The Privatization of Data, Private Data, and Data Privacy

Margaret Levenstein
Inter-university Consortium for Political and Social Research
Credible Sources: Libraries and the Integrity of Knowledge - Global Resources Forum 2019
Center for Research Libraries
Chicago, Illinois May 23, 2019

Two directions: open data and private data

- Big push for open data and open science
- Big increase in research use of private data
- Why these seemingly contradictory trends?
- What approaches support goals of
  - Increasing creation of new knowledge
  - Respecting human rights
    - Right to privacy and right to participate in research

Open data

- 2013 Office of Science and Technology Policy memorandum
  - Called for open access to data
- Foundations for Evidence-Based Policymaking Act of 2018
  - Encourages agencies to inventory and make data available
- ICPSR
  - Founded in 1962 to provide access to data
    - ANES made available to 22 members of consortium
      - Shift from researchers asserting exclusive or priority use of data they collected
    - ~ 800 consortium members, 10K studies, 250K users per year

Why open data?

- Data is power
  - Access to data is critical to democracy
- Data is costly to produce, but most of that is a fixed cost
  - Access to data is efficient
- Data is necessary for reproducibility and knowledge building
  - Access to data is science

Private data and data privacy

- Researchers increasingly make use of private data
  - Private because it *belongs* to a company that asserts control over it
  - Private because it contains information about individuals that they might not want to be public
- Academic journals in economics
  - Required data sharing, for transparency and reproducibility
  - Found 1/3 of empirical articles requested waiver
    - Data belonged to someone else
    - Data contained confidential information

Public data and privacy

- Increasing concern over the risk that “anonymized” will be re-identified
- Driving factor in Census Bureau announcements re changes in their production of public data products
  - Increasing computational power and availability of information about individuals and households
  - Differential privacy and the database reconstruction theorem
    - Citizenship question increases the sensitivity of the information at risk

Protection and access to private data

- Long-standing arrangements for access to confidential data
  - Each involves both a technological and a social component
- Local computing on secured, stand alone computers
  - Data use agreements
    - Enumerate researcher and institutional responsibilities and consequences
  - Encrypted CDs or download
  - Researcher responsible for disclosure review
- Physical enclaves
  - Data use agreements
    - Enumerate researcher and institutional responsibilities and consequences
  - Controlled computing environment
  - Third party disclosure review

Emerging arrangements for accessing confidential data

- Virtual data enclaves
  - Data use agreements
    - Researcher Passport
  - Controlled computing environment accessed from local computer
  - Third party disclosure review
- Secure on-line computing
  - Analysis of data that the researcher cannot see
  - Automated disclosure review, with minimally necessary noise infusion
  - Secure multi-party computing
    - Computationally very intensive
  - Requires highly processed and interoperable data
    - Difficult to use with non-designed data without large up-front investment whose appeal is
      - Digital traces of human activity
      - Available essentially immediately

