Introducing ICPSR

Workshop on Korean Data Services
11/15/2017

Lynette Hoelter, Director of Instructional Resources, ICPSR
Who is ICPSR?

• One of the world’s oldest and largest social science data archives
  • Founded in 1962 (Summer Program, 1963)
  • Current holdings: 10,000+ studies
  • Bibliography of 71,000+ data-related publications
• Membership organization of 22 institutions initially, now ~750 institutions world-wide
  • Entire collection, teaching materials, discounted SP fees available to those at member institutions
  • Large portion of collection freely available (externally funded)
The ICPSR Mission

ICPSR advances and expands social and behavioral research, acting as a global leader in data stewardship and providing rich data resources and responsive educational opportunities for present and future generations.
Notes about Membership Access

• To find out if your school is a member:
  http://www.icpsr.umich.edu/icpsrweb/membership/administration/institutions
What We Do – It’s About Data!

• Seek research data and pertinent documents from data producers (PIs, research agencies, government)
• Process (clean), describe (tag), and preserve the data and documents
• Share data
• Provide education, training, & instructional resources
• Assist grant writers (and authors) with data management plans and fulfillment of data sharing requirements
What Makes ICPSR Unique?

• Documentation to accompany data
• “Come here, get data, it’s been checked for errors and is well marked up so you understand what it is”
• No fishing expedition to find things you don’t want
• Explore the data and get to know it before you download
• Wide range of helpful resources
ICPSR in a Nutshell

- Lots of data
- Summer Program and instructional materials
- Data management and deposit
ICPSR Can Help with…

• Research
  • Data for secondary analysis
  • Help with data management plans, data sharing, showing impact

• Teaching and learning
  • Summer Program
  • Prepared exercises
  • Tools can be used in teaching methods/stats (and more)
Secondary Research

• Finding/using data for secondary analysis
  • Metadata allows for robust searching
  • See what has been done previously, demonstrate need
  • Tools to evaluate or analyze data
  • User support
• Deposit replication data/code
A Glimpse of the Catalog

- 8,230 studies (69,655 datasets) available “on demand”
- 1,479 studies (7,540 datasets) available as “restricted use”
- 1,254 studies have online analysis capabilities
- 266,206 files
- Social and behavioral sciences, broadly defined
Making Sense of That Much Data
Specialty Collections

- Arts/Culture (Nat’l Archive of Data on Arts and Culture)
- Criminal Justice (Nat’l Archive of Criminal Justice Data)
- Demography
  - Nat’l Archive of Computerized Data on Aging
  - Data Sharing for Demographic Research
  - Resource Center for Minority Data
  - Open Data Flint
- Education
  - Child Care and Early Education Research Connections
  - Civic Learning, Engagement and Action Data Sharing
  - Measures of Effective Teaching
  - NCAA Student-Athlete Experiences
- Health/Wellness
  - Archive of Data on Disability to Enable Policy
  - Health and Medical Care Archive
  - Nat’l Addiction and HIV Data Archive Program
- Self-Deposit (Open ICPSR) and At-Risk Government Data (Data Lumos)
Search and Evaluation Tools

• Built upon strong metadata at the study and variable levels
• New infrastructure → even better searching
• Finding and evaluating data:
  • Bibliography of Data-related Literature
  • Social Science Variables Database
  • Online Analysis for exploration
  • Study homepage
  • Study documentation
Bibliography of Data-Related Literature

- **Searchable database** of over 71,600 citations of known published and unpublished works linked to data archived at ICPSR
- Generates reference lists for each study
- Included in one integrated search on the ICPSR website
- Great place for students to start
Social Science Variables Database

Enables ICPSR users to:

• **Search & Compare Variables** across datasets
• Assists in:
  – Data discovery
  – Comparison/ harmonization projects
  – Data harvesting & data analysis
  – Question mining for designing new research
  – Research methods & substantive courses instruction

11/15/2017 Workshop on Korean Data Services
### Compare Variables

#### Question

**Q90_1**

Global View 2008: American Public Opinion and Foreign Policy; July 2008 Survey

If North Korea invaded South Korea:

**Question**

BASE: RANDOM 2/3 OF RESPONDENTS

Q90 (1-6): There has been some discussion about the circumstances that might justify using U.S. troops in other parts of the world. Please give your opinion about some situations. Would you favor or oppose the use of U.S. troops:

WA/1. If North Korea invaded South Korea

#### Responses

<table>
<thead>
<tr>
<th>VALUE</th>
<th>LABEL</th>
<th>UNWEIGHTED FREQUENCIES</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Favor</td>
<td>382</td>
<td>25.4%</td>
</tr>
<tr>
<td>2</td>
<td>Oppose</td>
<td>527</td>
<td>35.0%</td>
</tr>
<tr>
<td></td>
<td>Missing Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-2</td>
<td>Not Asked</td>
<td>563</td>
<td>37.4%</td>
</tr>
<tr>
<td>-1</td>
<td>Refused</td>
<td>33</td>
<td>2.2%</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1,505</td>
<td>100%</td>
</tr>
</tbody>
</table>

