Archiving and Providing Access to Confidential Social Science Data

Jared Lyle
George Alter

IDCC 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 200+ collections a year
Quantitative data

![Data Editor (Browse) - [auto.dta]](http://www.ssc.wisc.edu/sscc/pubs/stata_students1.htm)
Emerging sources and types of data

- Geo-spatial
- Video
- Administrative data
- Online text
- Transactions
- Clicks
- Sensors
Adaptive education software records every click.
SECTION 97.
…, a gaming establishment shall supply the Massachusetts gaming Commission with customer tracking data collected or generated by loyalty programs, player tracking software, player card systems, online gambling transactions or any other information system. ... The Commission shall convey the anonymized data to a research facility which shall make the data available to qualified researchers …


Casinos capture data too...
Safety Pilot Project

- 2800 cars, trucks, and buses with Vehicle Awareness Devices, sensors, and video
- ~1 Petabyte of data per year
MET Longitudinal Database

Scale
- 2 academic years
- 6 large school districts
- 6 grade levels
- 3,000 teachers
- 44,500 students
- 24,000 videos
- 22,500 observation sessions
- 900+ observers trained by ETS to score videos
- ~12GB of quantitative data
- ~10TB of video
Disclosure: Risk & Harm

- What do we promise when we conduct research about people?
  - That benefits (usually to society) outweigh risk of harm (usually to individual)
  - That we will protect confidentiality

- Why is confidentiality so important?
  - Because people may reveal information to us that could cause them harm if revealed.
  - Examples: criminal activity, antisocial activity, medical conditions...
What are We Afraid of...

- **Direct Identifiers**
  - Inadvertent release of unnecessary information (Name, phone number, SSN...)
  - Direct identifiers required for analysis (location, genetic characteristics,...)

- **Indirect Identifiers**
  - Characteristics that identify a subject when combined (sex, race, age, education, occupation)
Current Survey Designs Increase the Risks of Disclosing Subjects’ Identities

- Geographically referenced data
- Longitudinal data
- Multi-level data:
  - Student, teacher, school, school district
  - Patient, clinic, community
Who are We Afraid of?

- Parents trying to find out if their child had an abortion or uses drugs
- Spouse seeking hidden income or infidelity in a divorce
- Insurance companies seeking to eliminate risky individuals
- Other criminals and nuisances
- CIA, FBI, KGB, SABOT, SBL, SMERSH, KAOS, etc...
Sharing confidential data

- **Safe data**: Modify the data to reduce the risk of re-identification
- **Safe places**: Physical isolation and secure technologies
- **Safe people**: Training and Data use agreements
Safe Data

- 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
Further Resources: Safe Data

  http://www.fcsm.gov/working-papers/spwp22.html

- The American Statistical Association, Committee on Privacy and Confidentiality - Methods for Reducing Disclosure Risks When Sharing Data
  http://www.amstat.org/committees(pc)/SDL.html

- ICPSR's Confidentiality and Privacy web page
  http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/confidentiality/
Safe Places

- Secure Deposit Form
- Secure Processing Environment (SDE)
- Data protection plans
- Virtual data enclave
- Physical enclave
The Virtual Data Enclave (VDE) provides remote access to quantitative data in a secure environment.
Further Resources: Safe Places

- ICPSR “Instructions for Preparing the Data Protection Plan”
  http://www.icpsr.umich.edu/files/ICPSR/access/restricted/all.pdf

- “Introducing ICPSR’s Virtual Data Enclave (SDE)”

- ICPSR Physical Data Enclave
  http://www.icpsr.umich.edu/icpsrweb/content/ICPSR/access/restricted/enclave.html
Safe People

- Staff training
- Data use agreements
  - Responsible Use Statement
  - Research plan
  - IRB approval
  - Data protection plan
  - Behavior rules
  - Security pledge
  - Institutional signature
ICPSR Responsible Use Statement

Users of ICPSR data agree to a responsible use statement before downloading data from the Web site. It reads, in part:

• Any intentional identification of a RESEARCH SUBJECT (whether an individual or an organization) or unauthorized disclosure of his or her confidential information violates the PROMISE OF CONFIDENTIALITY given to the providers of the information. Therefore, users of data agree:

• To use these datasets solely for research or statistical purposes and not for investigation of specific RESEARCH SUBJECTS, except when identification is authorized in writing by ICPSR

• To make no use of the identity of any RESEARCH SUBJECT discovered inadvertently, and to advise ICPSR of any such discovery

• Agree not to redistribute data or other materials without the written agreement of ICPSR

Source: “Navigating Your IRB to Share Restricted Data” Webinar (http://bit.ly/Vi3RXd)
Restricted Data Use Agreement

- Sets requirements of the investigator and institution
- Defines ICPSR’s obligations
- Requires signatures from investigator and legal representative of researcher’s institution
- Incorporates by reference
  - Information entered into the access system
  - IRB approval or exemption for project
  - Data security plan

Source: “Navigating Your IRB to Share Restricted Data” Webinar (http://bit.ly/Vi3RXd)
Data Use Agreement

To avoid inadvertent disclosure of persons, families, households, neighborhoods, schools or health services by using the following guidelines in the release of statistics derived from the dataset.