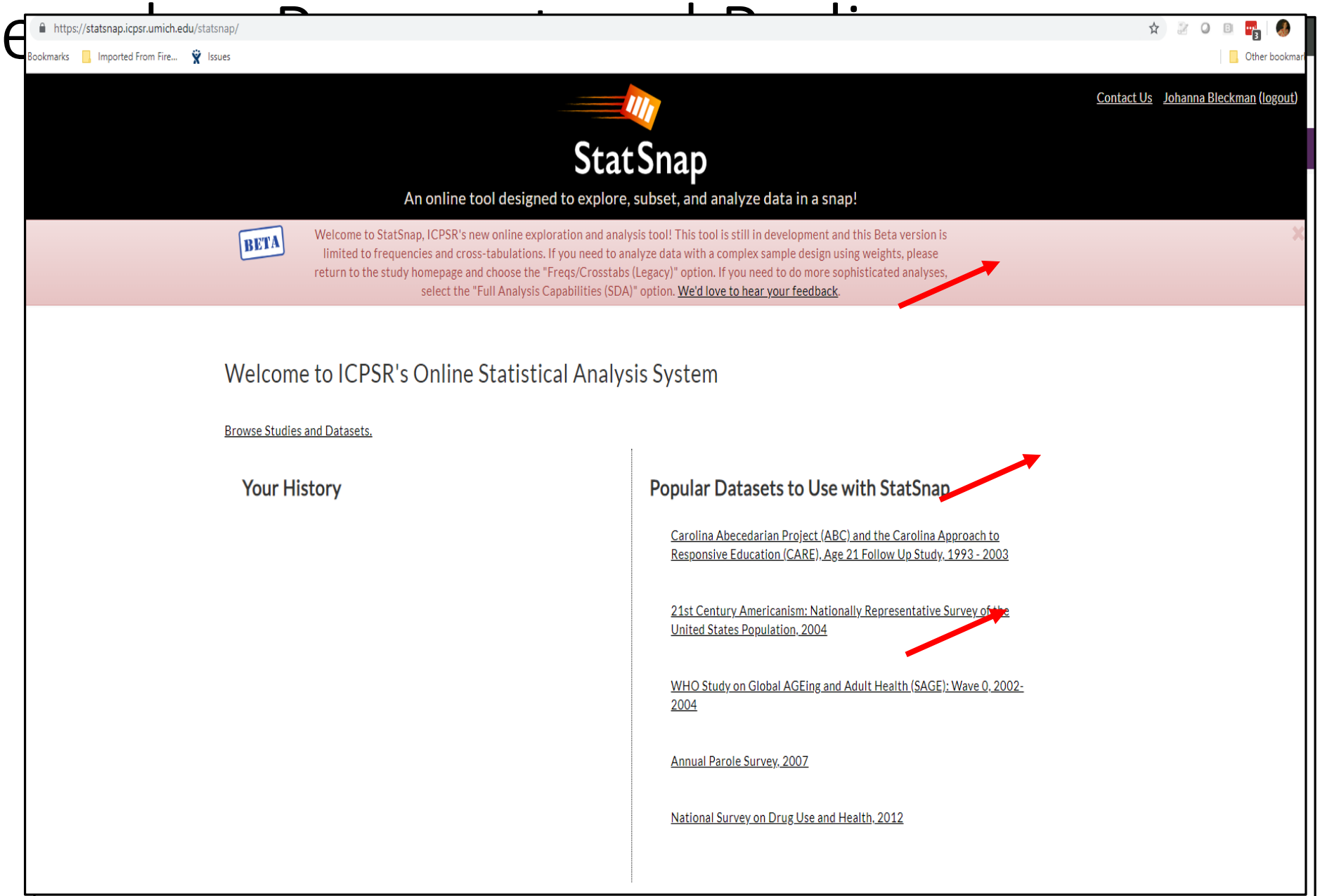
Researcher Passport

- 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
- Passports for safe people
  - Verified identities, institutional affiliation
  - Training
  - Experience (good and bad)
- Visas to control access
  - Permission to "enter" (access) specific data specifying
    - Passport holder
      - Project, Place, Period

