1. DGI-Konferenz, 62. DGI Jahrestagung

Semantic Web & Linked Data

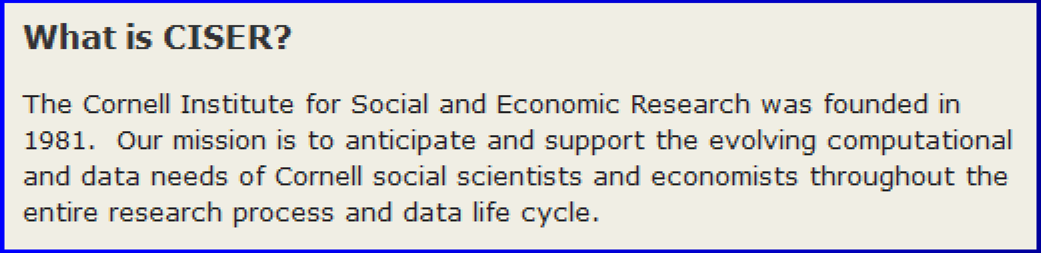
Elemente zukünftiger Informationsinfrastrukturen

Challenges and Opportunities in Social Science Research Data Management

Stefan Kramer
Research Data Management Librarian
Cornell Institute for Social and Economic Research

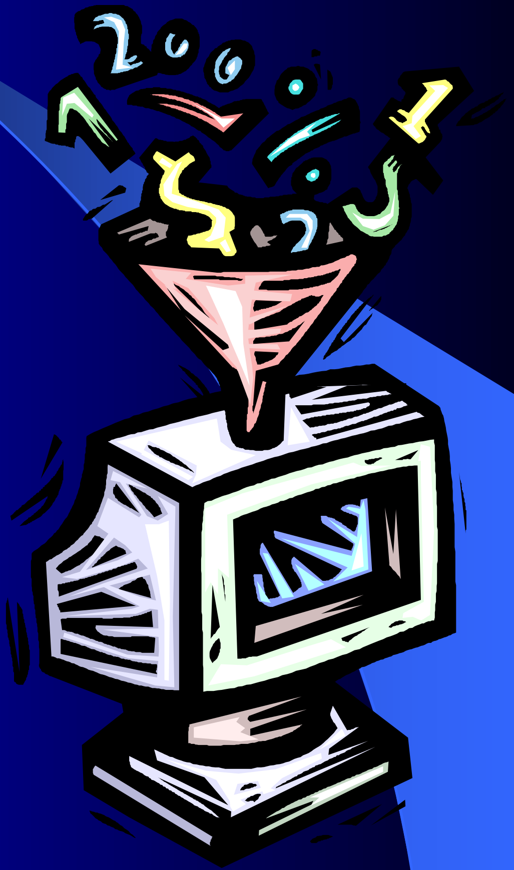
What is CISER?

