# UCLA Library Data Archive

*Doing Data since 1961*

Fall Quarter 2017

# Today's Topics

- UCLA Library Data Archives: who we are and what we do
- How to search for and access data for research
- Resources, tools and support available to students at UCLA
- Publishing and Data Management Plans

# UCLA Library Data Archive

**Social Science Data Archive**

**1120 Rolfe Hall**



[https://www.library.ucla.edu/location/social-science-data-archive](https://www.library.ucla.edu/location/social-science-data-archive)
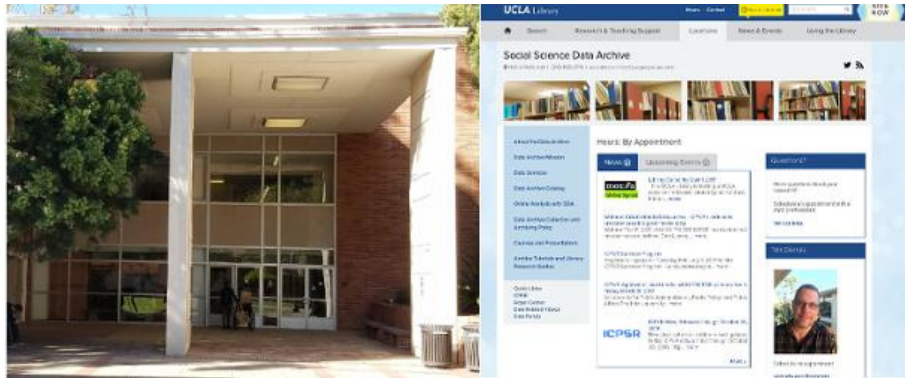
# Data Archives: Our Services and the Data Life Cycle

The Data Archive provides data services that are tied to the data life-cycle.

*"Data archiving is a process, not an end state where data is simply turned over to a repository at the conclusion of a study. Rather, data archiving should begin early in a project and incorporate a schedule for depositing products over the course of a project's life cycle and for the creation and preservation of accurate metadata, ensuring the usability of the research data itself. Such practices would incorporate archiving as part of the research method." Preserving Research Data, Jacobs and Humphrey (2004)*
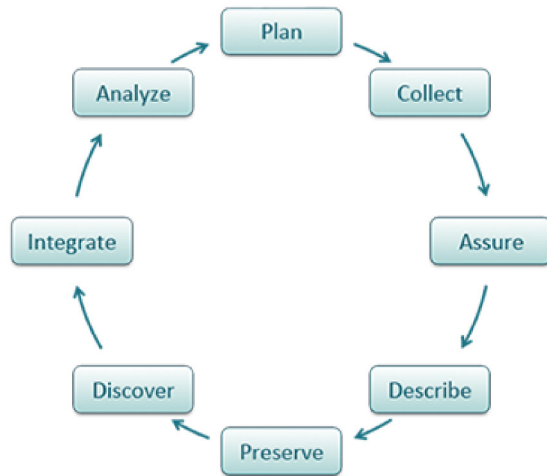
**illustration: DatOne Data Life Cycle**

# Data Archives: Our Services and the Data Life Cycle

https://www.library.ucla.edu/location/social-science-data-archive/data-services

## Plan

- Review data sources
- Help investigate archiving issues, costs, consent and disclosure risks
- create a data management plan

## Collect/Assure

- Project management and file organization
- Quality Assurance for collected data and acquisition workflows
- Consultation on restricted data access and security

## Describe

- Assist with filenames, standard terminology and data dictionaries
- Document analysis and file manipulations
- Help identify appropriate standards

# Data Archives: Our Services and the Data Life Cycle

## Preserve

- Help with backing up your Data
- Help with deciding what data to preserve
- Choosing stable file formats
- Help identifying suitable repositories