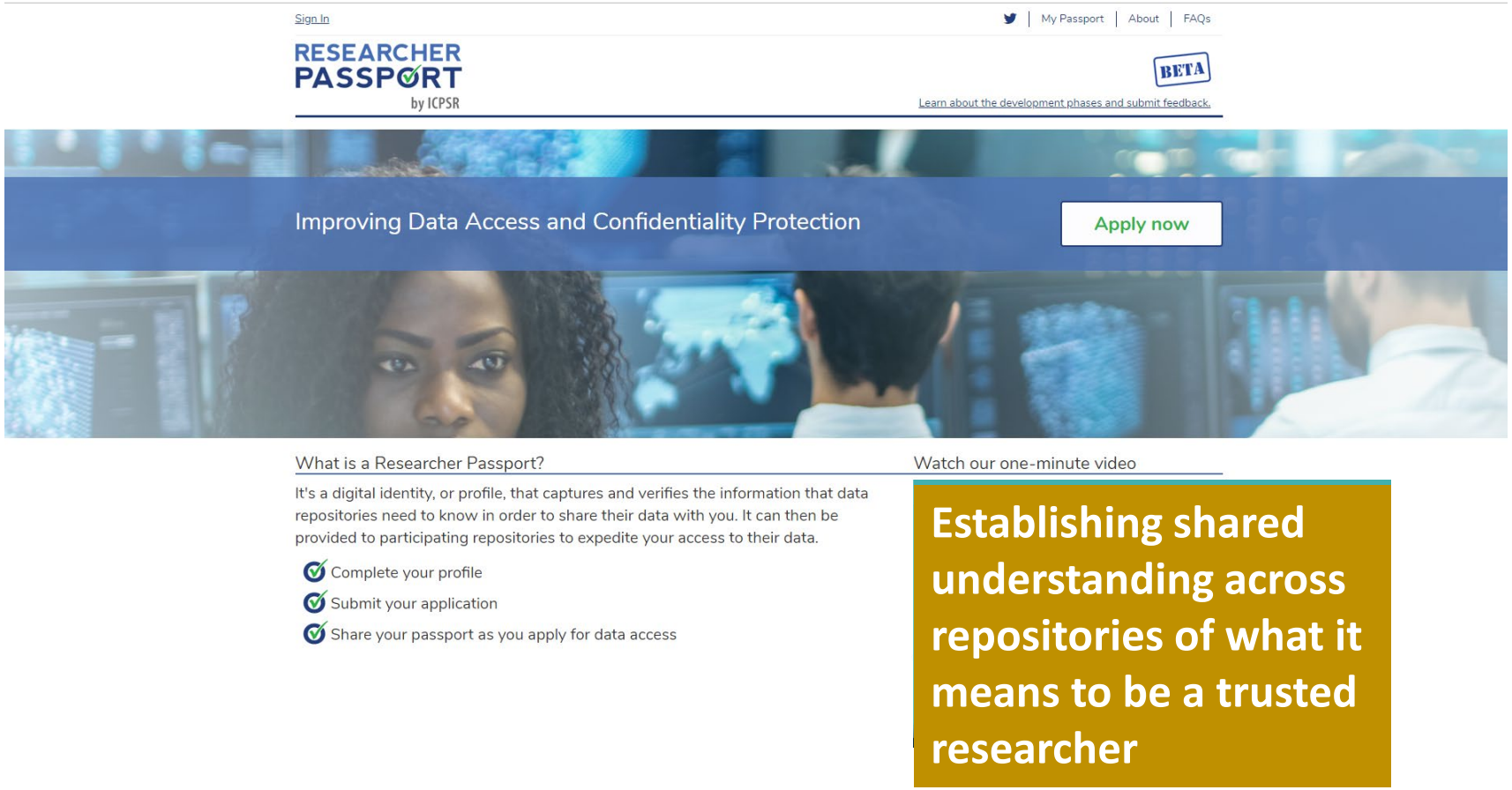
Researcher passport

- Digital identifiers associated with ICPSR MyData accounts
  - Open badges
  - Integrated into restricted data application process
- Repositories issue visas to access restricted data
  - Visa controls access to secure download or VDE
  - Visa issued after signed DUA
- Establishing content of training requirements
  - Training in confidentiality protection
  - Badges earned for access to data of different types or sensitivity
- Governance system to provide due process for allegations of misconduct