Cornell University
Cornell Institute for Social and Economic Research



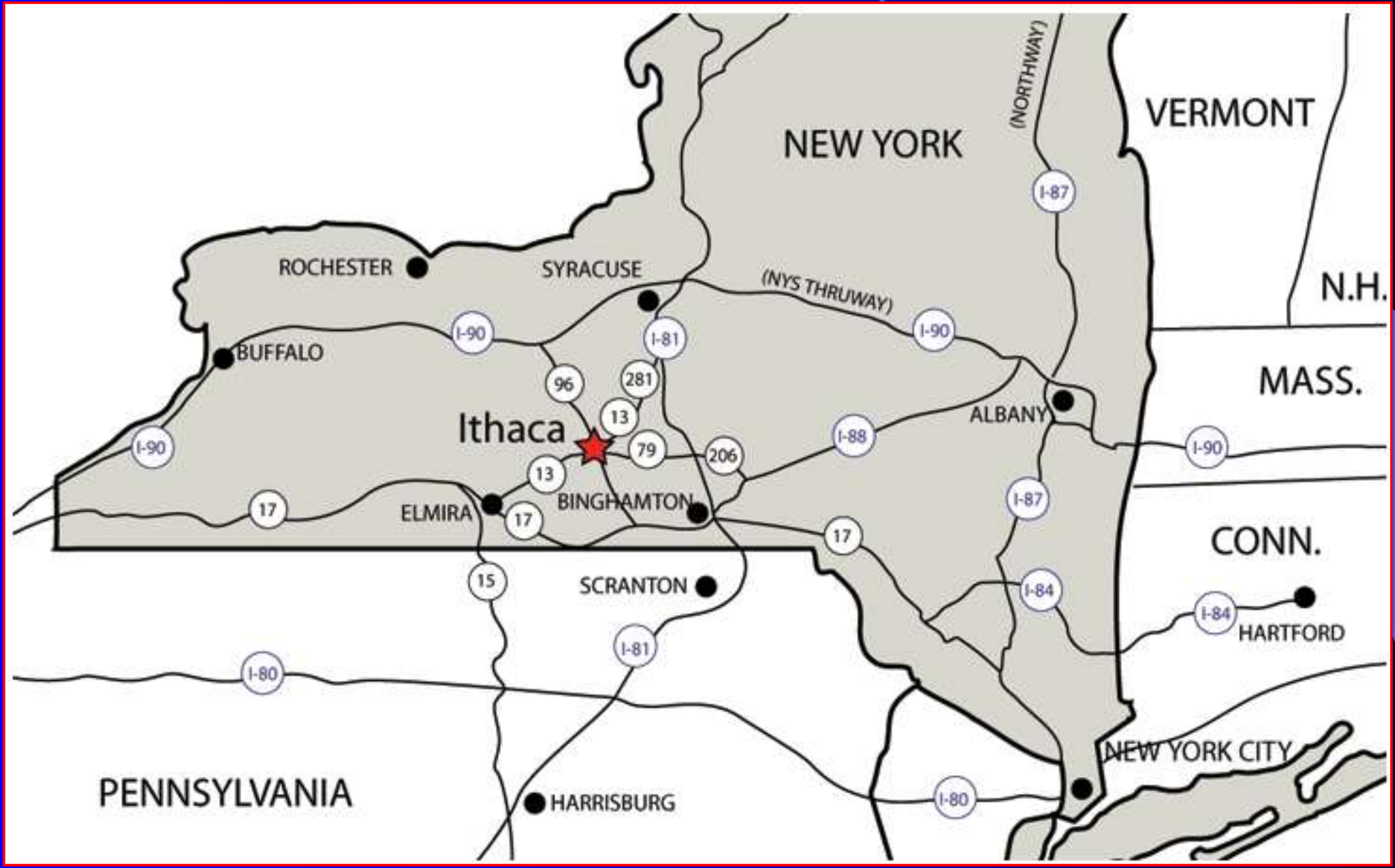
What is CISER?

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/About_CISER.shtml

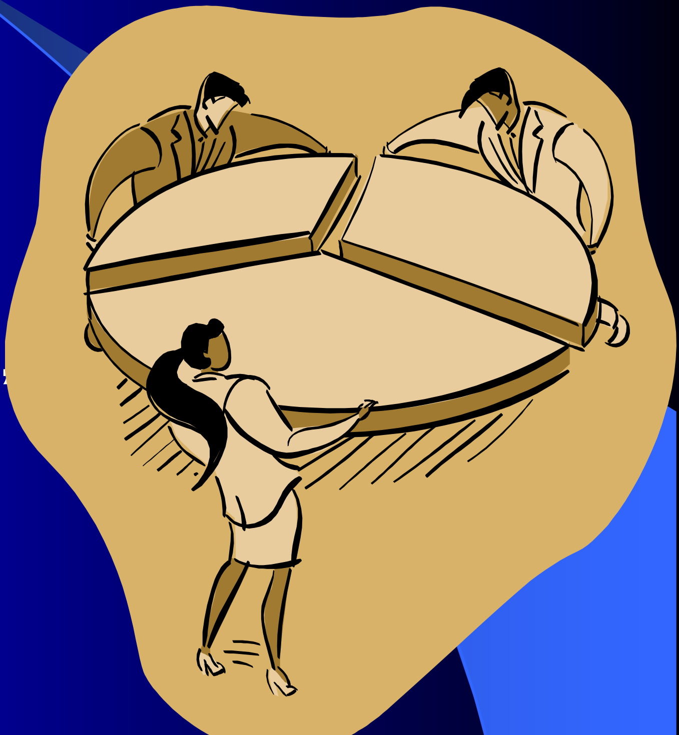
Where is Cornell University?