## Integration

- Consider compatability
- Document steps
- Capture the provenance of sources

## Analysis

- Help identify appropriate sofware
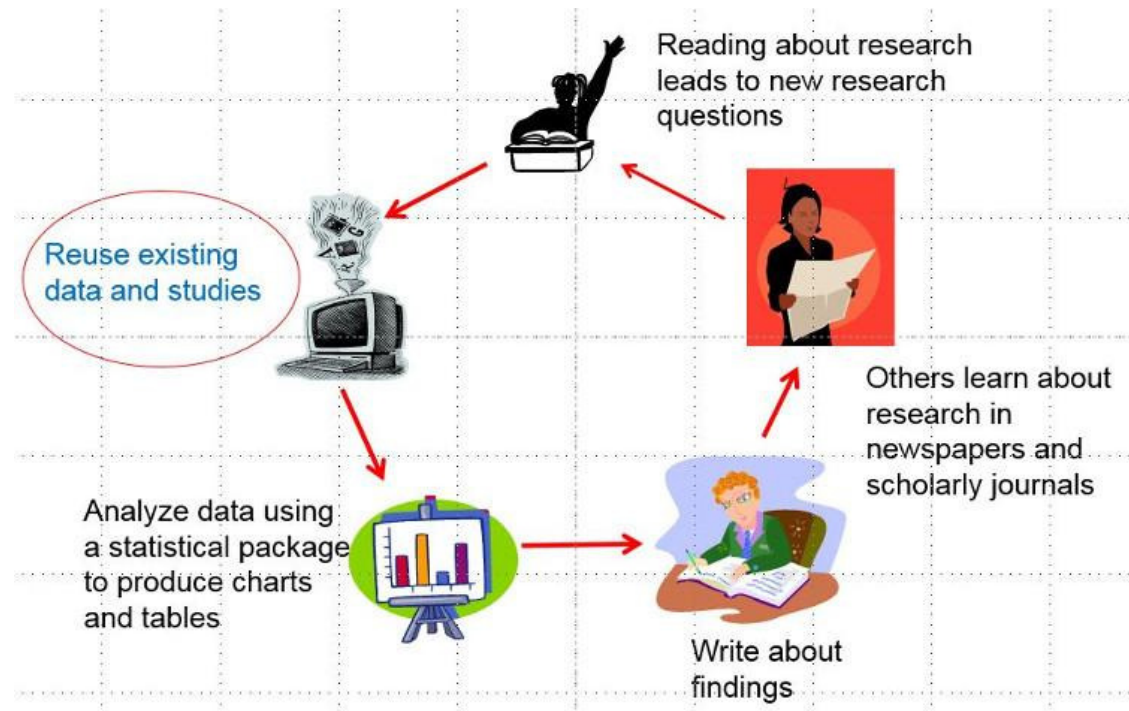- Consultation on coding best practices

# Data Archives: Our Services and Software

- Provide training: Software/Library carpentry / Software Carpentry R-language

- Host events and lectures: Data Archive News and Upcoming Events

- Statistical Consulting: Refer to UCLA Statistics Consulting Center

# Role of Data in Research: Data Reuse



Reading about research leads to new research questions

Reuse existing data and studies

Others learn about research in newspapers and scholarly journals

Analyze data using a statistical package to produce charts and tables

Write about findings

# Finding Out About Data Sources

- UCLA Library Data Archive
- Colleagues
- Public media (newspapers, news)
- Social media (twitter, etc)
- Scholarly literature
- Government reports and documents
- Archives and project websites
- Codebooks, questionnaires
- Exploration of datasets

# Data Sources

- ICPSR: http://www.icpsr.umich.edu/icpsrweb/ICPSR/
- UCLA Library Data Archive Catalog: https://dataverse.harvard.edu/dataverse/ssda_ucla
- UCLA Library DataDen: https://dataden.library.ucla.edu/jspui/
- Government Sites, Census Data: https://www.census.gov/
- IPUMS (Integrated Public Use Microdata Series): https://www.ipums.org/
- Public Policy Institute of CA: http://www.ppic.org/
- NBER (National Bureau of Economic Research): http://www.nber.org/data/
- CDC/NCHS (Centers for Disease Control and Prevention/National Center for Health Statistics): https://www.cdc.gov/nchs/surveys.htm
- Pew Research Center: http://www.pewresearch.org/
- Data.gov: https://www.data.gov/
- World Health Organization: http://www.who.int/healthinfo/statistics/en/
- International Monetary Fund: http://www.imf.org/en/Data

