDL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PDM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC BY-ODC-By</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>CC BY-SA</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>CC BY-ND</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
</tr>
<tr>
<td>CC BY-NC</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>CC BY-NC-SA</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>CC BY-NC-ND</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>Y</td>
</tr>
</tbody>
</table>

https://creativecommons.org/licenses/
Do I Need IRB Approval?

I am using anonymized / aggregated data
AND
I did not participate in data collection.
I cannot reidentify individuals.

Maybe. Under the OHRP Guidance, IRB oversight is not required. However, some organizations may want to see IRB review. Honors thesis and dissertations also require review.

I am using a dataset that requires a subscription, password, or permission for access.

Expedited. Datasets that have use restrictions may contain identifiers and are reviewed under the Exempt Category by the IRB.

I am using human subjects data that has not been deidentified.

Proposed project summaries can be emailed to the IRB for review ovpr-hsirb@wmich.edu
3. Secondary Data Funding
Funding Sources

National Institutes of Health (NIH)

- Data Harmonization, Curation and Secondary Analysis of Existing Clinical Datasets
- Small Grants for Secondary Analyses of Existing Data Sets and Stored Biospecimens
- NEI Research Grant for Vision-Related Secondary Data Analysis
- Secondary Analysis of Existing Datasets in Heart, Lung, and Blood Diseases and Sleep Disorders
- Secondary Analyses of Existing Datasets of Tobacco Use and Health
- Secondary Analyses of Existing Alcohol Research Data
- Secondary Analysis and Integration of Existing Data to Elucidate the Genetic Architecture of Cancer Risk and Related Outcomes
- Secondary Analysis and Integration of Existing Data Related to Acute and Chronic Pain Development or Management in Humans
Funding Sources (cont.)

National Science Foundation (NSF)
• Supporting Use of Existing Data and Samples in Atmospheric Sciences Research and Education

Department of Health and Human Services (HHS)
• Career Pathways Secondary Data Analysis Grants

Health Resources & Services Administration (HRSA)
• Autism Secondary Data Analysis Research Program

Russell Sage Foundation (RSF)
• Research Grants, particularly interested in novel uses of existing data

https://wmich.edu/research/funding
4. Secondary Data Resources
NIH’s All of US

Version Date: Apr 2, 2015

Principal Investigator(s):
Deborah Schildkraut, Tufts University; Ashley Grosse, Washington State University, Social and Economic Sciences Research Center

https://doi.org/10.3886/ICPSR27601.v3

Version V3

Project Description

Summary

The 21st Century Americanism survey was conducted to study (1) the multidimensional nature of American identity ("Americanism"); (2) resentment among Whites toward immigrants, Latinos, and Asians, fueled by perceptions that these groups violate the cherished norms that constitute American identity ("symbolic nativism"); (3) how perceptions of discrimination affect the process of "becoming American" among ethnic minorities ("reactive ethnicity"); and (4) the relationships among these issues and public opinion on policies that address ethnic change. The data collection began in July 2004 and was completed by October 2004. This nationally representative random-digit dial telephone survey has 2,800 respondents and includes oversamples of Blacks, Latinos, and Asians in the United States. It contains questions that allow for the examination of the causes and consequences of two facets of American identity: (1) how people define the normative content of American identity ("identity content"); and (2) the extent to which people think of themselves primarily as American rather than primarily as a member of a pan-ethnic (i.e., Latino or Asian) or national origin group ("identity attachment"). The survey can be used to test hypotheses regarding whether the alleged traditional consensus on what it means to be American is breaking down, or

Notes

• Curation and dissemination of this study is provided by the Institutional members of ICPSR, and data is available only to users at ICPSR member institutions. To determine if you are at a member institution, check the list of ICPSR member institutions, or learn more about becoming a member.

• One or more files in this data collection have special restrictions. Restricted data files are not available for direct download from the website; click on the Restricted Data button to learn more.
5. Reflections on Secondary Data
Thank you!

Finding Secondary Data Utilizing Library Resources
Monday, March 14 @ 12:00-1:00pm

Adapting Secondary Data for Analysis
Monday, March 21 @ 12:00-1:00pm

Daria Orlowska, Data Librarian and Assistant Professor
daria.orlowska@wmich.edu

Watch workshop recording here: