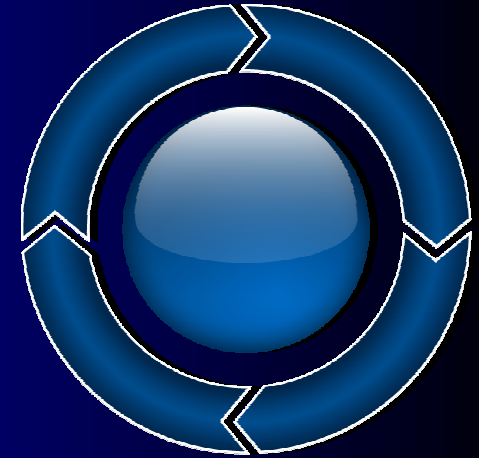


Opportunities and challenges across the research data lifecycle, and related activities at Cornell University



William C. Block & Stefan Kramer
Cornell Institute for Social and Economic
Research (CISER)

- **Bielefeld Colloquium on Knowledge Infrastructure, 2010-10-15**

What is CISER?

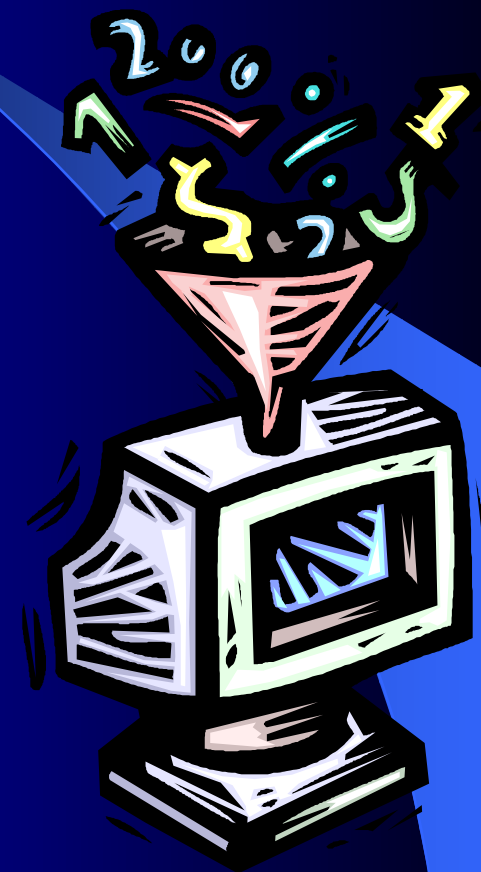


Cornell University
Cornell Institute for Social and Economic Research

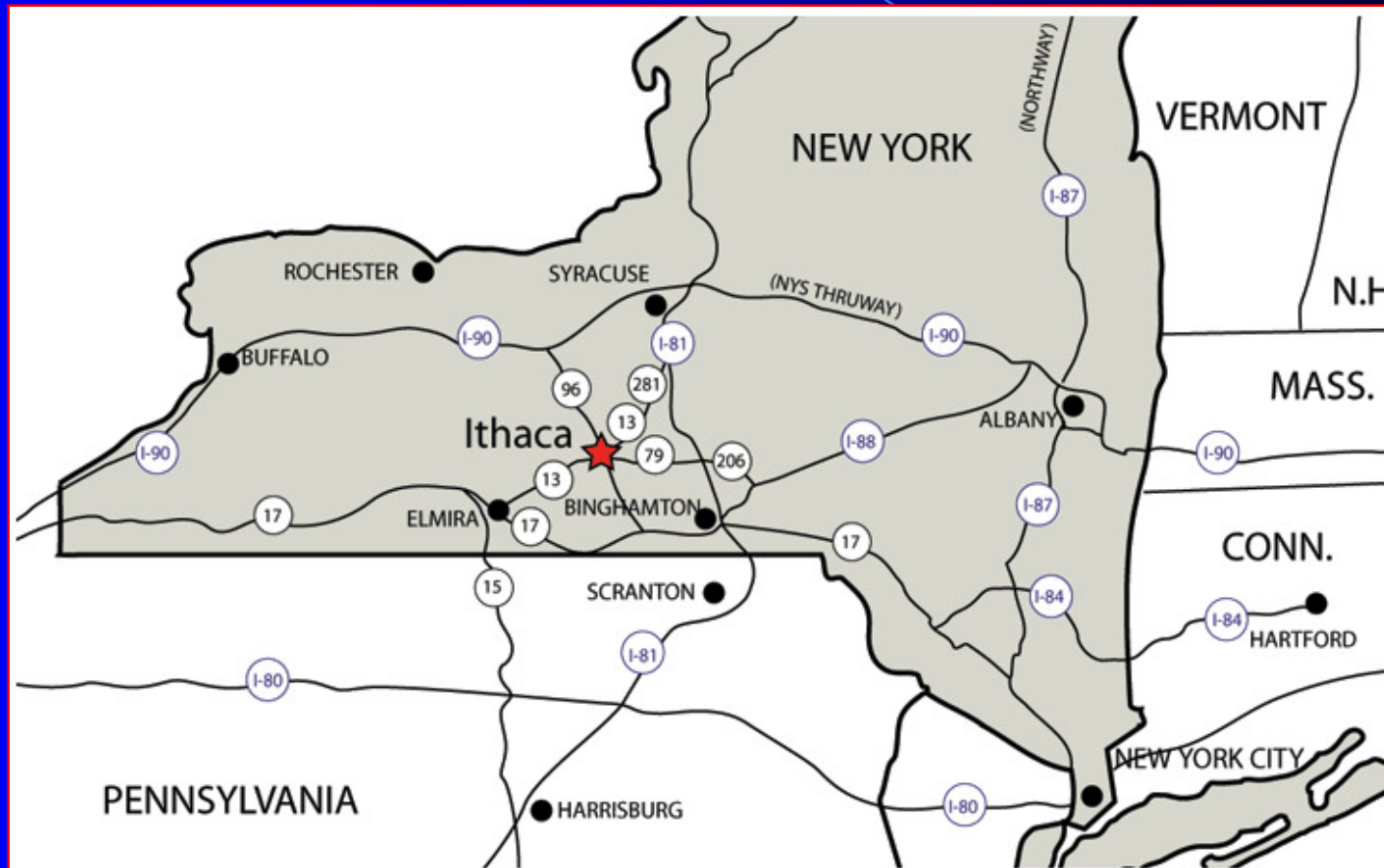
What is CISER?

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/>



Where is Cornell University?



Source: <http://www.cornell.edu/maps/state.cfm>

Some potential problems with *own* data (that's not (well) managed) for researchers

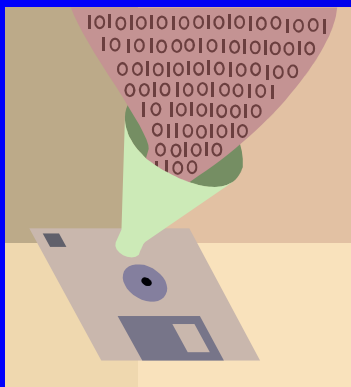
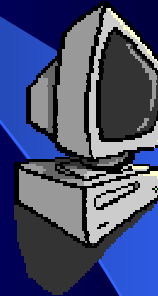


Where is it?



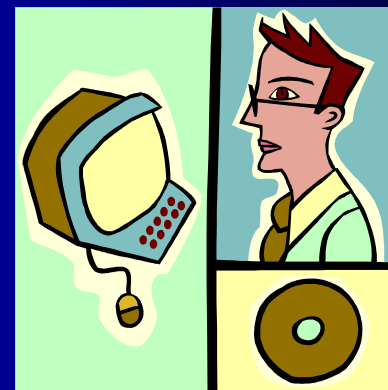
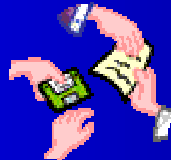
How safe is it, where(ever) it is?

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



Hmm, what format were those files in again?

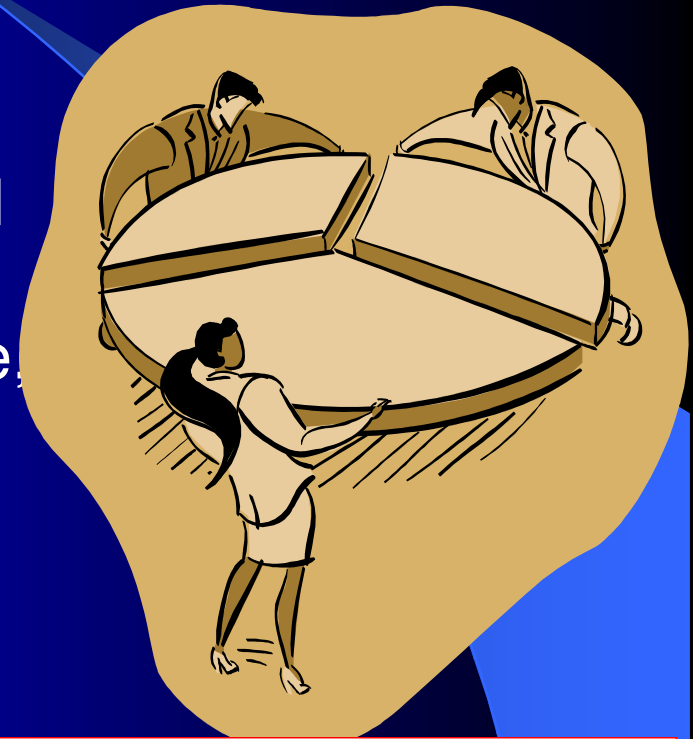
(How) can/may I give it to someone else?



Where is the graduate assistant who organized, analyzed, ... the data now?

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

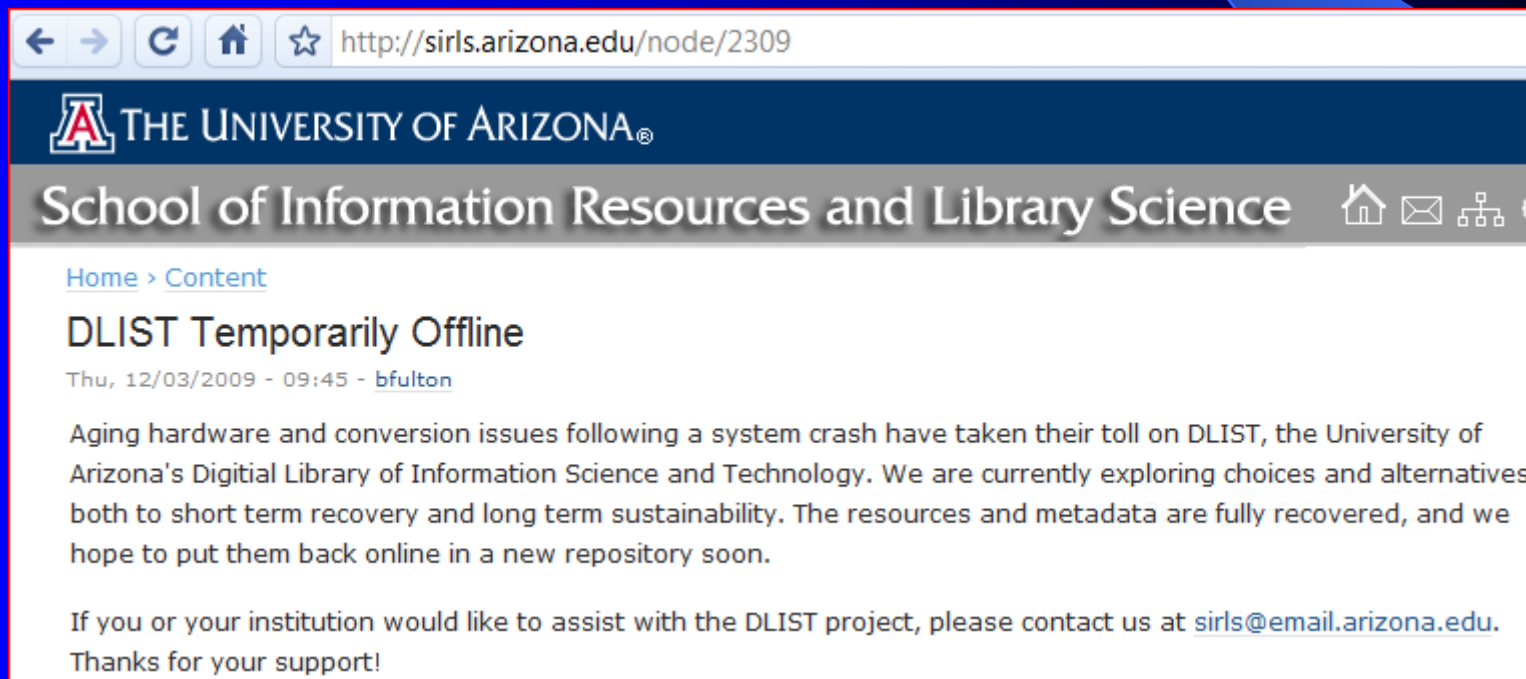


Bibliographic Citation: Hofferbert, Richard I. SOCIO-ECONOMIC, PUBLIC POLICY, AND POLITICAL DATA FOR THE UNITED STATES, 1890-1960 [Computer file]. Conducted by Cornell University Center for International Studies. ICPSR ed. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [producer and distributor], 1977. doi:10.3886/ICPSR00015

(Why cite the data?)