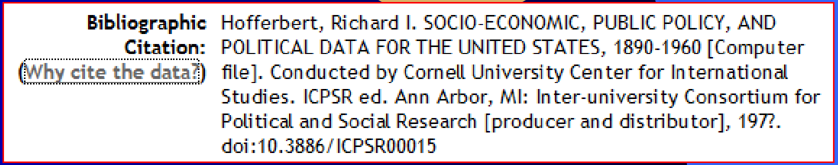


Source: http://www.cornell.edu/maps/state.cfm

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





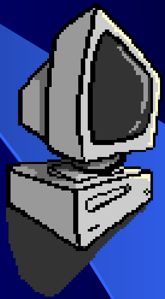
Bibliographic Citation: (Why cite the data?) Hofferbert, Richard I. SOCIO-ECONOMIC, PUBLIC POLICY, AND POLITICAL DATA FOR THE UNITED STATES, 1890-1960 [Computer file]. Conducted by Cornell University Center for International Studies. ICPSR ed. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [producer and distributor], 197?. doi:10.3886/ICPSR00015

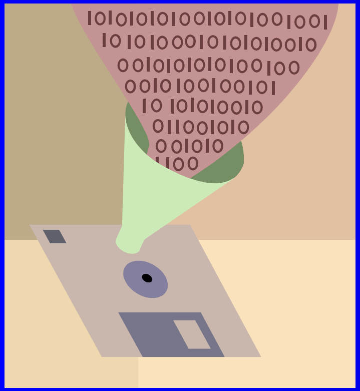
Some potential problems with ***own*** data (that’s not (well) managed) for researchers

Where is it?

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

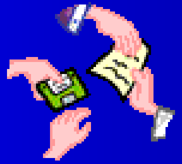
Can my computer and software still read/open/use it?

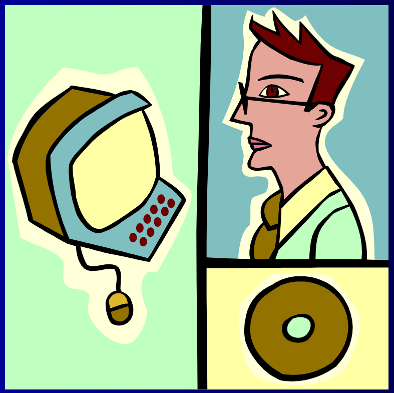




Hmm, what format were those files in again?

