



HARVARD
UNIVERSITY

Data Acquisition, Management, Security and Retention

Dustin Tingley

Associate Professor of Government

August 2014



- **Topics to be Covered:**
 - Data ownership
 - Data collection and management
 - Data security
 - Data workflow
 - Data sharing



• Data Ownership

- Sponsors/funders
 - Government-grants v. contracts
 - Private companies – usually seek to retain the right to the commercial use of data
- Third-party Data
 - Data Use Agreements



• Data Ownership (cont.)

- Pay attention to terms and conditions:
 - Who owns the data?
 - Who may access the data?
 - What rights do I have to publish?
 - Does collecting the data impose any obligations on me?
 - IRB
 - Security standards
 - Request review of any problematic terms by your sponsored programs administrator or the Office of the Vice Provost for Research prior to signing. Do not enter into agreements without approval from the institution



• Data Collection and Management

- Complying with federal and University requirements regarding the conduct of research, such as ensuring the appropriate use of animals, human subjects, recombinant DNA, radioactive materials, “select agents,” etc;
- Ensuring that research is conducted responsibly;
- Adhering to the terms and fulfilling any applicable scopes of work of project agreements and subcontracts and sub-awards;
- Protecting the rights of students, postdoctoral scholars, and staff, including, but not limited to, their rights to access data from research in which they participated;
- Securing intellectual property rights; and
- Making research records, including data and materials, available to colleagues, administrators, funders or others who have a legitimate need to examine the propriety of expenditures, or to examine data in order to verify their accuracy and/or to review and replicate research findings.



- **Data Collection and Management (cont'd)**
- Authorization/Permission
 - IRB
 - IACUC (animal care)
- Documentation (see subsequent discussion)
 - Hard-copy data
 - Electronic data



- **Retention of Research Records and Data**
 - Records must be retained 7 years after the end of a research project or activity.
 - Longer retention periods may apply:
 - In order to protect any intellectual property resulting from the research work
 - As required by external government or other funding source or sponsor
 - If needed in connection with pending litigation, or fact-finding for research integrity, or human subjects or animal use purposes.
 - If the research involves children or individuals with mental incapacity



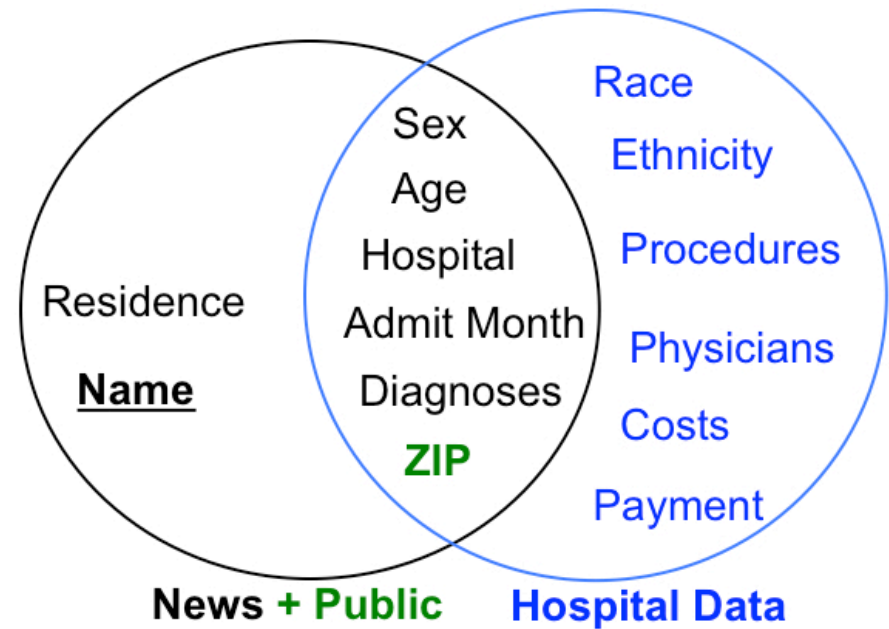
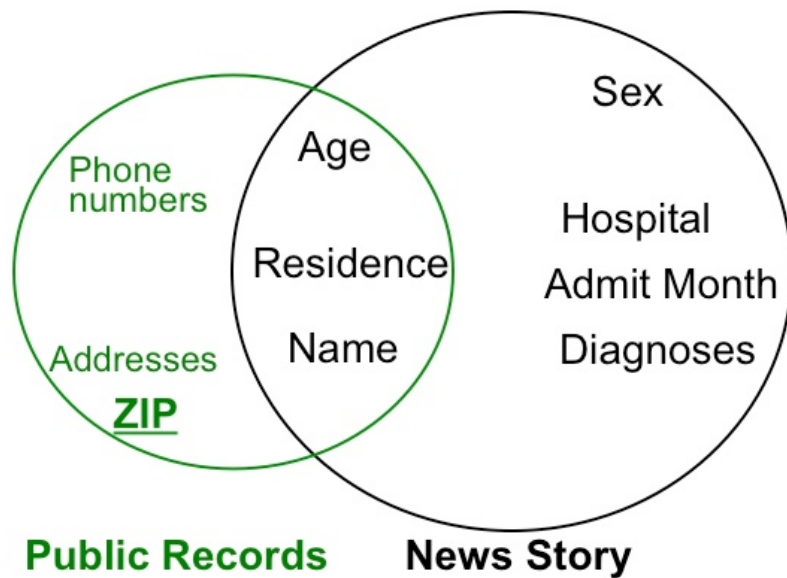
- **Retention of Research Records and Data (cont'd)**
 - **7 year retention requirement does not apply:**
 - To documents or data subject to IRB destruction or de-identification mandates.
 - To confidential or third-party data subject to Data Use Agreements that require return or destruction of data prior to the expiration of the retention period.
- My view: just put your data in a repository.