Rese


https://statsnap.icpsr.umich.edu/statsnap/
Bookmarks
Imported From Fire...
Issues
Other bookmar
Contact Us
Johanna Bleckman (logout)
StatSnap
An online tool designed to explore, subset, and analyze data in a snap!
BETA
Welcome to StatSnap, ICPSR's new online exploration and analysis tool! This tool is still in development and this Beta version is limited to frequencies and cross-tabulations. If you need to analyze data with a complex sample design using weights, please return to the study homepage and choose the "Freqs/Crosstabs (Legacy)" option. If you need to do more sophisticated analyses, select the "Full Analysis Capabilities (SDA)" option. We'd love to hear your feedback.
Welcome to ICPSR's Online Statistical Analysis System
Browse Studies and Datasets.
Your History
Popular Datasets to Use with StatSnap
Carolina Abecedarian Project (ABC) and the Carolina Approach to Responsive Education (CARE), Age 21 Follow Up Study, 1993 - 2003
21st Century Americanism: Nationally Representative Survey of the United States Population, 2004
WHO Study on Global AGEing and Adult Health (SAGE): Wave 0, 2002-2004
Annual Parole Survey, 2007
National Survey on Drug Use and Health, 2012

Researcher Passport



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

Building models of access to private data

- Trusted intermediaries
  - Credentialed researchers
  - Privacy protecting technologies
  - Cooperation from data custodians?
- Public sector
  - Foundations for Evidence-Based Policy Act of 2018
    - Federal Statistical Research Data Center network
  - State and local governments
    - Patchwork of arrangements
- Private sector

Researcher access to private sector data

- Abraham, Levenstein and Shapiro (2019) study of commercial data used in economic research
  - Types of agreements depended on type of data
    - Aggregators of transactions
    - Individual companies
    - Service providers
    - Information services
    - Social media
    - Webscraping

Survey results: Cost

No meaningful disclosable results

- Pilots at relatively low cost
- Production costs (would be) significant
- Data obtained through webscraping are free

Survey results: Sharing

- Most agreements do not allow data posting
- Most agreements do not allow data sharing
- Most data could be available to others with separate agreements
- Most data with agreements have limits on use
- DUAs generally not shareable

Survey results: Archiving

Archiving of data, roughly evenly split between

- Provider
- User (sometimes government agency)
- Not archived

Concern that agency archiving practices not well-suited to track provenance and versions of commercial data

Concerns for firms

- Cost (doing this requires time and resources)
- Information has business value; would prefer to sell or monetize data
- Disclosure has risks
  - PII disclosure risk
  - SEC/material information in the financial market sense
  - Business strategy disclosure risk
- Concern that data will be used for regulation or enforcement
- Firms are not monolithic; the chief economist might think it is a great idea, the CFO might not
- Fatigue: Too many requests

Solutions?

- Templates and standards
  - For agreements
  - For data and meta-data
  - For transmission
  - Universities, funders, learned societies, journals must support standards
- Credible burden reduction by leveraging business information systems
- Trusted intermediaries
  - Archive and access stale data for research