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

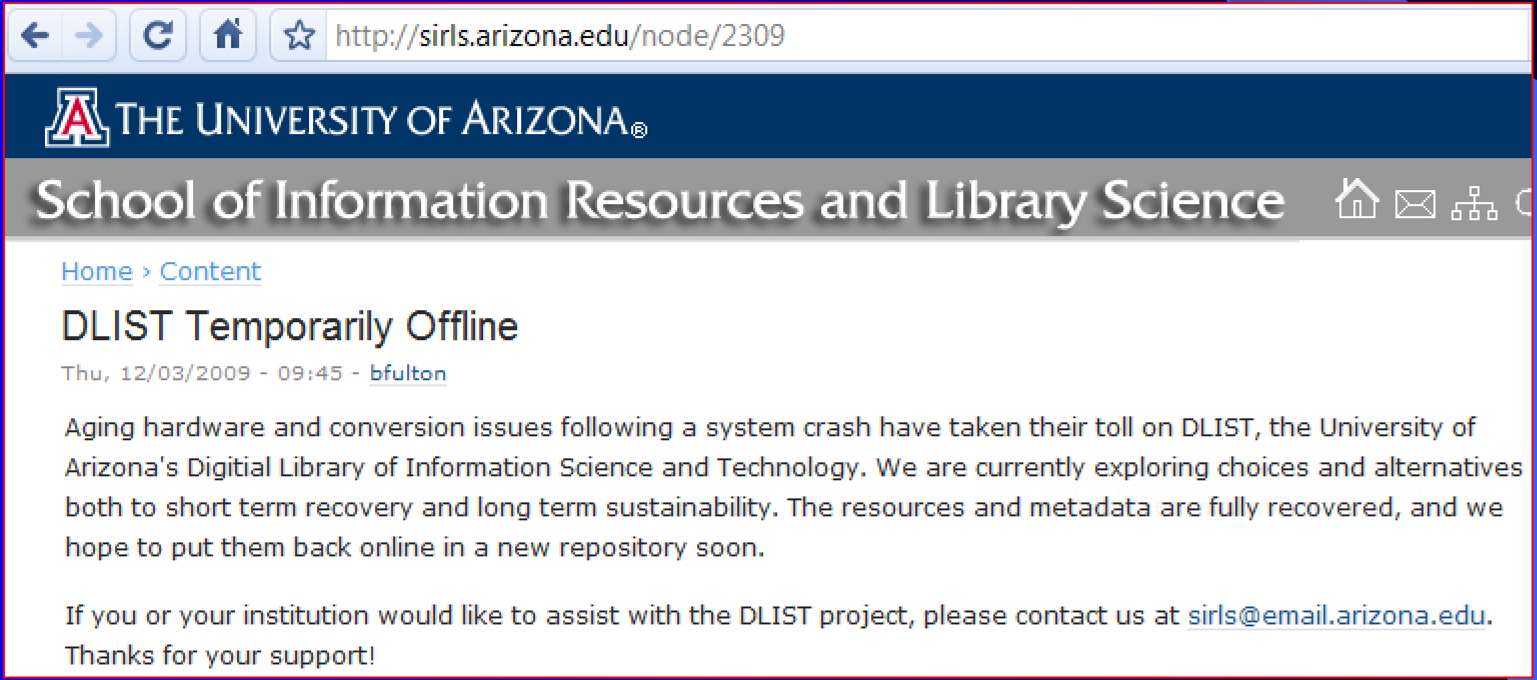




Where is the graduate assistant who organized, analyzed, … the data now?

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



http://sirls.arizona.edu/node/2309

THE UNIVERSITY OF ARIZONA®

School of Information Resources and Library Science

Home › Content

DLIST Temporarily Offline

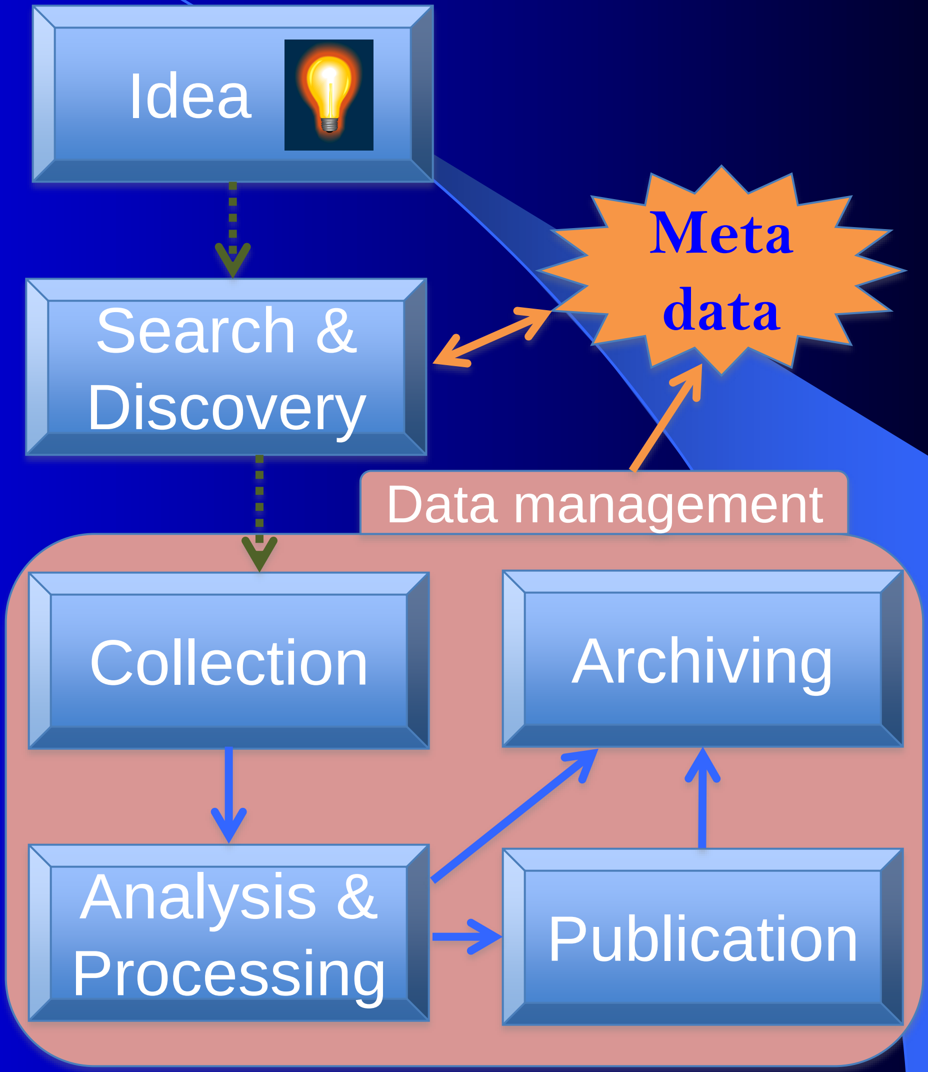
Thu, 12/03/2009 - 09:45 - bfulton

Aging hardware and conversion issues following a system crash have taken their toll on DLIST, the University of Arizona's Digitial 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.

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!