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 screenshot shows a web browser window with the address bar containing <http://sirls.arizona.edu/node/2309>. The page header includes the University of Arizona logo and the text "THE UNIVERSITY OF ARIZONA®" and "School of Information Resources and Library Science". The main content area features a blue link "Home > Content" followed by the heading "DLIST Temporarily Offline" and a timestamp "Thu, 12/03/2009 - 09:45 - [bfulton](#)". The body text reads: "Aging hardware and conversion issues following a system crash have taken their toll on DLIST, the University of Arizona's Digital Library of Information Science and Technology. We are currently exploring choices and alternatives both to short term recovery and long term sustainability. The resources and metadata are fully recovered, and we hope to put them back online in a new repository soon." At the bottom, it says: "If you or your institution would like to assist with the DLIST project, please contact us at sirls@email.arizona.edu. Thanks for your support!"

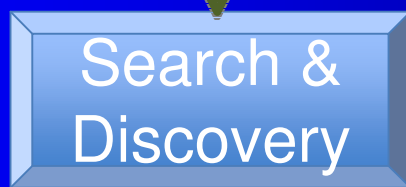
Lifecycle of social science research data

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



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

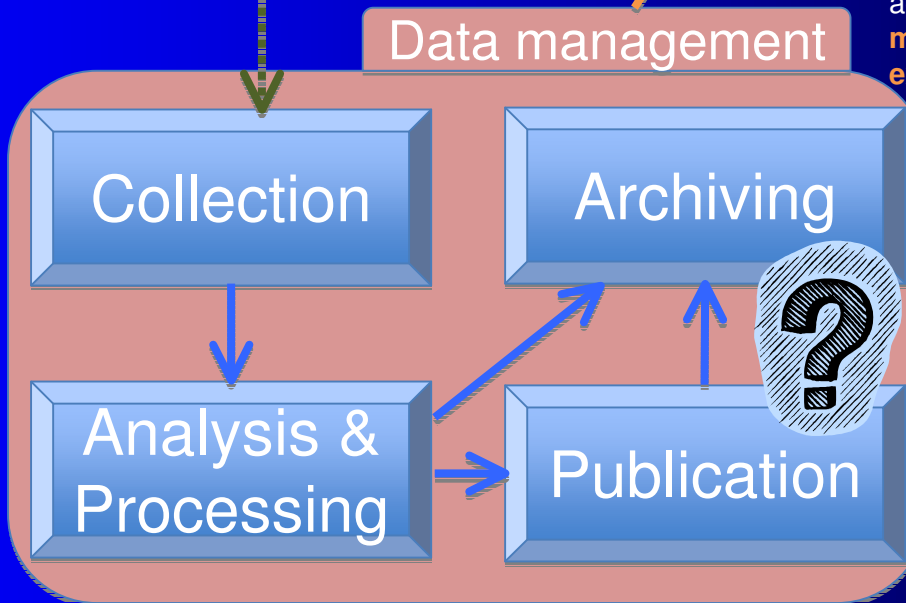
Existing data sources are sought and explored – also happens for basic research needs



Meta data

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

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



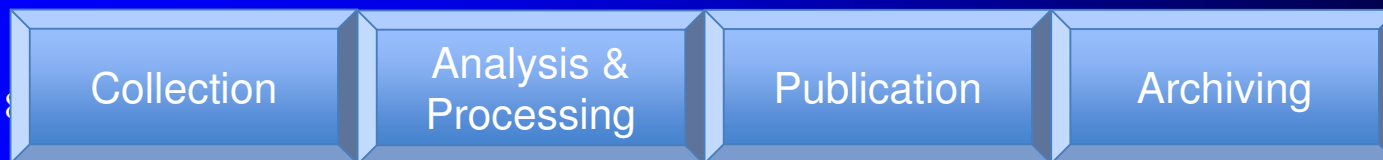
Final datasets are deposited for long-term preservation – e.g., into institutional or domain repository

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

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
 - 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)
- Better to start at earlier stages of data lifecycle than try to “retrofit” later!



