ICPSR

SHARING DATA TO ADVANCE SCIENCE

Research Infrastructures: Ensuring trust and quality of data

Margaret C. Levenstein

Director, Inter-university Consortium for Political and Social Research

The initiatives described here are supported by the National Science Foundation (1744065 and 1525662) and the Sloan Foundation.

Data in the wild

- Organic or non-designed (found) data create new challenges for quality and trust
  - Not just increase in scale
- Data changes in real time
  - Requires snapshots, versioning
- No survey instrument or documentation of study design to provide metadata for re-use or discovery
  - Or even informed use of data the first time
  - Requires development of standards (e.g., extend DDI)
  - Citizen-scientist engagement

Research Infrastructures: ensuring trust and quality of data

- Provenance
- Preservation
- Privacy

- All more challenging in the new world of “found” data

Research Infrastructures: ensuring trust and quality of data

- Provenance
- Preservation
- Privacy

Research Infrastructures: ensuring trust and quality of data

- Provenance
  - Adapting (and using) standards for new kinds of data
    - Linked data
    - Social media and web-based data
- Preservation
- Privacy

Research Infrastructures: ensuring trust and quality of data

- Provenance
- Preservation
- Privacy

Research Infrastructures: ensuring trust and quality of data

- Provenance
- Preservation
  - Tension between openness and preservation
    - Feasibility
      - Individual researchers and institutions
    - Incentives
- Privacy

Research Infrastructures: ensuring trust and quality of data

- Provenance
- Preservation
- Privacy

Research Infrastructures: ensuring trust and quality of data

- Provenance
- Preservation
- Privacy
  - Safe data can be achieved in different ways
    - Important to be able to use sensitive data in safe ways or sensitive subjects and vulnerable populations are ignored
  - Match researchers to appropriate data and computing environment
  - Sanitize (synthesize) data for less trusted users
    - Critical for training purposes
  - Secure computing environment and differential privacy of *output* for trusted researchers

ICPSR initiatives: ensuring trust and quality of data

- LinkageLibrary
- SOMAR
- Researcher passport

Data linkage challenges

- Linked data present challenges for both confidentiality and reproducibility
  - Linkage more accurate with more detailed information
    - Need standards for safe, ethical ways to enhance data with new linkages
  - Linked data easier to re-identify, even after removing unique identifiers
    - Need safe places to analyze linked data
  - Linkage strategies introduce differences in datasets that are often not well documented


LINKAGE LIBRARY
Maintaining datasets to support the data linkage community

LINKAGE LIBRARY

- Encourage researchers to share linked (or linkable) data, and linkage strategies
  - Algorithms, code
- Compare approaches across projects, datasets, disciplines
  - Improve linkage practices
  - Improve transparency

SOMAR: Social Media Archive

- Addresses 4 communities who:
  - Study social media use specifically
  - Leverage social media data to understand people and society
  - Study social science methods
  - Investigate new methods for curation, publication, confidentiality and quality assessment, and long-term management of research data
- Archive enables historical and longitudinal analyses often missing from rapidly changing social medial platforms

SOMAR: Social Media Archive

- Archive data where possible
- Archive workflows and code where data sharing is prohibited
  - Eg: Twitter IDs and code for rehydrating
- Curation and metadata
  - Provenance, dates, hashtags, confidentiality protection

Researcher Passport

Establishing shared understanding of what it means to be a trusted researcher

Researcher Passport

- Researcher Passport: Improving Data Access and Confidentiality Protection
  - ICPSR's Strategy for a Community-normed System of Digital Identities of Access
  - https://deepblue.lib.umich.edu/handle/2027.42/143808
  - Identifies inconsistent language and policies that impede access
  - Facilitate sharing of *proprietary* data
- Passports for safe people
  - Verified identities, institutional affiliation, open badges
  - Training
  - Experience (good and bad)
- Visas to control access
  - Permission to "enter" (access) specific data specifying
    - Passport holder
      - Project, Place, Period

Questions

- How do we solve coordination problems?
  - Research across domains requires use of interoperable standards. How do we get that?
- Openness is limited by paywalls, but without resources long term preservation and access are not sustainable.
  - What's the appropriate balance between openness and sustainable preservation?

More information

- ICPSR help@icpsr.umich.edu
- Researcher Credentialing
  - Johanna Bleckman at Bleckman@umich.edu
- LinkageLibrary
  - Susan Leonard at hautanie@umich.edu
- SOMAR
  - Libby Hemphill at LibbyH@umich.edu

The initiatives described here are supported by the National Science Foundation (1744065 and 1525662) and the Sloan Foundation.

ICPSR

- Founded in 1962 by 22 universities, now consortium of 800 institutions world-wide
- Focus on social and behavioral science data, broadly defined
- Current holdings
  - 10,000 studies, quarter million files
  - 1500 are *restricted studies*, almost always to protect confidentiality
  - Bibliography of Data-related Literature with 75,000 citations
- Approximately 60,000 active MyData (“shopping cart”) accounts
- Thematic collections of data about addiction and HIV, aging, arts and culture, child care and early education, criminal justice, demography, health and medical care, and minorities