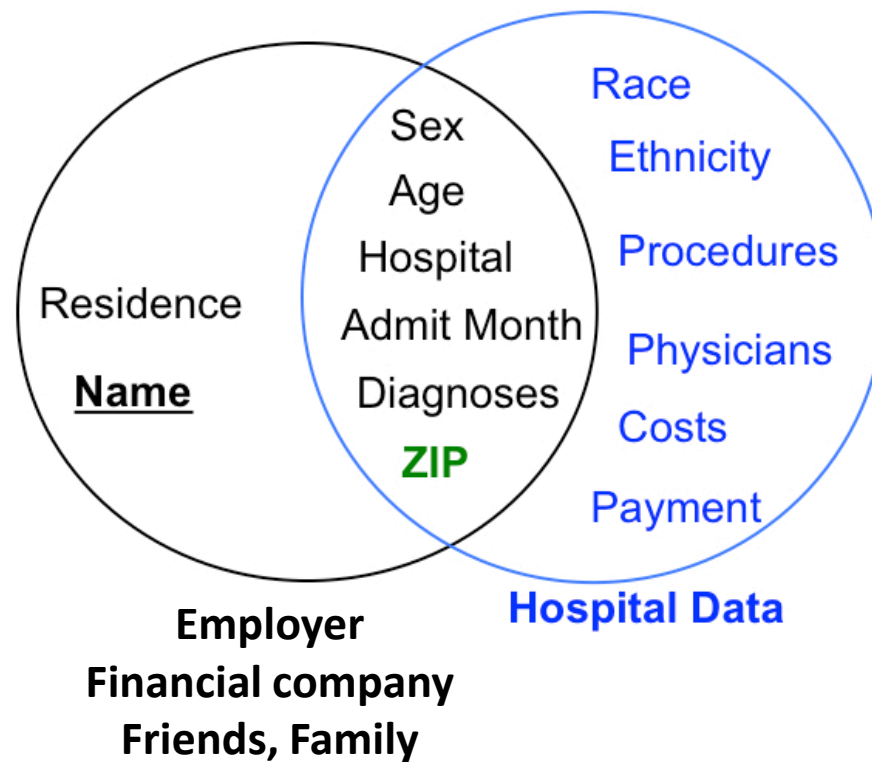


• **Data Security**

- Data security is essential to the protection of research participants' privacy
- Researchers should safeguard research data and findings.
- Unauthorized individuals must not access the research data or learn the identity of research participants

Washington State Health Database







- **Data Security**

Research data may be subject to security terms set forth by:

- **Contracts (Sponsor or Provider)**

- **IRB/HUIT**

- **HIPAA or other Federal Regulations**

- **State**



• Data Security Categories

- ❑ Level 1: De-identified research information about people and other non-confidential research information
- ❑ Level 2: Benign information about individually identifiable people – but confidentiality has been promised
- ❑ Level 3: Sensitive information about individually identifiable people
- ❑ Level 4: Very sensitive information about individually identifiable people
- ❑ Level 5: Extremely sensitive information about individually identifiable people



- **Data Security Categories (cont'd):**

- Level 1: no specific University requirements
- Level 2: Password protection
- Level 3: Must not be directly accessible from the Internet (i.e. email) unless the data is encrypted
- Level 4: servers must be located only in physically secure facilities under University control
- Level 5: information must be stored off network and used only in physically secured rooms controlled by University. No master keys are allowed



- **Identifying Information:**
 - Broadly defined
 - Not just name, address, social security number, etc.
 - Includes any item or **combination of items** that could lead directly or indirectly to the identification of a research participant



Data workflow

Collecting data

- Keep a log book (perhaps via a google projects page)

Processing data

- Point and click is your enemy!
- Always process data with script files

Analyzing data

- Point and click is your enemy!
- Always analyze data with script files

Be aware of “abstraction violations” in code. If you are copy and pasting code, you probably are doing something inefficiently.