# Search Strategies and Defining Your Research

- General to specific; broad to narrow
- Names of investigators; names of studies
- Studies described in literature, news and websites

| Geography | Physical areas | Government units | Political boundaries | Other – school districts, etc. | |
|---|---|---|---|---|---|
| Time period | Current | | Historical | Over time | |
| Population & Units of analysis | Gender, age, etc. | Race, ethnicity, ancestry, etc. | Persons | Households | Other – schools, hospitals, etc. |
| Context | What is your hypothesis? What will the data tell you? How will you interpret it? | | | | |
| Variables in the dataset | What variables do you need to answer your research question? | | | | |

# How do I know this study useful?

- Raw data are **not** eye-readable
- To decipher data you will use **Questionnaires and Codebooks**

# What is a Questionnaire?

- A **questionnaire** is a *research instrument* consisting of a series of *questions* and other prompts for the purpose of gathering information from respondents.

- Questionnaires cover separate topics such as:

  - Preferences (e.g. political party)
  - Behaviors (e.g. food consumption)
  - Facts (e.g. gender)

- Questionnaires can include indexes or scales:

  - Latent traits (e.g. personality traits)
  - Attitudes (e.g. towards immigration)
  - An index (e.g. Social Economic Status)

# Questionnaire: ABC New Poll

TNS          Job #175374
Horsham, PA          July 23, 2008          ← **Date of poll**
FINAL

ABC News/Discovery Channel/Stanford University Poll #1067: Global Warming Poll

**Title of Poll**

Hello, I'm (NAME), calling for the ABC News public opinion poll. We're not selling anything, just doing an opinion poll on interesting subjects for the news. Out of all the people age 18 or older who are AT HOME RIGHT NOW, can I please speak to the
YOUNGEST (male/female)? (IF NO MALE/FEMALE IN THE HOUSEHOLD, ASK: Then may I please
speak to the YOUNGEST (female/male)?)

**Question text**

1. How would you rate the condition of the natural environment in the world today –
excellent, good, fair, poor, or very poor?

1 Excellent
2 Good          5 Very poor
3 Fair          DK/No opinion
4 Poor          NA/Refused

**Response categories and values**

# What is a Codebook?

- An overview of the study or dataset
- Variable names and descriptions
- Column locations (mostly this is for raw data)

## May also include:

- Test of questions if the data comes from a survey
- Example of the Questionnaire
- Detail on who responded to the survey

# Codebook: General Social Survey

| | | |
|---|---|---|
| **Variable name** | → | **HLTHINFO** R SOUGHT HEALTH INFORMATION PAST YEAR |

**Text of this Question or Item**

792. In the past year, that is, between (INSERT CURRENT MONTH) 1999/2001 and (INSERT CURRENT MONTH) 2000/02 have you looked for information about a health concern or medical problem?

**Question text** →

| % Valid | % All | N | Value | Label |
|---|---|---|---|---|
| 49.8 | 3.5 | 1,628 | 0 | NOT AT ALL |
| 20.6 | 1.5 | 675 | 1 | 1 OR 2 TIMES |
| 17.1 | 1.2 | 559 | 2 | 3-5 TIMES |
| 12.5 | 0.9 | 408 | 3 | 6 OR MORE TIMES |
| | 0.0 | 13 | 8 | DK |
| | 0.1 | 27 | 9 | NA |
| | 92.9 | 43,200 | . | (No Data) |
| 100.0 | 100.0 | 46,510 | | Total |

**Number of responses** ←

**Values and labels** ←

**Properties**

Data type: numeric
Minimum code defined as valid: 0
Maximum code defined as valid: 3
Missing-data codes: -1,8,9
Record/column: 1/3360

# Codebook and Raw Data



| Q | VAR | DECK | COL |
|---|-----|------|-----|
| 2 | 119 | 05 | 7-8 |

NOW, I'D LIKE TO ASK YOU SOME QUESTIONS ABOUT WHERE YOU LIVE.
HOW LONG HAVE YOU LIVED IN THE LOS ANGELES AREA?