#### Time Period

2008

---

#### Question

**Q30_1**

Global View 2010: American Public Opinion and Foreign Policy

If North Korea invaded South Korea:

**Question**

**Responses**

<table>
<thead>
<tr>
<th>VALUE</th>
<th>LABEL</th>
<th>UNWEIGHTED FREQUENCIES</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>Refused</td>
<td>46</td>
<td>1.8</td>
</tr>
<tr>
<td>1</td>
<td>Favor</td>
<td>538</td>
<td>20.7</td>
</tr>
<tr>
<td>2</td>
<td>Oppose</td>
<td>711</td>
<td>27.4</td>
</tr>
<tr>
<td></td>
<td>Missing Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>-</td>
<td>1,302</td>
<td>50.1</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>2,597</td>
<td>100%</td>
</tr>
</tbody>
</table>

#### Time Period

2010-06-11-2010-06-22
Online Analysis Tool (SDA)
“List studies for which online analysis is available”
Your query returned 458 studies. Series results are grouped under the most relevant study. Ungroup series results to view all results.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
</table>
| **1.** | **Facebook Interaction Data and Survey of People: 2013 [Republic of Korea]** (ICPSR 4650)  
Wee, Jeun; Lee, Joohwan |
| **2.** | **Public Opinion and Foreign Policy in the United States, China, India, Australia, and South Korea, 2006** (ICPSR 4650)  
Chicago Council on Global Affairs; Asia Society; East Asia Institute (EAI) |
| **3.** | **World Health Organization Collaborative Study on Social and Health Aspects of Aging in Fiji, Korea, Malaysia, and the Philippines, 1983-1985** (ICPSR 6057)  
Andrews, Gary R. |
| **4.** | **World Health Organization Regional Office for Southeast Asia Collaborative Study on Social and Health Aspects of Aging in Burma, Indonesia, North Korea, Sri Lanka, and Thailand, 1990** (ICPSR 6242)  
Andrews, Gary R. |
| **5.** | **Soft Power in Asia Survey, 2008** (ICPSR 25342)  
Bouton, Marshall; Kul, Steven; Page, Benjamin; et al. |
| **6.** | **Global Views 2004: South Korean Public Opinion and Foreign Policy** (ICPSR 4135)  
Chicago Council on Foreign Relations; The East Asia Institute  
16 more results in American Public Opinion and United States Foreign Policy Series |
Rodgers, Yana |
| **8.** | **Are Prices Countercyclical? Evidence From East Asian Countries** (ICPSR 1168)  
**Workshop on Korean Data Services** |
Korean General Social Survey (KGSS), 2010 (ICPSR 34666)

Principal Investigator(s): Kim, Sang-Wook. Sungkyunkwan University (South Korea), Survey Research Center

Summary:

The Korean General Social Survey (KGSS) is the South Korean version of the General Social Survey (GSS), closely replicating the original GSS of the National Opinion Research Center at the University of Chicago. Each round of the KGSS typically includes the topical module surveys of the International Social Survey Programme (ISSP), and/or the East Asian Social Survey (EASS), an international survey network of four GSS-type surveys from countries in East Asia (including China, Japan, Taiwan, and South Korea). Respondents were ... more info

Series: Korean General Social Survey (KGSS) Series

Access Notes

- Data in this collection are available only to users at ICPSR member institutions. Please log in so we can determine if you are with a member institution and have access to these data files.

Dataset(s)

Korean General Social Survey (KGSS), 2010 - Download All Files (16.3 MB)

Documentation: Codebook.pdf
Download: SAS SPSS Stata R ASCII Excel/TSV
ASCII + SAS Setup SPSS Setup Stata Setup
Analyze Online: simple crosstab/frequency SDA

Study Description

Citation


Persistent URL: https://doi.org/10.3886/ICPSR34666.v2
## Search Tools/Strategies

<table>
<thead>
<tr>
<th>You Have/Know…</th>
<th>You Want</th>
<th>Start Here</th>
</tr>
</thead>
<tbody>
<tr>
<td>General topic area</td>
<td>Ideas of studies that might work, work others have done</td>
<td>Thematic collections, general search, bibliography</td>
</tr>
<tr>
<td>Info about the study (PI, title)</td>
<td>To find that study</td>
<td>General search</td>
</tr>
<tr>
<td>Variables/questions of interest</td>
<td>A study containing those questions, an example of question wording for creating a survey, frequency distributions</td>
<td>Search/Compare Variables (Social Science Variables Database)</td>
</tr>
<tr>
<td>The study of interest</td>
<td>Variables included, what has been done with it, information about methodology</td>
<td>Study homepage</td>
</tr>
<tr>
<td>The study of interest</td>
<td>Whether the subgroup in which you are interested is large enough</td>
<td>Codebook</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Online analysis (SDA)</td>
</tr>
<tr>
<td>That you want to use data in class</td>
<td>Something students can use online</td>
<td>View all studies for which online analysis is available</td>
</tr>
<tr>
<td>A journal publication</td>
<td>To replicate the results</td>
<td>Replication datasets, OpenICPSR</td>
</tr>
</tbody>
</table>
User Support