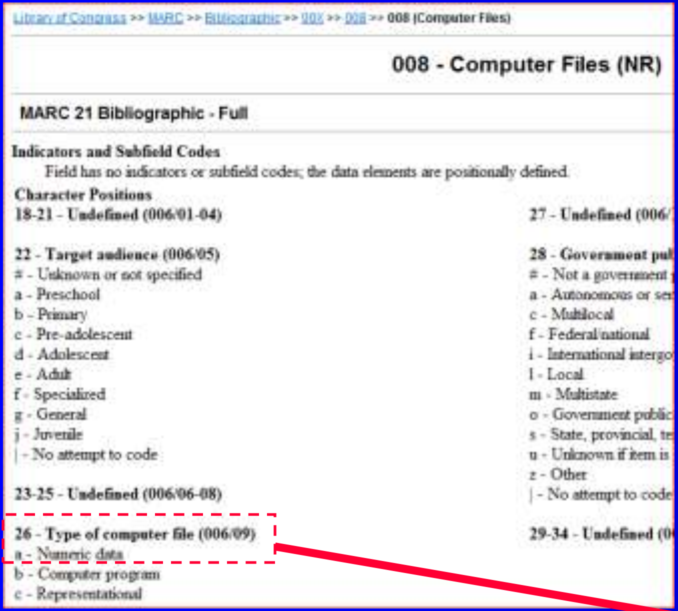
The lifecycle of research data

- 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



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)

22 - Target audience (006/05)
- Unknown or not specified
a - Preschool
b - Primary
c - Pre-adolescent
d - Adolescent
e - Adult
f - Specialized
g - General
j - Juvenile
| - No attempt to code

23-25 - Undefined (006/06-08)

26 - Type of computer file (006/09)
a - Numeric data
b - Computer program
c - Representational

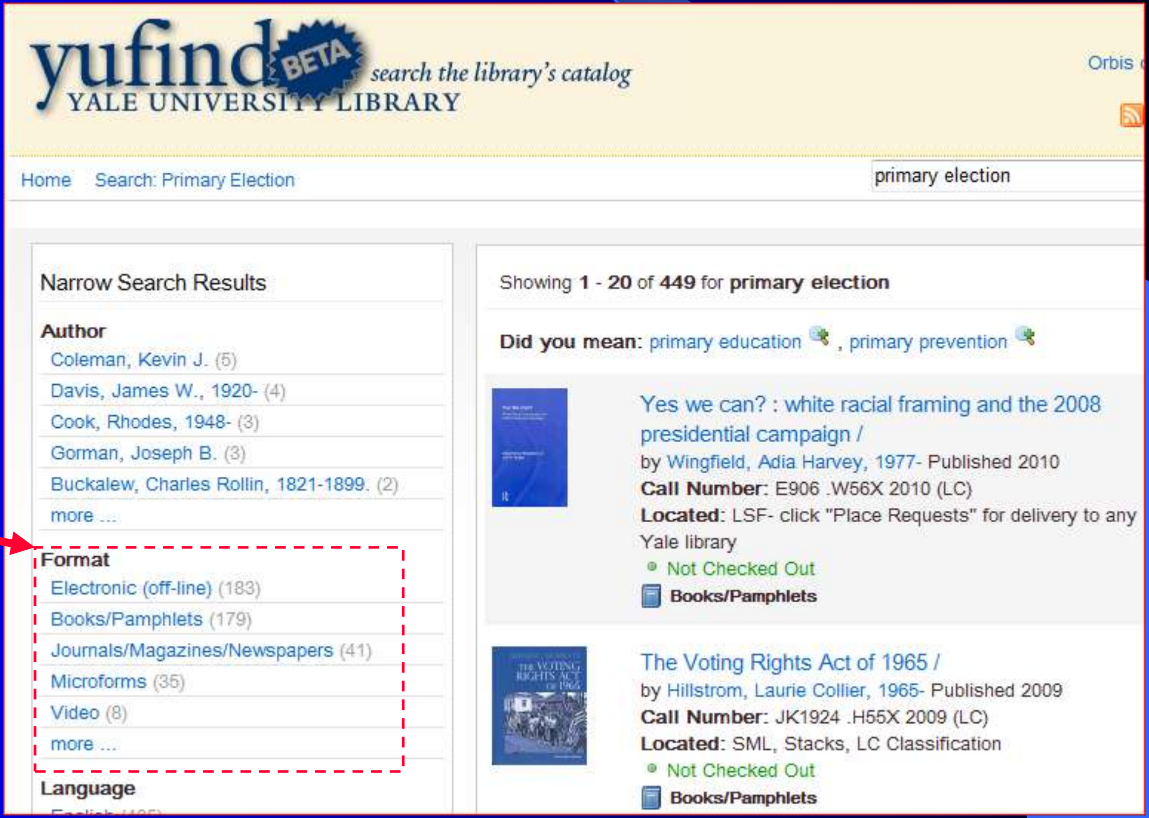
27 - Undefined (006/

28 - Government pu
- Not a government
a - Autonomous or se
c - Multilocal
f - Federal/national
i - International intergo
l - Local
m - Multistate
o - Government public
s - State, provincial, te
u - Unknown if item is
z - Other
| - No attempt to code