00. less than 1 month
01. 1 month to less than 2
etc.
11. 11 months to less than 12
12. less than 1 year—number of months not specified
13. 1 year to less than 2
etc.
96. 96 or more years
*97. refused
*98. DK
*99. NA

| Q | VAR | DECK | COL |
|---|-----|------|-----|
| 3 | 120 | 05 | 9-10 |

HOW LONG HAVE YOU LIVED AT YOUR PRESENT ADDRESS?

00. less than 1 month
01. 1 month to less than 2
etc.
11. 11 months to less than 12
12. less than 1 year—number of months not specified
13. 1 year to less than 2
etc.
96. 96 or more years
*97. refused
*98. DK
*99. NA

Dump of Data -- This is an example of a raw data file.

U C L A   CAMPUS COMPUTING NETWORK

3068                    10 250                    $00002142000000581000000000000
00002142000000000000000000000000000000000000000000000000000000
-0000000000000001-0000000000000001000000000000032890-00000000000001000
0000000000022070-000000000000001           -00000000000001000000000
00000001000000000000020350000011400000025000000250000000940000012800
000000210000002700000016000001200000011000000900000043000001700000
0000002700000011000000050000001700000030000001150000002000000300000
00000025000000310000000300000017000000210000001500000019000000160000
0000015100000010000000023000000130000000700000002000000014000000250000
000000030000000020000000000000010000000000000003000000020000000300000
000000000000010000000100000060000000200000008000000020000000010000
00000230000000090000000100000001700000045000000800000024000000130000
00000040000000420000001400000016000000050000001000000000000000000000

# ICPSR Data Site

**Inter-university Consortium for Political and Social Research**
http://www.icpsr.umich.edu

# ICPSR: Find Data

http://www.icpsr.umich.edu/icpsrweb/ICPSR

# ICPSR: Register to Access Data

(ORCID Identifier: https://orcid.org/)

# ICPSR: Search by Topics

## Browse by Topic

Users have two options for identifying studies related to a particular topic.

- *Topic Classifications* address topics broadly, and are meant to be a good starting point for exploring the collection.
- The Thesaurus contains both broad and narrow terms that are more specific to the subject matter of individual studies.
- Our Thematic Collections are full websites that focus on particular subject areas.

## Topic Classifications

I. Census Enumerations: Historical and Contemporary Population Characteristics
    A. United States
        1. Decennial Censuses
            a. 1790-1960 Censuses
            b. 1970 Census
            c. 1980 Census
            d. 1990 Census
            e. 2000 Census
            f. 2010 Census
        2. American Housing Survey Series
        3. Current Population Survey Series
        4. Other Census, Including County and City Data Books
    B. Nations Other Than the United States
II. Community and Urban Studies
    A. Studies of Local Politics
        1. United States
        2. Nations Other Than the United States
    B. Detroit Area Studies
    C. Historical Urban and Community Structure Studies
III. Conflict, Aggression, Violence, Wars
    A. Conflict and Stability Within Nations
    B. Conflict Between and Among Nations
IV. Economic Behavior and Attitudes
    A. Continuing Series of Consumer Surveys

# ICPSR: Search for and Compare Variables

## Search/Compare Variables

The Social Science Variables Database (SSVD) enables ICPSR users to examine and compare variables and questions across studies or series. The SSVD currently includes over 4 million variables, representing about 76% of ICPSR's holdings that have quantitative data described in statistical syntax.