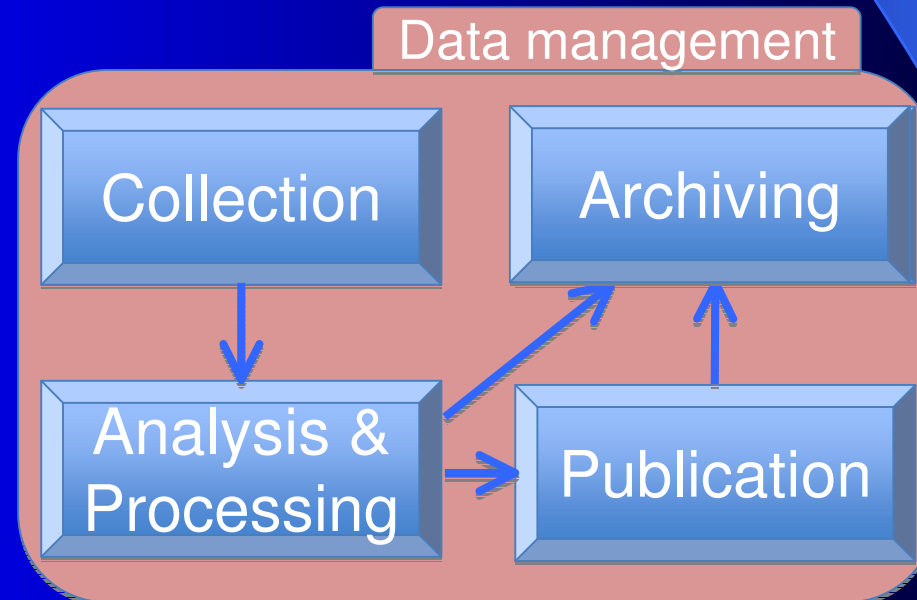
Researcher buy-in is essential for data archiving

“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.**”

ICPSR Guide to Social Science Data Preparation and Archiving: 4th Edition, p. 3

<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.*



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



Collection

Analysis &
Processing

Publication

Archiving

Possible improvement in the *search & discovery* stage of research data lifecycle from the university library end

Library of Congress >> MARC >> Bibliographic >> 00X >> 008 >> 008 (Computer Files)

008 - Computer Files (NR)

MARC 21 Bibliographic - Full

Indicators and Subfield Codes
Field has no indicators or subfield codes; the data elements are positionally defined.

Character Positions

18-21 - Undefined (006/01-04)	27 - Undefined (006/05-08)
22 - Target audience (006/05)	28 - Government publication (006/09-12)
# - Unknown or not specified	# - Not a government publication
a - Preschool	a - Autonomous or semi-autonomous
b - Primary	c - Multilocal
c - Pre-adolescent	f - Federal/national
d - Adolescent	i - International intergovernmental
e - Adult	l - Local
f - Specialized	m - Multistate
g - General	o - Government publication
j - Juvenile	s - State, provincial, territorial
- No attempt to code	u - Unknown if item is a government publication
	z - Other
	- No attempt to code

23-25 - Undefined (006/06-08)	29-34 - Undefined (006/13-18)
26 - Type of computer file (006/09)	
a - Numeric data	
b - Computer program	
c - Representational	

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

<http://yufind.library.yale.edu/yufind/>

The screenshot shows the yufind search interface. At the top, the Yale University Library logo is visible with the text "yufind BETA search the library's catalog". A search bar contains the text "primary election". Below the search bar, the results are displayed under the heading "Narrow Search Results".

Author

- Coleman, Kevin J. (5)
- Davis, James W., 1920- (4)
- Cook, Rhodes, 1948- (3)
- Gorman, Joseph B. (3)
- Buckalew, Charles Rollin, 1821-1899. (2)
- more ...

Format

- Electronic (off-line) (183)
- Books/Pamphlets (179)
- Journals/Magazines/Newspapers (41)
- Microforms (35)
- Video (8)
- more ...

Language

- English (183)

On the right side of the interface, there is a section titled "Showing 1 - 20 of 449 for primary election". Below this, there are suggestions for "Did you mean: primary education", "primary prevention". Two search results are shown:

- Yes we can? : white racial framing and the 2008 presidential campaign /** by Wingfield, Adia Harvey, 1977- Published 2010
Call Number: E906 .W56X 2010 (LC)
Located: LSF- click "Place Requests" for delivery to any Yale library
Not Checked Out
Books/Pamphlets
- The Voting Rights Act of 1965 /** by Hillstrom, Laurie Collier, 1965- Published 2009
Call Number: JK1924 .H55X 2009 (LC)
Located: SML, Stacks, LC Classification
Not Checked Out
Books/Pamphlets

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)

ACADEMIC MEDIA & TECHNOLOGY
SOCIAL SCIENCE RESEARCH SERVICES
YALE UNIVERSITY LIBRARIES
SOCIAL SCIENCE LIBRARIES AND INFORMATION SERVICES

