



New Developments in Data Sharing, Remote Access, Secure Data, and Documentation at the Cornell Institute for Social and Economic Research (CISER)

William C. Block and Lars Vilhuber

4th Workshop on Data Access

26 March 2012

Luxembourg



Outline of Today's Presentation

- CISER Secure Data Services
 - Cornell's NSF-Census Research Network (NCRN) Project
 - Cornell's Research Data Management Service Group (RDMSG)
 - Improvements to CISER Data Archive
 - Latest Developments at CISER and Cornell
- 
- A decorative graphic consisting of several concentric circles in a light blue color, located on the right side of the slide.

CISER

Before we begin...what is CISER?



CISER's Mission:

....anticipate and support the evolving computational and data needs of Cornell social scientists and economists throughout the entire research process and data life cycle.

CISER

CISER Broad Range of Services include:

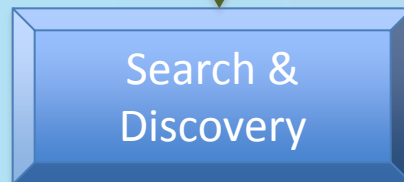
- Data Programming Services
- Consulting and Workshops
- Data Archive
- Data Management Plans
- Restricted Data Services
- Research Computing
- Administrative Support



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



For some projects, existing data sources may be sought and explored

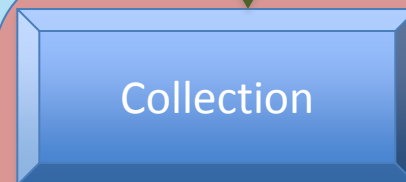


By search tools utilizing metadata from data stores, **new research data** becomes available for finding and exploring by researchers

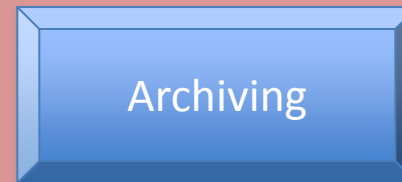
Metadata

Ideally begins early in data lifecycle to assure long-term preservation and access of data.

Measurement instruments are designed, developed, acquired; data are collected through appropriate methodologies

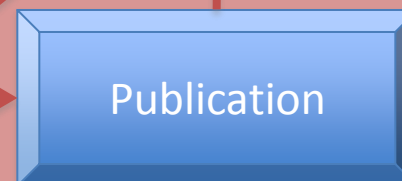
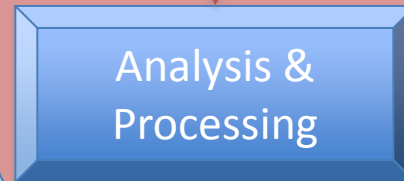


Data management

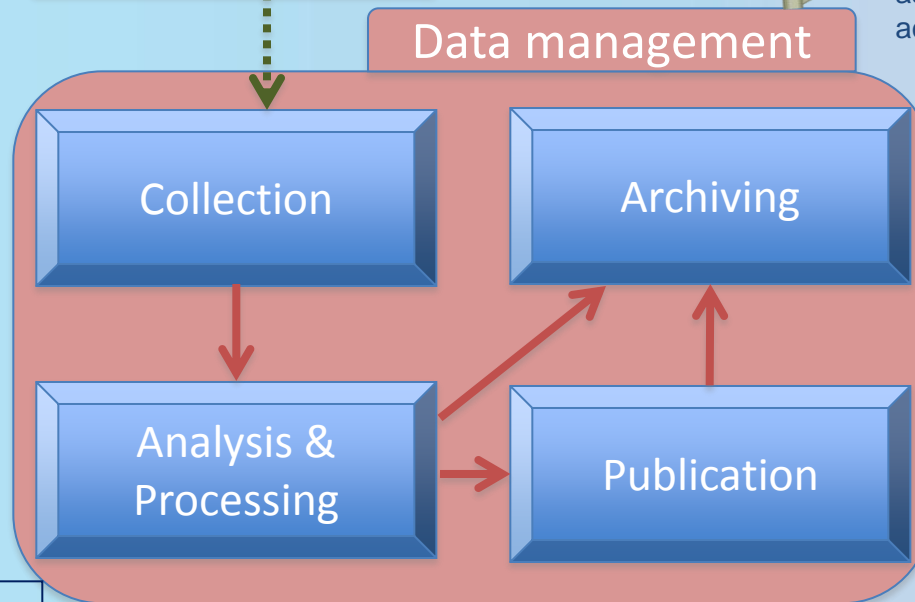


Final datasets are deposited for long-term preservation – e.g., into institutional or domain repository