- FAQs
- YouTube Channel
  - Tutorials
  - Webinars
  - Presentations
- Phone/Email
- Local support (ORs)
When Finished

• Cite data in publication
• Deposit data/code for replication
• Send in citation
Primary Research

• Determining need
• Developing questionnaire
• Data management plans for proposals
• Budget for archiving
• Letter of support
• Informed consent so data can be shared
• Guide to Social Science Data Preparation and Archiving
• Sharing data
Useful Examples

Elements of a Data Management Plan

<table>
<thead>
<tr>
<th>Element</th>
<th>Description</th>
<th>Recommended?</th>
<th>NSF Mapping</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data description</td>
<td>A description of data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Existing data</td>
<td>A survey</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Format</td>
<td>Formats and archival</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metadata</td>
<td>A description of data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Storage and backup</td>
<td>Storage used for data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Security</td>
<td>A description of data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Responsibility</td>
<td>Names of data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intellectual property rights</td>
<td>Ownership of intellectual property</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access and sharing</td>
<td>A description of data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Audience</td>
<td>The potential of data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Selection and retention periods</td>
<td>A description of data</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Confidentiality

The success of the social science research enterprise relies on the willingness of research participants to take part in surveys. Thus, it is critically important to protect the identities of research subjects. Disclosure risk is a term used to describe the risk that private information that otherwise would not have been revealed will be revealed, perhaps in an unexpected way. It has become easier to link research data to other data sources, which may reveal, or at least make it easier to identify, research participants.

ICPSR's Approach to Confidentiality

ICPSR accepts data with identifying information. ICPSR staff can work closely with data providers to create confidentiality of individuals and organizations.

Disclosure Risk

With the exception of deposits placed into ICPSR’s restricted-use data files, ICPSR’s policy is to provide data in a non-confidential form.

Identifiers

Two kinds of identifiers are often found in data files:

- Direct Identifiers: These are variables that identify individuals or organizations, such as names, addresses, telephone numbers, and social security numbers.
- Indirect Identifiers: These are variables that may be used to identify an individual, such as dates of birth, dates of marriage, names of siblings, and so on.

Recommended Informed Consent Language for Data Sharing

Language to Avoid

Promises in the informed consent can appear to limit an investigator’s ability to share data with the research community. In reality, investigators can inform study participants that they are scientists with an obligation to protect confidentiality and still share the study data with the broad scientific community. Many effective means exist to create public-use data files or share restricted-use data files under controlled conditions. That is, data can be modified to reduce the risk of disclosure or shared with additional safeguards while preserving their value for science.

Model Language

Here are two model statements investigators may use in informed consents to describe protection of confidentiality that also allows data sharing.

Sample 1. Study staff will protect your personal information closely so no one will be able to connect your responses and any other information that identifies you. Federal or state laws may require us to show information to university or government officials (or sponsors), who are responsible for monitoring the safety of this study. Directly identifying information (e.g., names, addresses) will be safeguarded and maintained under controlled conditions. You will not be identified in any publication from this study.

Sample 2. The information in this study will be used only for research purposes and in ways that will not reveal who you are. Federal or state laws may require us to show information to university or government officials (or sponsors) who are responsible for monitoring the safety of this study. You will not be identified in any publication from this study.

Known Concerns and Recommended Alternatives

Concern 1

Terms such as “anonymous” and “de-identified” are undefined and left open to interpretation. Some data are collected anonymously as directly identifying information is never obtained. De-identification may involve more than removing direct identifiers. Indirect identifiers in the file may still be used in combination to isolate a subject that is unique on certain characteristics. Even using the “safe harbor” method of de-identification by removing 18 specified elements still requires the covered entity to affirm it has no “actual knowledge that the remaining information alone or in combination with other information can be used to possibly identify the subject” (source [2, 156KB]).

Recommendations:

- Use descriptive sentences that state what information will not be shared:
  - “Any personal information that could identify you will be removed or changed before files are shared with other researchers or results are made public.”
  - “Your answers to the questions I ask will be anonymous. That is, I will not ask for your name, and we will not attach your name or jail number to your answers.”
  - “The Personally Unidentified Study Data does not include your name, address, telephone or social security number.”

- Use descriptive sentences that state what may be retained in data if shared with other researchers:
  - “Personally Unidentified Study Data may include your date of birth, initials, and dates you received medical care. Personally Unidentified Study Data also may include the health information used, created, or collected in the research study.”
Data Sharing: Deposit Options

- Uncurated: OpenICPSR
  - Self-deposit, immediate publishing with persistent ID
  - What you deposit is what others get

- Curated: ICPSR
  - Thematic archive (Funders set acquisitions policies)
  - Queue for Members Archive
  - Fee for curation, data available to all
Questions? Comments?

Lynette Hoelter
lhoelter@umich.edu