29-34 - Undefined (0

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

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



yufind BETA search the library's catalog
YALE UNIVERSITY LIBRARY

Orbis

Home Search: Primary Election

primary election

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

Showing **1** - **20** of **449** for **primary election**

Did you mean: primary education, primary prevention

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

Search/browse functions for numeric data in social sciences

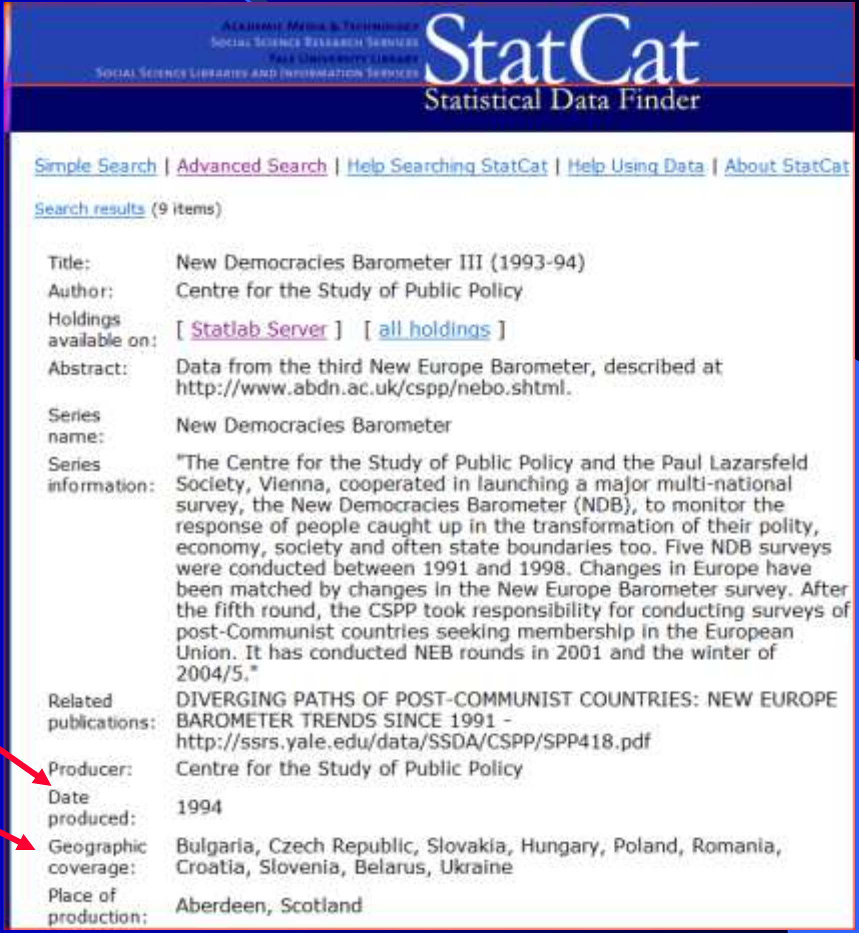
Not (easily) offered by most search systems, incl. library catalogs, but often needed by data searchers, in addition to topic:

Time span (example: 1970-present)

Time frequency (example: annually)

Geographic extent (example: all of United States)

Geographic granularity (example: county level)


Academic Media & Technology
Social Science Research Services
Yale University Library
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