1. In no table should all cases in any row or column be found in a single cell.
2. In no case should the total for a row or column of a cross-tabulation be fewer than ten.
3. In no case should a quantity figure be based on fewer than ten cases.
4. In no case should a quantity figure be published if one case contributes more than 60 percent of the amount.
5. In no case should data on an identifiable case, or any of the kinds of data listed in preceding items 1-3, be derivable through subtraction or other calculation from the combination of tables released.
Data Use Agreement

The Recipient Institution will treat allegations, by NAHDAP/ICPSR or other parties, of violations of this agreement as allegations of violations of its policies and procedures on scientific integrity and misconduct. If the allegations are confirmed, the Recipient Institution will treat the violations as it would violations of the explicit terms of its policies on scientific integrity and misconduct.
What are the consequences of violating the terms of use agreement for ICPSR data?

Subjects who participate in surveys and other research instruments distributed by ICPSR expect their responses to remain confidential. The data distributed by ICPSR are for statistical analysis, and they may not be used to identify specific individuals or organizations. Although ICPSR takes steps to assure that subjects cannot be identified, users are also obligated to act responsibly and not to violate the privacy of subjects intentionally or unintentionally.

If ICPSR determines that the terms of use agreement has been violated, one or more of the steps will be taken which may include:

- ICPSR may revoke the existing agreement, demand the return of the data in question, and deny all future access to ICPSR data.
- The violation may be reported to the Research Integrity Officer, Institutional Review Board, or Human Subjects Review Committee of the user’s institution. A range of sanctions are available to institutions including revocation of tenure and termination.
- If the confidentiality of human subjects has been violated, the case may be reported to the Federal Office for Human Research Protections. This may result in an investigation of the user’s institution, which can result in institution-wide sanctions including the suspension of all research grants.
- A court may award the payment of damages to any individual harmed by the breach of the agreement.
Customizing the RUDDDA

- Customize RUDDDA with information supplied by the data provider
  - Institution’s legal name and address
  - Official name AND familiar reference of project or dataset
  - Contact name of legal representative
  - Preference for electronic or hard copy

---

THE <UNIVERSITY/ORGANIZATION NAME>

AND

THE REGENTS OF THE UNIVERSITY OF MICHIGAN on behalf of the NATIONAL ADDICTION AND HIV DATA ARCHIVE PROGRAM, INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH

RESTRICTED-USE DATA DEPOSIT AND DISSEMINATION AGREEMENT

This Agreement, dated as of the [date], 2012 by and between <University/Organization Name>, having an office at <full address of university/organization> (hereinafter, the “Provider”) and the REGENTS OF THE UNIVERSITY OF MICHIGAN on behalf of the NATIONAL ADDICTION AND HIV DATA ARCHIVE PROGRAM, INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH (NAHDAP/ICPSR), having its principal place of business at The University of Michigan, Institute for Social Research, 1131 Perry Building, (330 Packard Street), Ann Arbor, Michigan, 48106-2321, defines the terms of the confidential relationship between the aforementioned parties.

WITNESSETH:

WHEREAS the National Institute on Drug Abuse (hereinafter referred to as NIDA) is the Federal focal point for research on drug abuse and addiction. NIDA’s mission is to lead the Nation in bringing the power of science to bear on drug abuse and addiction.

WHEREAS the National Addiction and HIV Data Archive Program, Inter-university Consortium for Political and Social Research (hereinafter referred to as NAHDAP/ICPSR) is a member-based consortium that provides data archiving and dissemination services to organizations such as NIDA;

WHEREAS the <NIDA Project or Dataset Name> (hereinafter referred to as <project/data familiar form or acronym>) is a project funded by the National Institute on Drug Abuse (hereinafter referred to as NIDA) that generates confidential data on participants and therefore requires protection of participant identities;

WHEREAS, in order to make <project/data familiar form or acronym> data available to qualified researchers and to assure that the <project/data familiar form or acronym> restricted-use dataset is appropriately protected, Provider intends to provide a copy of the <project/data familiar form or acronym> Dataset and related documentation to NAHDAP/ICPSR for the purpose of allowing NAHDAP/ICPSR to examine, process, and distribute the materials;

NOW, THEREFORE, for and in consideration of the premises and mutual promises and

---

Source: “Navigating Your IRB to Share Restricted Data” Webinar (http://bit.ly/Vi3RXd)
Accessing Restricted-Use Data

• Use online data access request system
  – Link in Access Notes on study homepage

• Must provide:
  – Name, department, and title of investigator
  – Description of the proposed research
  – Approval or exemption from IRB
  – Names of research staff accessing data
  – CVs and signed confidentiality pledges
  – Information on data formats needed and data storage technology

Source: “Navigating Your IRB to Share Restricted Data” Webinar (http://bit.ly/Vi3RXd)
Further Resources: Safe People

• Example NAHDAP Restricted Data Use Agreement
  http://www.icpsr.umich.edu/files/NAHDAP/GenericRDAAgreement.pdf

• NAHDAP “Restricted-Use Data Deposit and Dissemination Procedures”

• “Navigating Your IRB to Share Restricted Data” Webinar
  http://bit.ly/Vi3RXd
Sharing confidential data

- **Safe data**: Modify the data to reduce the risk of re-identification
- **Safe places**: Physical isolation and secure technologies
- **Safe people**: Training and Data use agreements
Thank you

lyle@umich.edu