ICPSR | INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH

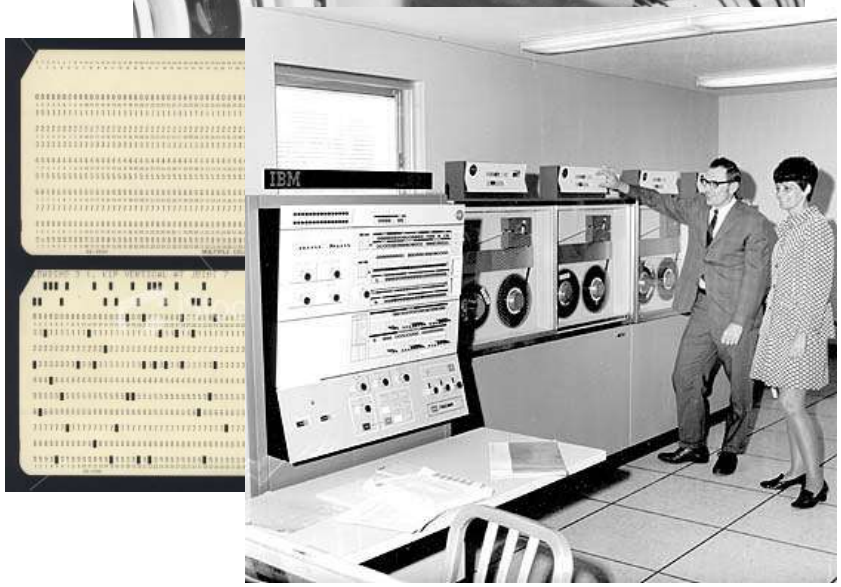
A PARTNER IN SOCIAL SCIENCE RESEARCH

Restricted Data in the Social Sciences

Prepared for the Breakout session on Privacy Issues vs. Public Access, NDIIPP, July 9, 2008

Evolution of Social Science Data

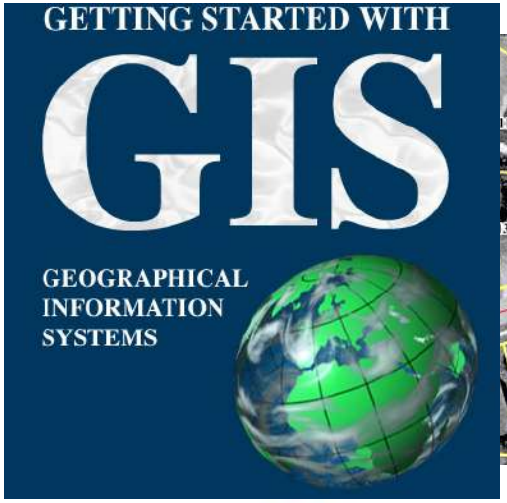
- Since the 1960's ICPSR has been disseminating public-use, quantitative data files

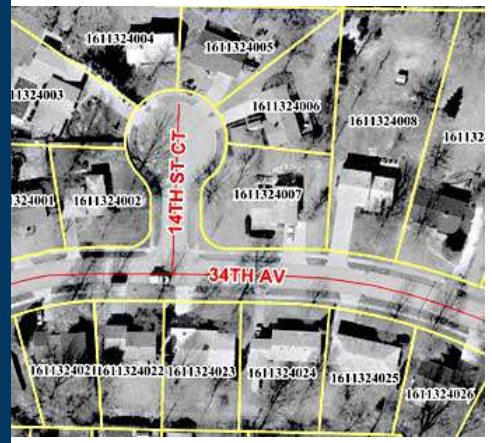


Evolution of Social Science Data

- Long-term investigations – many data points for a given individual
- Data about places, contexts
- Commercial data accessible for merging

WISCONSIN LONGITUDINAL STUDY



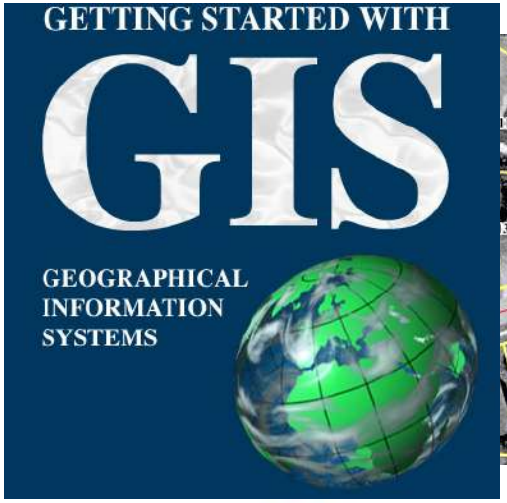


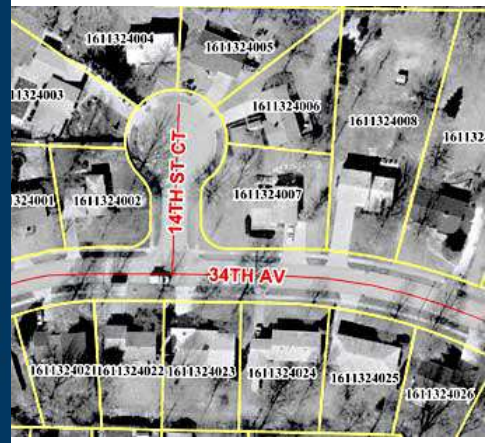
Google

Evolution of Social Science Data

- Increasing pressure to balance access to data with protection of confidentiality
- Large public investment in social science data