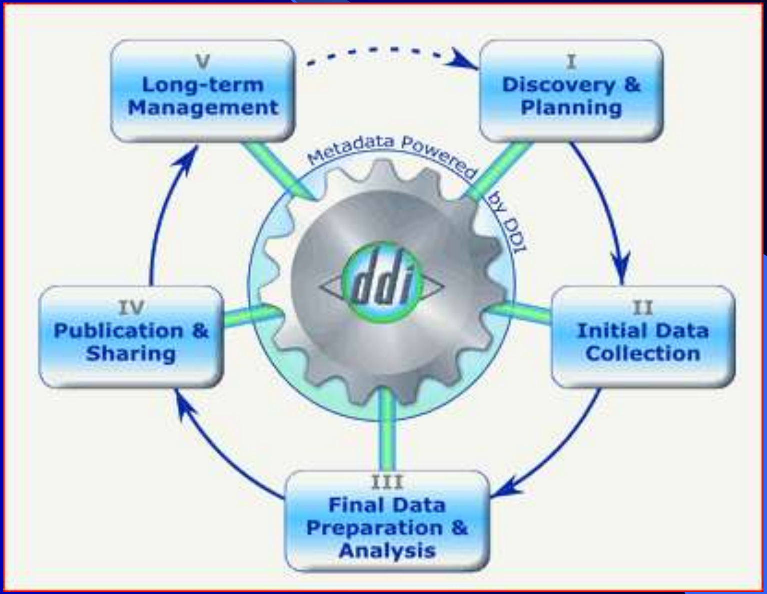
Researchers and metadata creation/maintenance

- 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

Increase the REACH of Your Data
The Data Documentation Initiative
www.ddialliance.org
<ddi>

Data Documentation Initiative (DDI)

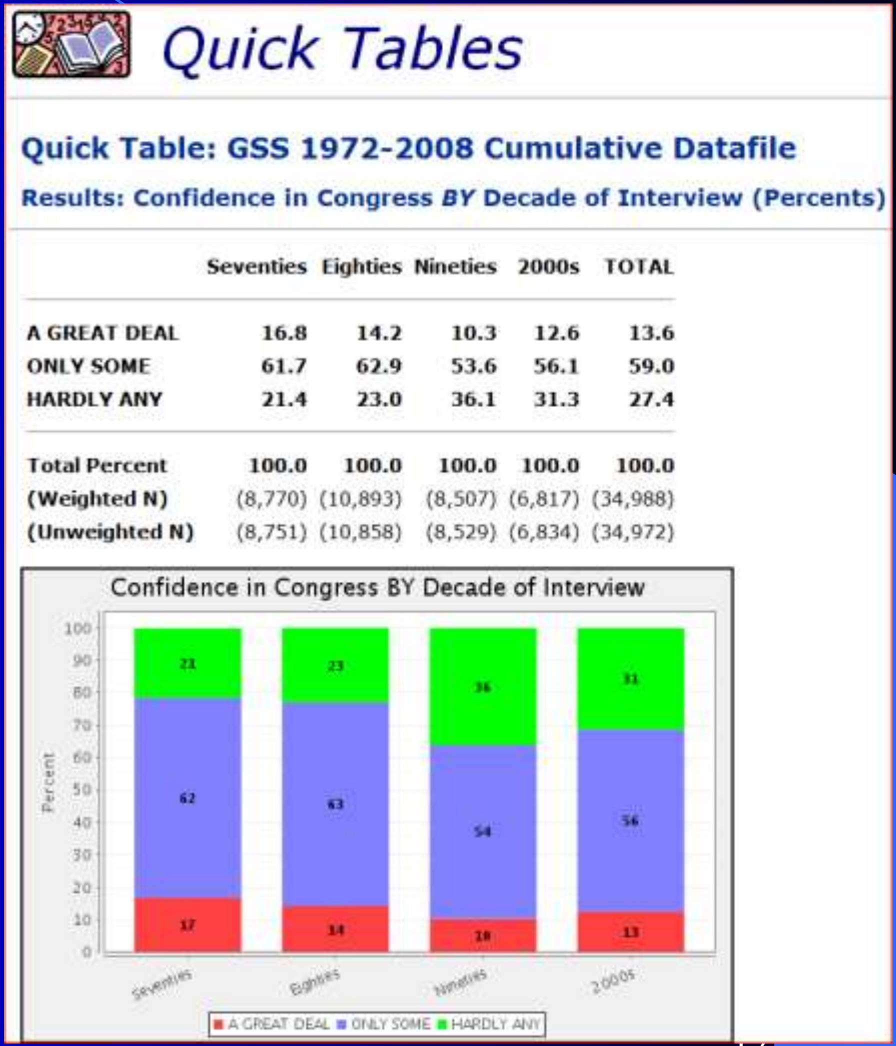
- Earlier versions of DDI focused on *codebooks* – the "manual" for datasets
- DDI 3 designed to support the data lifecycle with metadata
- Powerful – but also complex! Used by national statistical agencies, data archives, etc.
- Tools for using DDI being developed – choosing the right ones for specific 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