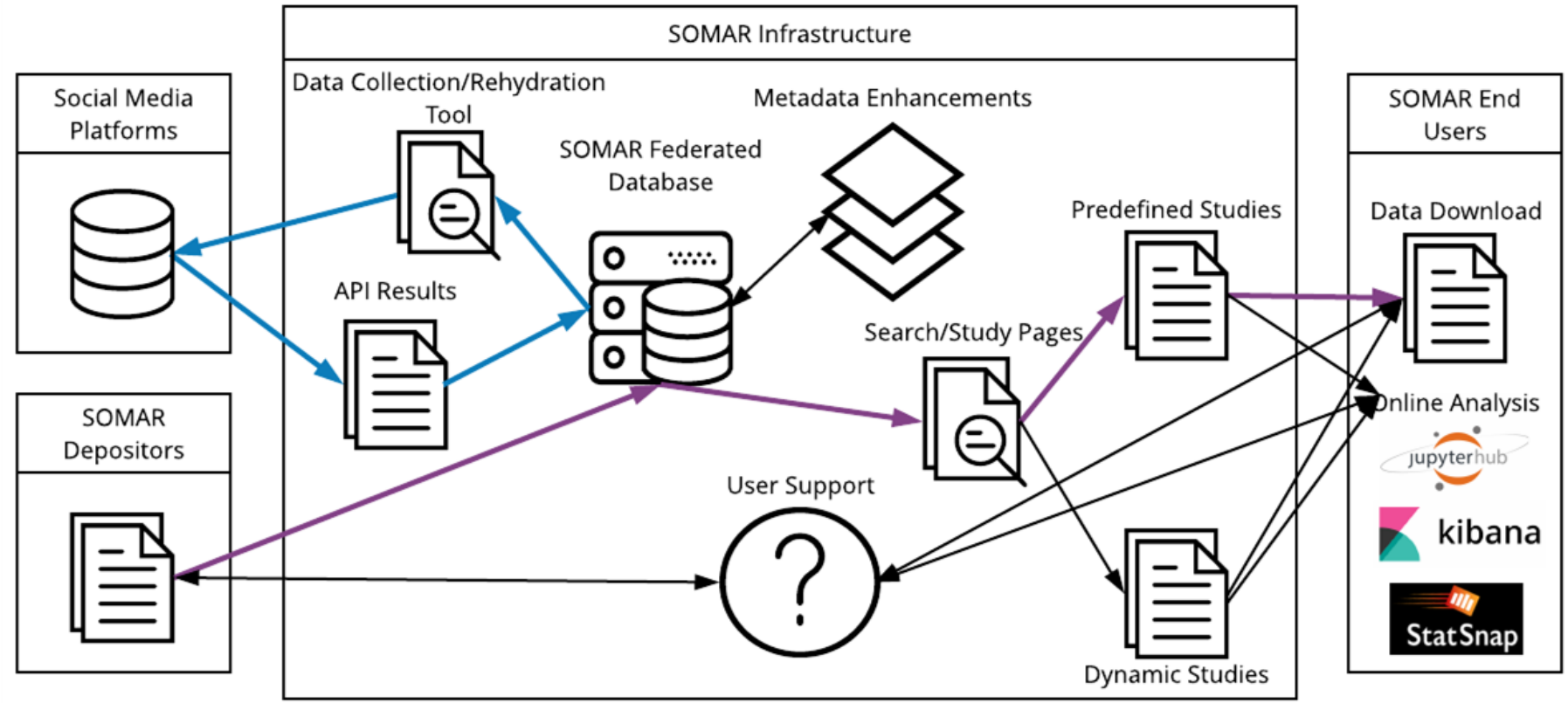

SOMAR
Social Media Archive @ ICPSR

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


SOMAR Infrastructure
Social Media Platforms
SOMAR Depositors
Data Collection/Rehydration Tool
API Results
SOMAR Federated Database
Metadata Enhancements
Search/Study Pages
Predefined Studies
User Support
Dynamic Studies
SOMAR End Users
Data Download
Online Analysis
jupyterhub
kibana
StatSnap

SOMAR Challenges

- Technical infrastructure
- Ethical and legal infrastructure
- Metadata enhancements
- Adoption

Data as scientific asset

- Building, and building on, data communities
  - https://sr.ithaka.org/publications/data-communities/
  - SOMAR, LinkageLibrary are new examples
  - ICPSR's topical archives in aging and criminal justice are decades old examples
- Cross-disciplinary, cross-university collaborations around research problems and questions
  - Motivation for researchers and data owners to share
  - Requires systematic support from all parts of the research eco-system, including libraries, deans, editors

Sustainable data

- Libraries make resources sustainable
  - Provide curation and preservation of resources
  - Limit access to defined cohort, usually tied to how resources are provided
    - Local public libraries to local residents
    - Universities to students, faculty, and staff
- Data access needs to be sustainable
  - Privatization is one model
    - Paying customers get access
  - Building sustainability into data communities is a different model
    - Universities are the right level of aggregation for some, but not most, data
    - Institutional repositories work for storage and preservation, not access and re-use
      - Research is cross-institution and cross-discipline
    - Need pipelines to curate and share data, analogous to inter-library loan