Background

- Cornell Institute for Social and Economic Research (CISER):
  - Data and Computing Support for Social and Economic Researchers at Cornell University
  - Founded in 1981
  - Extensive Data Archive of Social and Economic Data
- Bill Block:
  - Director of CISER
  - Historical Demographer
  - 20 years with the Minnesota Population Center, IPUMS, NHGIS, IHIS
- Stefan Kramer:
  - Research Data Management Librarian at CISER (since May)
  - before that: Social Sciences Data Librarian at Yale U., Dir. of Library Services at Fielding Graduate U., user services specialist at U. of WA Network Information Center
Challenges

There are many, but...

- Growing CISER’s ability to meet the data needs of Cornell researchers
  - World class researchers
  - ever-more complex and large data questions
- Metadata lies at the heart of our strategy...forms the basis for our talk today
The Lifecycle of Social Science Research Data: Enabling Discovery through Metadata and Search Tools

William C. Block & Stefan Kramer
Cornell Institute for Social and Economic Research (CISER)
Lifecycle of social science research data

Research study is conceived and planned, methodologies selected, funding sources explored

Existing data sources are sought and explored – also happens for basic research needs

Research instruments are designed; data are collected through surveys, interviews, etc. – and from existing data sources

Collected data are merged, cleaned, analyzed, subsetted, coded, harmonized, linked, etc.

By search tools utilizing metadata from data stores, new research data becomes available for finding and exploring by researchers

Ideally begins early in data lifecycle to assure long-term preservation and access of data. One activity is metadata preparation and its exposure to external search tools

Final datasets are deposited for long-term preservation – e.g., into institutional or domain repository

Final datasets are made publicly accessible – e.g. via researcher’s and/or department’s and/or journal publisher’s web site
Data management

- Includes activities through the data lifecycle to assure that data remain or become understandable, usable, accessible, and findable – by the researchers compiling and analyzing the data themselves, and others for re-use or verification – such as:
  - Establishing naming and labeling conventions for variables, files, directory structures
  - Documenting newly recoded and computed variables
  - Creating policies about retention of files (data, analyses commands, table & chart output) and associated documentation, questionnaires, etc.
  - Determining appropriate file formats for analysis & processing (current research project use) and long-term preservation
  - Migrating files to different formats to preserve their usability with available software
  - Creating and maintaining metadata (about the data) that can eliminate duplication of work (e.g., having to repeat entry of text in questionnaire design and later in statistical analyses scripts) and make data discoverable without need to open proprietary-format data files

- Better to start at earlier stages of data lifecycle than try to “retrofit” later!
Researchers and metadata creation/maintenance

- Researchers will tend to describe their data only as much as necessary for their own use, for current project
- But: no one knows their data better than they do
- Needed: easy-to-use tools, and outreach to researchers, for sustainable metadata production – some actions may be performed by researchers, others by their institution’s data service providers
Researcher buy-in is essential for data archiving

“Archives that preserve and disseminate social and behavioral data perform a critical service to the scholarly community and to society at large, ensuring that these culturally significant materials are accessible in perpetuity. The success of the archiving endeavor, however, ultimately depends on researchers’ willingness to deposit their data and documentation for others to use.”

http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf

Ideally, the archiving endeavor achieves researcher buy-in in all lifecycle stages involving data management activities – not just at the final point of archival deposit.
Challenges of finding data 1: institutional catalogs may contain pointers to data, but are focused on other types of content.

Searching for texts (or images, or videos) differs from common search needs for social science research data.
Challenges of finding data 2: there are many data-focused archive catalogs … but often as “information silos”

Different search inputs, different search outputs, no easy way to search all at once, and not in “data-targeting” ways
Desirable search or browse functions for numeric data in social sciences

Not (easily) offered by most data catalogs, but often needed by data searchers, in addition to topic ... such as:

- Time span (example: 1970 - present)
- Time frequency (example: annually)
- Geographic extent (example: all of United States)
- Geographic granularity (example: county level)
- Methodology, sample (example: survey of adults aged 18-24)

http://ssrs.yale.edu/statcat/
Data Documentation Initiative (DDI)

- **DDI 3** designed to support the social science data lifecycle with metadata
- Powerful – but also **complex**! Used by national statistical agencies, data archives, etc.
- **Tools** for using DDI being developed – choosing the right ones for specific institutional needs is key
- Has the elements to capture information targeted in social science data searches

Exposing and indexing the holdings of data archives and publications in standardized metadata formats could enable web-scale discovery through new cross-collection search engine functions built to exploit that metadata.

Better Search & Discovery

Search for data
about: ___
From (year): ___
To (year): ___
In (geography): ___
at the level of: ___
Collected via: ___
etc., etc.: ___
Linking of research data with papers, articles, dissertations, etc.

- Data is one “raw material” behind published research
- Bidirectional links between research results and research data would enhance discovery of both – finding publications could help find data and vice versa
- Challenge: creating and maintaining these links
Thank you for your time & attention!

William C. Block  
block@cornell.edu

Stefan Kramer  
stefan.kramer@cornell.edu