Quick Tables

Quick Table: GSS 1972-2008 Cumulative Datafile

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

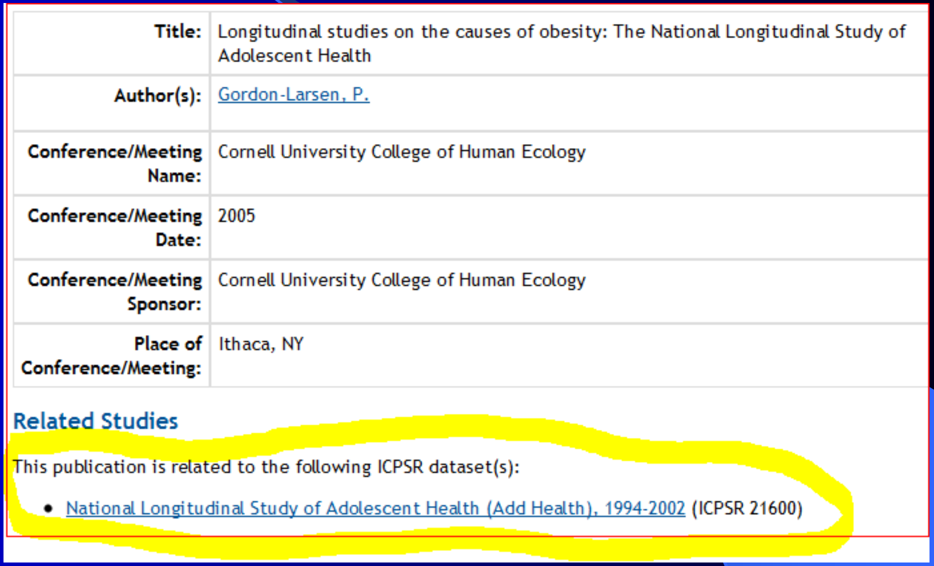
	Seventies	Eighties	Nineties	2000s	TOTAL
A GREAT DEAL	16.8	14.2	10.3	12.6	13.6
ONLY SOME	61.7	62.9	53.6	56.1	59.0
HARDLY ANY	21.4	23.0	36.1	31.3	27.4
Total Percent	100.0	100.0	100.0	100.0	100.0
(Weighted N)	(8,770)	(10,893)	(8,507)	(6,817)	(34,988)
(Unweighted N)	(8,751)	(10,858)	(8,529)	(6,834)	(34,972)

Confidence in Congress BY Decade of Interview

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*



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)

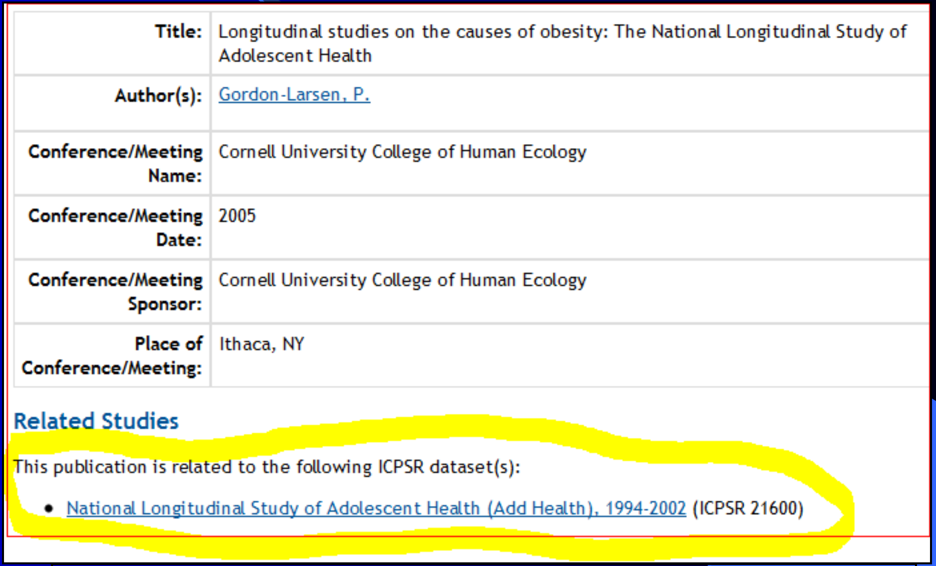
From ICPSR’s Bibliography of Data-Related Literature (http://www.icpsr.umich.edu/icpsrweb/ICPSR/citations/)

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



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/)

Thank you for your time & attention!

Stefan Kramer

stefan.kramer@cornell.edu

http://www.linkedin.com/in/kramerstefan