Collected data are merged, cleaned, analyzed, subsetted, 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 Archive - *Collection and Services*

- Established 30 years ago
- Collection of datasets to support quantitative research
- Consulting services to match user needs with appropriate data
- Provides Cornell social science researchers a repository for providing others access to, and long-term preservation of, their numeric/statistical research data

CISER Research Computing

- DELL R810 servers, Diskless boot using DELL's Advanced Infrastructure Management (AIM)
- Windows 2008 R2
- 40-core (80-core when hyper threaded)
- 512Gb RAM
- 10Gb internal network utilizing Force10 equipment with failover capability
- 96 Tb of raw disk on CISERRSCH and CRADC domains/each
- 144 Tb of raw backup disk for CRADC domain (disk to disk backup)
- Ez-backup service for CISERRSCH domain (daily backup)
- Currently service ~2,200 accounts
- Virtual servers available upon request for specialized projects
- Secured by CISCO ASA 5520 (CRADC) and 5540 (CISERRSCH)

CISER Secure Data Services

- Newly-named umbrella term for “a comprehensive suite of services for researchers using restricted data.”
- Resolve confusion between CRADC and CRDC
- CISER is Cornell’s official custodian of restricted access data sets
- Customized environment to meet data security standards from individual data providers
- Support across entire Project Lifecycle

CISER Secure Data Services—Multiple Modes of Access

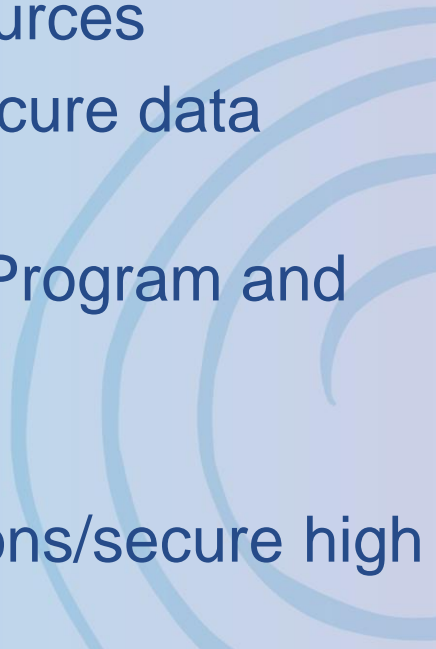
- Secure Remote Access (to CISER servers from your computer*)
- Dedicated Stand-alone computers/Secure Rooms at CISER
- Access to remote-servers from secure rooms at CISER (pending IAB agreement)
- Cornell Census Research Data Center (U.S. Government statistical data)
- Cornell Virtual RDC

*Most commonly understood CRADC feature

CISER Secure Data Services—Examples of Secure Data and Providers

- U.S. Census Bureau
- Agency for Healthcare Research and Quality (AHRQ)
- Adolescent Health (Add Health)
- New York State Department of Health
- Health and Retirement Study (HRS)
- German Institute for Employment Research (IAB)
- Many other government agencies, research organizations, private companies...

CISER Secure Data Services—Support Across the Data Lifecycle

- Identification of appropriate data resources
 - Writing security plans/working with secure data providers
 - Coordinate with Office of Sponsored Program and Institutional Review Board
 - Data Management Plans
 - Complete range of software applications/secure high capacity computing
- 
- A decorative graphic consisting of several concentric circles in a light blue color, located on the right side of the slide. The circles are centered around a point towards the bottom right and expand outwards, creating a spiral-like effect.

Cornell's Node of the NSF-Census Research Network (NCRN)

- Cornell NSF-Census Research Node: Integrated Research, Support, Training, and Data Documentation
- 1 of 8 Nodes funded in the Network (\$1.2M – \$3M each)
- Investigators: John Abowd, William Block, Lars Vilhuber, and Ping Li
- The Comprehensive Census Bureau Metadata Repository (CCBMR)
- Socio-economic / official statistics often have need for confidentiality restrictions/privacy.

The Death Knell for Public-use Data

- Sounded by young scholars pursuing research programs that mandate inherently identifiable data: geospatial relations, exact genome data, networks of all sorts, linking administrative records.
- These researchers acquire authorized restricted access to the confidential identifiable data and perform their analyses in secure environments.
- But they don't leave behind the scientific trail that has made public-use files so important.