The SSVD does not include our entire collection. For a comprehensive search of all datasets, use our main data search, which indexes all of the documentation available for our studies, including the variable descriptions.

| college major | GO |
|---|---|

- Browse by Study
- Browse by Series

## Comparing Variables

Compare variables using the checkboxes and the "Compare" button on the search results page.

| Results 1 - 25 | | Results per page: | 25 |
|---|---|---|---|

| Compare | Name | Label/Question |
|---|---|---|
| ☑ | HEROIN | drug of choice:heroin |
| | | D3. Indicate the drugs of choice (select all... |
| | | Taken from: Criminal Justice Drug Abuse Tr |
| ☑ | HEROIN | USED HEROIN AS MAIN DRUG |
| | | Taken from: Services Research Outcomes St |
| ☑ | HEROIN | HEROIN – ABUSER USING FIRST TIME FORCE |

## Search Tips

- Enter words or strings that are likely to appear in a variable name, label, question, and value labels:

  *Presidential election* will return variables dealing with all presidential elections

  *Presidential election Obama* will return only variables dealing with the 2008 and 2012 presidential elections

- Use quotes to search for specific phrases:

  "life satisfaction"
  "minority rights"
  "community programs"

- The minus sign may be used to remove certain types of results:

  *"Presidential debate" -"Bill Clinton"* will eliminate the debates from 1992 and 1996 in which Clinton participated

- A Boolean "and" is implied in the search.

- The search automatically does stemming (more info ↗); there's no need to type in an asterisk. It's also

# ICPSR: Compare Variables

## Compare Variables

Compare (3)   Clear All (3)   Back to search results

| ☑ MAJOR1 | ☑ MAJOR2 | ☑ OC12 |
|---|---|---|
| General Social Survey, 1972-2014 [Cumulative File] | General Social Survey, 2012 Merged Data, Including a Cultural Module [United States] | Longitudinal Study of American Youth, 1987-1994, 2007-2011: LSAY Merged Cohort File (Base File) |
| **COLLEGE MAJOR 1** | **COLLEGE MAJOR 2** | **COLLEGE MAJOR** |
| Question | Question | Question |
| "What was your major or field of study when you received your (RESPONDENT'S COLLEGE DEGREE) degree? IF THE RESPONDENT RECEIVED MORE THAN ONE GRADUATE LEVEL DEGREE, ASK ABOUT THE HIGHEST DEGREE OBTAINECOUNT UP TO TWO MENTIONS." | What was your major or field of study (RESPONDENT'S COLLEGE DEGREE) degree? <br><br> Added text: IF THE RESPONDENT RECEIVED DEGREE, ASK ABOUT THE HIGHEST DEGREE RATHER THAN AN M.B.A). [COUNT UP TO 2 | COLLEGE MAJOR |
| Responses | Responses | Responses |

| VALUE | LABEL | | VALUE | LABEL | | VALUE | LABEL | UNI FRE |
|---|---|---|---|---|---|---|---|---|
| 0 | IAP | | 0 | Inapplicable | | -99 | Not asked | 87 |

# ICPSR: Study Page and Data

# ICPSR: Reading About Data

# ICPSR: Publications linked to Data

# ICPSR: Publications Linked to Data

FIND DATA     SEARCH/COMPARE VARIABLES     DATA-RELATED PUBLICATIONS     RESOURCES FOR STUDENTS     HELP

Please take a short survey ↗ to help us improve our data-related literature resources.

## Citation
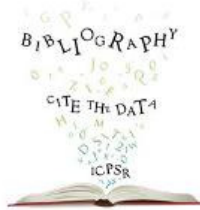
The following citation is from the ICPSR Bibliography of Data-related Literature:

Plomin, Robert, Scheier, Michael F., Bergeman, C.S., Pedersen, N.L., Nesselroade, J.R., McClearn, G.E. . Aug 1992. **Opt pessimism and mental health: A twin/adoption analysis**. *Personality and Individual Differences*. 13, (8), 921-930.
doi:10.1016/0191-8869(92)90009-E
Full Text Options: DOI ↗ WorldCat ↗ Google Scholar ↗
Export Citation: RIS EndNote XML

## Related Studies

This publication is related to the following dataset(s):

- Swedish Adoption/Twin Study on Aging (SATSA), 1984, 1987, 1990, 1993, 2004, 2007, and 2010

# Publishing and Data Management Plans

- What is a Data Management Plans
- Why do I need this?
- What does this have to do with Publishing?
- What tools are available to help?

