1. DGI-Konferenz, 62. DGI Jahrestagung
Semantic Web & Linked Data
Elemente zukünftiger Informationsinfrastrukturen

Challenges and Opportunities in Social Science Research Data Management

Stefan Kramer
Research Data Management Librarian
Cornell Institute for Social and Economic Research
What is CISER?

The Cornell Institute for Social and Economic Research was founded in 1981. Our mission is to anticipate and support the evolving computational and data needs of Cornell social scientists and economists throughout the entire research process and data life cycle.

More at:
http://ciser.cornell.edu/About_CISER.shtml
Where is Cornell University?

Source: http://www.cornell.edu/maps/state.cfm
Sharing & preserving data: why (would researchers want to)?

- Collaboration with fellow researchers on current projects
- Future use/access by others (public/limited, open/restricted) and self
- Making research findings replicable, help avoid duplication
- Requirements from funding agencies, journal publishers, own institution
- May help in tenure/promotion process
- Making research data citable
Some potential problems with *own* data (that’s not (well) managed) for researchers

Where is it?

- How is it?
- (How) can/may I give it to someone else?

Hmm, what format were those files in again?

Can my computer and software still read/open/use it?

Where is the graduate assistant who organized, analyzed, … the data now?
Working with faculty to deposit data

- In local/institutional AND domain/subject repositories … e.g. eCommons@Cornell AND ICPSR
- Domain/subject repositories are not infallible, so institutional repositories provide a “backup” too
The lifecycle of research data

- Creation, (re)use and discovery of research data often follows a predictable flow.
- Supporting this flow with integrated tools and services can make social science research based on data more efficient and effective.

Source: http://hdl.handle.net/1813/17472
Example: enhancing search & discovery stage of research data lifecycle

Source: http://www.loc.gov/marc/bibliographic/bd008c.html

http://yufind.library.yale.edu/yufind/
Search/browse functions for numeric data in social sciences

Not (easily) offered by most search systems, incl. library catalogs, but often needed by data searchers, in addition to topic:

- Time span (example: 1970-present)
- Time frequency (example: annually)
- Geographic extent (example: all of United States)
- Geographic granularity (example: county level)
Researchers and metadata creation/maintenance

- Researchers will tend to describe their data 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 long-term access and preservation – some actions to be performed by researchers, others by institutional data service providers.
Data Documentation Initiative (DDI)

- Earlier versions of DDI focused on *codebooks* – the “manual” for datasets
- **DDI 3** designed to support the 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 going to be key

Source: http://www.ddialliance.org/
Making research data available for web-based analysis

- Most repository platforms make content, incl. datasets, available for downloading.
- But for many audiences, such as introductory methodology classes or “the public,” analysis of downloaded data is asking too much (lacking software or skills).
- Possible solution: web-based analysis, exploration, visualization of locally created data, e.g. through Berkeley SDA or Google Fusion Tables.

Output generated via http://sda.berkeley.edu/quicktables/quickconfig.do?gss08
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.

From ICPSR’s Bibliography of Data-Related Literature (http://www.icpsr.umich.edu/icpsrweb/ICPSR/citations/)
A few other issues in research data management and dissemination:

- Policies for data submission and dissemination - privacy, human subjects protection
- Ownership of datasets – does researcher, university, funding agency, … own?
  - What if datasets from multiple sources are merged/joined?
- Versioning and derivatives -
  - Keep every version of a dataset (one with originally collected information, next with added recoded/computed variables, etc.)?
  - Keep original dataset file (e.g., Microsoft Access from phone interview), also files generated from that (e.g., Stata for statistical analyses, etc.)?
- Capturing the data’s transformation, analysis, and analysis output
  - Relationships of data, command/syntax, and output files
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

From ICPSR’s Bibliography of Data-Related Literature (http://www.icpsr.umich.edu/icpsrweb/ICPSR/citations/)
Thank you for your time & attention!

The End

Stefan Kramer
stefan.kramer@cornell.edu
http://www.linkedin.com/in/kramerstefan