StatCat

Statistical Data Finder

[Simple Search](#) | [Advanced Search](#) | [Help Searching StatCat](#) | [Help Using Data](#) | [About StatCat](#)

[Search results](#) (9 items)

Title: New Democracies Barometer III (1993-94)
Author: Centre for the Study of Public Policy
Holdings available on: [[Statlab Server](#)] [[all holdings](#)]
Abstract: Data from the third New Europe Barometer, described at <http://www.abdn.ac.uk/cspp/nebo.shtml>.
Series name: New Democracies Barometer
Series information: "The Centre for the Study of Public Policy and the Paul Lazarsfeld Society, Vienna, cooperated in launching a major multi-national survey, the New Democracies Barometer (NDB), to monitor the response of people caught up in the transformation of their polity, economy, society and often state boundaries too. Five NDB surveys were conducted between 1991 and 1998. Changes in Europe have been matched by changes in the New Europe Barometer survey. After the fifth round, the CSPP took responsibility for conducting surveys of post-Communist countries seeking membership in the European Union. It has conducted NEB rounds in 2001 and the winter of 2004/5."
Related publications: DIVERGING PATHS OF POST-COMMUNIST COUNTRIES: NEW EUROPE BAROMETER TRENDS SINCE 1991 - <http://ssrs.yale.edu/data/SSDA/CSPP/SPP418.pdf>
Producer: Centre for the Study of Public Policy
Date produced: 1994
Geographic coverage: Bulgaria, Czech Republic, Slovakia, Hungary, Poland, Romania, Croatia, Slovenia, Belarus, Ukraine
Place of production: Aberdeen, Scotland

<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



Source: <http://www.ddialliance.org/>

```
<xs:element name="Geography" type="GeographyType"/>  
<xs:element name="StartDate" type="BaseDateType"/>  
<xs:element name="EndDate" type="BaseDateType"/>  
<xs:element name="DataCollectionFrequency" type="DataCollectionFrequencyType"/>  
1 <xs:element name="SamplingProcedure" type="r:IdentifiedStructuredStringType"/>
```

Increase the
REACH
of Your Data

The Data Documentation
Initiative
www.ddialliance.org

<ddi>

The graphic features a glowing globe with a hand reaching out to touch it, set against a background of data lines and a gear. The text is in a serif font, with 'REACH' in large, bold letters.

*Challenges of finding data 1: institutional catalogs **may** contain pointers to data, but are focused on other types of content*

Example of a library catalog



Cornell University
Library

CATALOG

Not geared towards data

Home to search Articles, Databases, e-Journals, Images

NEW SEARCH

PATRON INFO

REQUESTS

PREFERENCES

SAVED SEARCHES

BOOKBAG

INTERLIBRARY LOAN

Database Name: Cornell University Library

Basic Search

Guided Keyword Search

Search for:

Quick Limit:

None
English
Networked Resources
Serials

Search by:

Title
Journal Title
Journal Title Abbreviation
Author
Subject Heading
Call Number
Author--Sorted by Title
Relevance Keyword
Command Keyword

Typical search for books or journal articles targets author, title, subject, publ. date or issue (depending on topical or known-item searching)

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”

CISER Data Archive: Online Catalog

PRB INFORM EMPOWER ADVANCE **Population Reference Bureau**

IPUMS

Data.gov Catalog

Use the Data.gov catalog to search for datasets. Click on the name of a dataset to view additional metadata for that dataset. If you agree to the **Data Policy**, Data.gov offers data in three ways: through the “raw” data catalog, use the geodata catalog. The “Raw” Data Catalog provides an instant download of machine readable, platform-independent datasets while the Tools Catalog provides hyperlinks which may lead to agency tools or agency web pages that allow you to use datasets.

“RAW” DATA CATALOG **TOOL CATALOG** **GEODATA CATALOG**

Search “raw” data by keywords

Search “raw” data by file type
XML CSV/Text KML/KMZ Shapefile RDF Other

Search “raw” data by single/multiple category

- All Categories
- Agriculture
- Arts, Recreation, and Travel
- Banking, Finance, and Insurance

Search “raw” data by single/multiple agency

- Antitrust Division (DOJ/ATR)
- Broadcasting Board of Governors (BBG)
- Bureau of Economic Analysis (DOC/BEA)
- Bureau of Indian Education (DOI/BIE)
- Bureau of Industry and Security (DOC/BIS)

SEARCH

Google Search

not case sensitive when you
tton. Do **not** include leading

readings

Related

Different search inputs, different search outputs, no easy way to search all at once, and not in “data-targeting” ways

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*

Consumer Expenditure Survey, 2004: Diary Survey

Bibliographic Information:

US. Dept. of Labor, Bureau of Labor Statistics. -- ICPSR FastTrack version -- Washington, DC: U.S. Dept. of Labor, Bureau of Labor Statistics, 2005 [producer]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, 2006 [distributor]. Note: This is the ICPSR FastTrack version, which has received minimal processing. Codebook: ECON-051D(2004).