# What is a Data Management Plan?

A data management plan is a document that describes what you will do with your data *during* your research and *after* you complete your research.

# Why do you need this?

- Funding Agencies
- Publishing
- Sharing and Archiving Data
- Regardless of the requirements, good data management is an essential skill for researchers.

Example: Williams Institute, UCLA School of Law / UCLA DataDen

# Basic Elements of a Data Management Plan

- What are your data?
- What formats will you be using?
- How will you describe this data?
- What intellectual property and privacy rights are associated with this data? (Will you need IRB approval?)
- How will you share this data? If you don't plan on sharing it, why not?
- How much will your data management cost?

See also ICPSR Elements of a Data Management Plan

# Getting Help with Data Management Plans

## ICPSR Resources for Data Management

- ICPSR Framework for Creating a Data Management Plan
- Data Management and Curation
- ICPSR Guide to Archiving Social Science Data for Institutional Repositories, 5th edition

## Other Resources

- UCLA Data Archive
- UCLA Library: Research Guides
- UKDA tools for creating and managing data (Pdf document)
- MANTRA – online learning tool/tutorial
- JISC Digital Media Advice

# Data Management and Confidentiality

If your research involves human subjects, you will need to consider both legal and ethical obligations in managing and sharing your data.

**Confidentiality** refers to the agreement between the researcher and the participant about how the participant's identifiable private information will be handled, managed, and disseminated.

As a researcher, you need a clear view about how to protect the privacy of your research subjects.

# Questions?
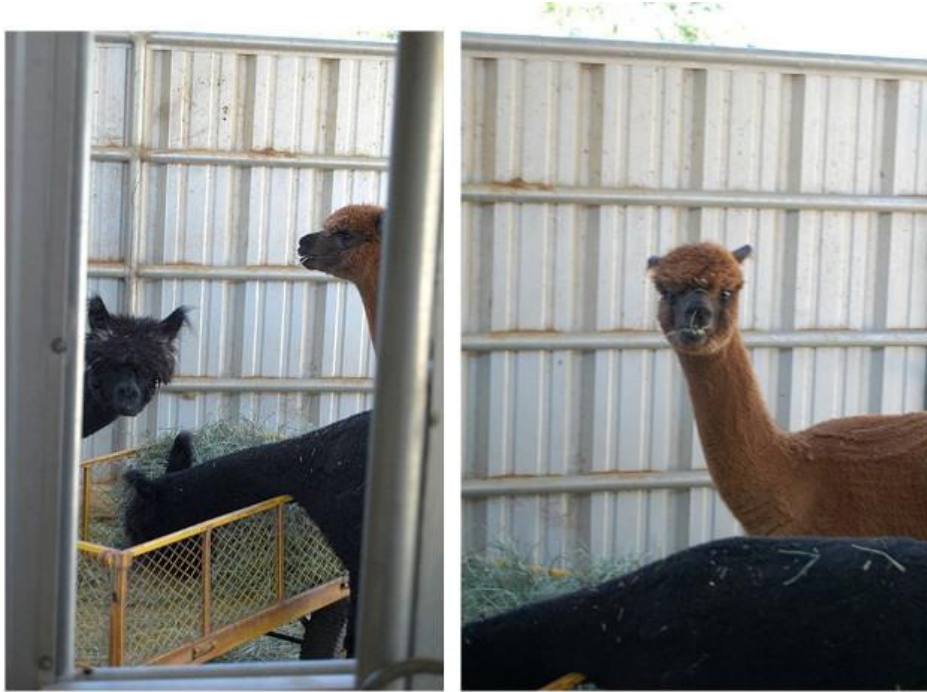
# Problems ?

# The Data Archive is here to help

1120-H Rolfe
ucla-data-archive@googlegroups.com
310-825-0716

Tim Dennis, Director of the Data Archive
Jamie Jamison, Archive Staff

presentation