WISCONSIN LONGITUDINAL STUDY

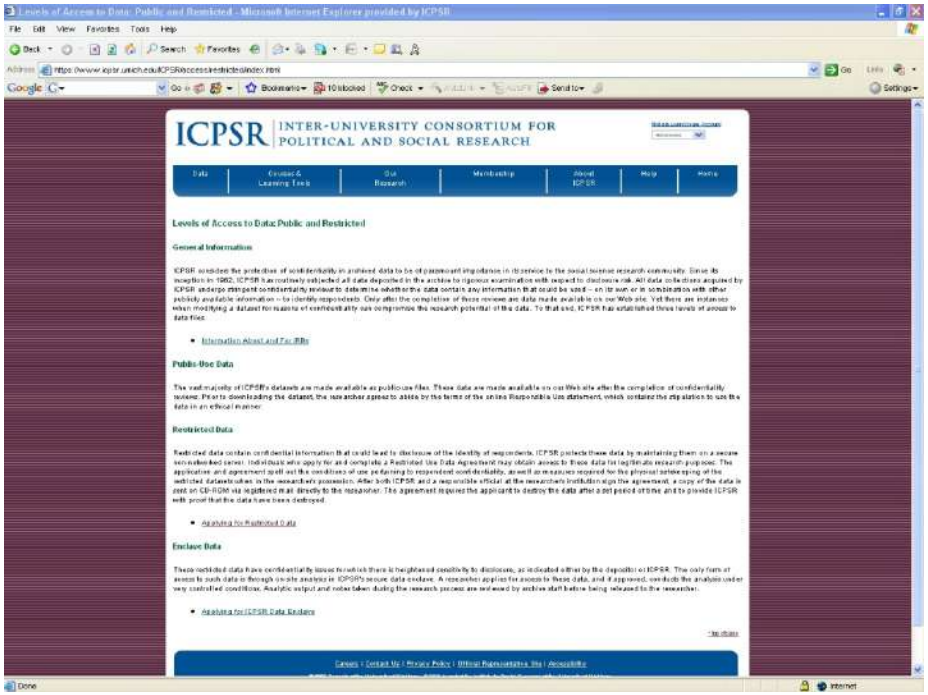




Google

Evolution of Social Science Data

- Around 2000, ICPSR developed restricted data dissemination procedures
- Not long after ICPSR's Data Enclave was established



Examples of Sensitive Data

- Exploring Women's Histories of Survival of Violence and Victimization in a Midwestern State, 2004-2005
  - Sensitive Topic
  - Narrow Geography
  - Qualitative information
- Michigan Student Study, 1990-1994
  - Known timeframe
  - Narrow sampling frame
  - Detailed demographic and educational data

Disclosure Analysis

- Attributes that can be known by an outsider (age)
  - Usually not attitudes or beliefs
- Availability of geographic data.
  - The more specific the geography, the more attention must be paid to disclosure risk.

Two Types of Disclosure Risk

- Type I disclosure: when an intruder has knowledge that a given person (or organization) is included in a survey and the intruder attempts to find this record.

- Type II disclosure: when an intruder does not know the identity ahead of time and uses externally available resources (linking databases) to attempt to find survey respondents.

Common Methods of Reducing Disclosure Risk

- Release a sample (of a larger data collection)
- “Collapse” data categories (e.g., combine categories, top- and bottom-coding, converting continuous variables to categories)

Methods of Reducing Disclosure Risk

- Suppress data that illuminate unique cases (e.g., blanking variables with sensitive information or removing unique cases altogether)

- Perturb values of the data (e.g., adding random noise or distortion, microaggregating, swapping cases, suppressing, and re-imputing data)

Restricting Access

- Restricted-Use Data
  - Restricted-use agreement
  - Application, Bonafied researcher with a research plan, IRB approval, Data security plan, Data mailed, User's institution legally bound in the agreement, Data destroyed after agreed upon time
- Secure Data Enclave
  - On-site analysis at ICPSR, non-networked computing environment, monitored, output reviewed

Data-PASS Results

- Data-PASS has dealt with primarily public-use data
- ICPSR has uncovered through LEADS a non-trivial number of “at risk” studies that have disclosure risk problems (e.g. audio clips, video clips, qualitative interviews)
- Out of scope for Data-PASS, but ICPSR will pursue many of these collections outside of Data-PASS.

Thank you!

For more information:

www.icpsr.umich.edu

Amy Pienta
apienta@umich.edu