File Information:

For All Parts					
Type of File	Directory \ File Name	Records	LRECL	RECFM	Size / Size Zipped
Questionnaire	U:\ArchiveData\econ051d\survey04.pdf	n/a	8,101	V	491 KB / 429 KB
Codebook	U:\ArchiveData\econ051d\drvdoc04.pdf	n/a	3,874	V	528 KB / 465 KB
Documentation	U:\ArchiveD			V	289 KB / 274 KB
Universal Classification Codes					
Type of File	Directory \ F			RECFM	Size / Size Zipped
Codelist	U:\ArchiveD			V	18 KB / 7 KB
Annual Income					
Type of File	Directory \ F			RECFM	Size / Size Zipped
Data	U:\ArchiveD			C	2 MB / 237 KB

Better Search & Discovery

```

<r:Citation>
  <r:Title>Consumer Expenditure Survey, 2004: Diary Survey</r:Title>
  <r:Creator>U.S. Dept. of Labor, Bureau of Labor Statistics</r:Creator>
  <r:Publisher>U.S. Dept. of Labor, Bureau of Labor Statistics, 2005</r:Publisher>
  <r:PublicationDate>2006-02</r:PublicationDate>
</r:Citation>
<s:Abstract id="">
  
```

Meta data

Google directory

Search only in Data Archives Search the Web

Wolfram

Enter what you want to calculate

SCIRU for scientific information

Advanced search | P

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

Title:	Longitudinal studies on the causes of obesity: The National Longitudinal Study of Adolescent Health
Author(s):	Gordon-Larsen, P.
Conference/Meeting Name:	Cornell University College of Human Ecology
Conference/Meeting Date:	2005
Conference/Meeting Sponsor:	Cornell University College of Human Ecology
Place of Conference/Meeting:	Ithaca, NY
Related Studies	
This publication is related to the following ICPSR dataset(s):	
• National Longitudinal Study of Adolescent Health (Add Health), 1994-2002 (ICPSR 21600)	

From ICPSR's Bibliography of Data-Related Literature
(<http://www.icpsr.umich.edu/icpsrweb/ICPSR/citations/>)

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](#)

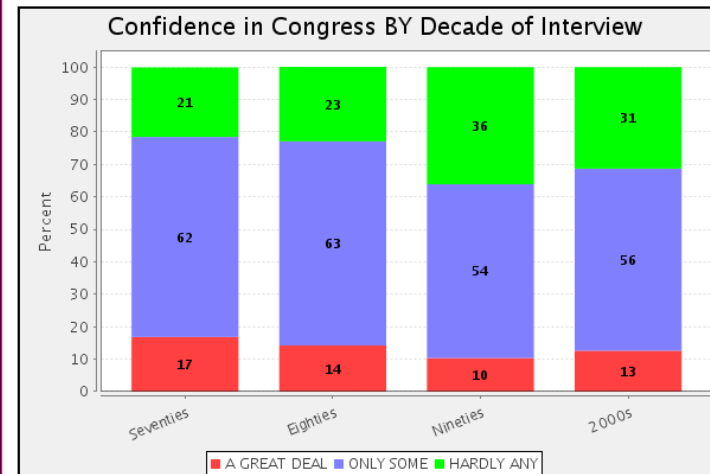


Quick Tables

Quick Table: GSS 1972-2008 Cumulative Datafile

Results: Confidence in Congress *BY* Decade of Interview (Percents)

	Seventies	Eighties	Nineties	2000s	TOTAL
A GREAT DEAL	16.8	14.2	10.3	12.6	13.6
ONLY SOME	61.7	62.9	53.6	56.1	59.0
HARDLY ANY	21.4	23.0	36.1	31.3	27.4
Total Percent	100.0	100.0	100.0	100.0	100.0
(Weighted N)	(8,770)	(10,893)	(8,507)	(6,817)	(34,988)
(Unweighted N)	(8,751)	(10,858)	(8,529)	(6,834)	(34,972)