The Comprehensive Census Bureau Metadata Repository (CCBMR)

- Facilitates access to detailed metadata
 - Restricted-access data from outside an RDC while enabling fine-grained control over confidential information for the (Longitudinal Business Database (LBD), American Community Survey (ACS), American Housing Survey (AHS), Longitudinal Employer-Household Dynamics (LEHD))
 - Public-use datasets inside restricted-access areas (IPUMS, CPS)
- Expands the notion of metadata to include user-generated components (notes, programs, *etc.*)

NCRN Research Program

- Develop and implement state-of-the-art statistical learning algorithms for the U2W imputation task

Explore boosting, logistic regression, support vector machines, and conditional random fields

Boosting is potentially promising

- In addition to the distance attribute, explore other available data such as the demographic information (gender, occupation, race, education, *etc.*)
- Use statistical learning to improve the integration of the establishment demography

NCRN Training Component

- CCBMR
- VirtualRDC and Synthetic Data Server
- Instructional material converted to online, self-paced course (see INFO 7470)
- All tools developed in the NCRN nodes accessible (in cooperation with all other nodes)

CISER

Cornell's Research Data Management Service Group (RDMSG)

- “comprehensive data management planning and services”
- CISER is a founding member
- Present university-wide coherent set of services available to researchers
- Unified web presence for specialized services
- Source of standard language/guidance on data management plans

Cornell RDMSG Service Areas

- Data Management Planning Overview
- Data Storage and Backup
- Metadata
- Data Analysis
- Collaboration Tools
- High performance computing
- Privacy and confidentiality
- Intellectual property and copyright
- Data publication
- Data Archiving and Curation



CISER

- Some recent enhancements made to the CISER Data Archive

Search Archive Holdings

1) Choose which **field(s)** you want to search:

☒ Title
☒ **Abstract ***
☐ Producer
☐ Principal Investigator
☐ ICPSR Number
☐ CISER Codebook Number

* Abstract note: Our abstracts are prepared by the ICPSR staff. (47)

2) Retrieve listings for files on:

☒ CISER Research Computing System

3) Search for:

4) Search type: ☒ all of the above

Search Results

Search: TITLE, ABSTRACT = election

[go back](#) , [create a new search](#), or use this form to **search within your 200 results:**


Limit search results to studies that were ☐ created / ☐ enhanced by Cornell

Search for files on: ☒ CISER Research Computing System ☒ CDROM / DVD

Studies on the CISER Research Computing System: [Studies on CDROM / DVD \(below\)](#)

198 results found.

ABC News General Election Exit Poll, November 1988

 ABC News. Radnor, PA: Chilton Research Services, 1988 [producer]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. Codebook: POL-120 .

File access:

- CUWebLogin authorization

Click on traffic light for more information.

with voters participating in the November 8, 1988 general election. Voters at the time were asked to complete questionnaires as they exited the polling places. All respondents were asked to provide information on their age, sex, race, education, occupation, and United States representative... [more](#)

Future Plans for the CISER Data Archive

- Assignment of **persistent identifiers** (e.g., DOIs) to datasets to make them more easily citable in the scholarly literature, and reliably linkable
- **Web-based ingest function** for CU researchers to submit datasets and metadata for archiving and access
- **Variable-level search functionality** to allow end users to find datasets that contain specific variables of interest to their research
- Export and exposure of standards (DDI, Dublin Core, ...) conformant metadata

Latest Developments at Cornell and CISER

- Labor Dynamics Institute Founded in the ILR School

Mission: To create and make accessible novel data on the dynamics of the labor markets, we work with research networks and statistical agencies, developing appropriate statistics to inform policy makers, researchers, and simply people seeking knowledge. We emphasize and meet the requirements of stakeholders: users as well as providers, balancing the utility of the data with the confidentiality of the people and businesses whose activities the data describe.

- National Registry of Catastrophic Youth Sports Injuries (Pilot Project Proposed at CISER)

Latest Developments at Cornell and CISER (Continued)

- Improvements planned to the Cornell Virtual RDC

- Planned upgrade of internode network (10x faster)
- Easier transfers to XSEDE (web-based interface)

- IASSIST 2012 (June, Washington D.C.)
Combined Session on Expanding Data Access and Protecting Confidentiality in a Shrinking World and Secure Data Services:
Unlocking the Power of Restricted Data

Latest Developments at Cornell and CISER (Continued)

- Secure Data Agreement with Institut für Arbeitsmarkt- und Berufsforschung (IAB; Institute for Employment Research)
 - Access to remote-servers from secure rooms at CISER (pending IAB agreement)
 - Secure Remote Access (to CISER servers from your computer; Germany scientific use files)



Thank you for your
time & attention!

Questions?

SecureData@cornell.edu

ciser.cornell.edu