10

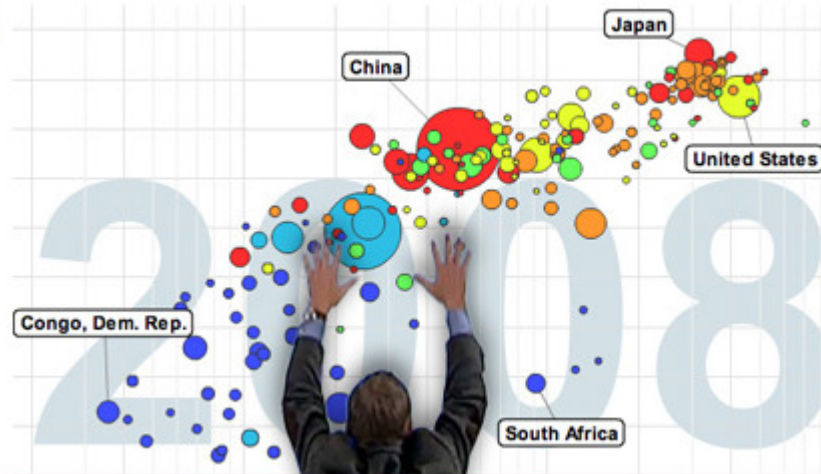
Making research data available for web-based *visualization*

- Could universities load locally created social science research data into a system like ... [Gapminder](#)?

Explore the world

Gapminder World shows the world's most important trends.

- › Wealth & Health of Nations
- › CO₂ emissions since 1820
- › Africa is not a country!
- › Is child mortality falling?
- › Where is HIV decreasing?



Existing infrastructure at Cornell, aligned with the data lifecycle

- Planning and collection
 - Survey Research Institute
 - fee-for-service survey unit at Cornell
 - study design, data collection
 - Increasing partnership with CISER
 - Cornell National Social Survey

Existing infrastructure at Cornell, aligned with the data lifecycle

- Planning and collection (cont.)
 - Colectica Designer
 - Commercial product
 - survey design
 - study documentation
 - data collection
 - statistical command file generation
 - built on DDI 3

Existing infrastructure at Cornell, aligned with the data lifecycle

- Planning and collection (cont.)
 - Research Data Management Services Group
 - New development/collaboration
 - Potentially very significant
 - Return to in a minute

Existing infrastructure at Cornell, aligned with the data lifecycle

- **Analysis and Collaboration**

- CISER

- Research Computing
- Restricted Data Environment
 - CRADC
 - Census RDC
- Help Desk/Workshops
- Close collaboration with Cornell Statistical Consulting Unit

Existing infrastructure at Cornell, aligned with the data lifecycle

- **Analysis and Collaboration**

- DataStaR

- Data staging repository
- Collaboration, data sharing during analysis stage
- Publishes data and metadata elsewhere
- Project of Mann Library at Cornell

Existing infrastructure at Cornell, aligned with the data lifecycle

- **Archiving and Discovery**

- CISER Data Archive
- eCommons@Cornell
 - CU's institutional repository
 - DSpace, text documents, limited capacity for data
- ICPSR

Relationship building

- **Research Data Management Services Group (RDMSG)**
 - Grows out of pending NSF requirement regarding data management plans
 - Partners include: CISER, CUL, DISCOVER Research Services Group, Astronomy, CAC
 - “Meeting Funders’ Data Policies: Blueprint for a RDMSG”

Relationship building

- **Cornell Population Program**
 - “Young” Population Program (formed 2008)
 - Looking for “Signature” Data product
 - Cornell National Equivalencies File (CNEF)

Relationship building

- **Cornell National Social Survey (CNSS)**
 - annual Cornell National Social Survey is conducted by the Survey Research Institute (SRI). It polls adults aged 18 and over on a wide range of current public policy and socioeconomic topics
 - **CISER creates integrated data, web extraction system**
 - **Value increases over time**
 - **Connection to Cornell Researchers**

Relationship building

- **Relationships with Individual Faculty/Researchers**
 - Mildred Warner
 - Concerned about at-risk data
 - Connections to data sources/urgent timing

CISER Staffing Initiatives

- **Building on the CISER Data Archive (in existence since 1981):**
- **CISER Research Data Management Librarian (Stefan Kramer)**
- **CISER Research Associate**
 - **Good research/data skills**
 - **Available for hire to Cornell research projects**
- **Cornell Programming Staff**
 - **Front end/backend programming**

Thank you for your time & attention!



William C. Block
block@cornell.edu

Stefan Kramer
stefan.kramer@cornell.edu