- **Sharing data?**
 - NIH believes all data should be considered for data sharing, I agree
 - Data should be made as widely and freely available as possible while safeguarding the privacy of participants, and protecting confidential and proprietary data
 - Data sharing plan required for applications requesting >\$500,000/year from NIH
- **Why is sharing/archiving good for YOU!**
- **Data Sharing/Archiving Resources: Dataverse, ICPSR**

http://thedata.harvard.edu
http://thedata.harvard.edu

The screenshot displays the Harvard Dataverse Network homepage. At the top, it features logos for IQSS (The Institute for Quantitative Social Science), Harvard Library, and research DATA collaborative. The main header includes the text "Share, Cite, Reuse, Archive Research Data" and "Scientific data for reproducible research". Below this, the site is powered by the Dataverse Network PROJECT v. 3.4. A search bar is present with a "Search" button and links for "Advanced Search" and "Tips". A prominent statistics banner shows "52,013 Studies, 723,402 Files, 762,639 Downloads". The page is divided into sections for "Dataverses" (with a "Create Dataverse" button) and "Studies" (with a "Create Study" button). Each section includes a brief description and a list of "RECENTLY RELEASED" items with their titles and dates.

Harvard Dataverse Network

52,013 Studies, 723,402 Files, 762,639 Downloads

Dataverses Create Dataverse
530 Dataverses

Studies 52,013 Studies, 723,402 Files, 762,639 Downloads

RECENTLY RELEASED DATAVERSES

Korzennik, Sylvain	Jun 5, 2013
The Golden Rice	Jun 2, 2013
Márquez, Javier	May 31, 2013

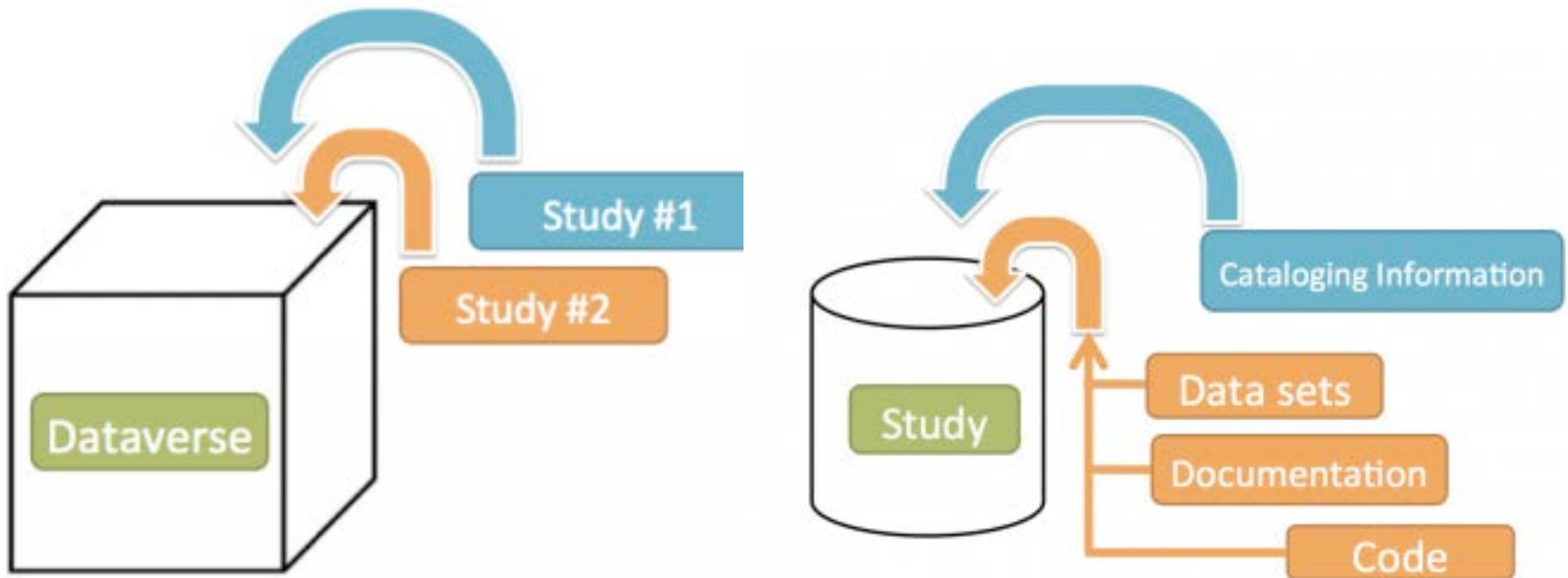
RECENTLY RELEASED STUDIES

2012 Survey of the Performance of American Elections by Stewart, Charles	Jun 6, 2013
Effects of Organized Family Planning Programs on U.S. Adolescent Fertility, 1970-1975 by Jacqueline Darroch Forrest	Jun 6, 2013

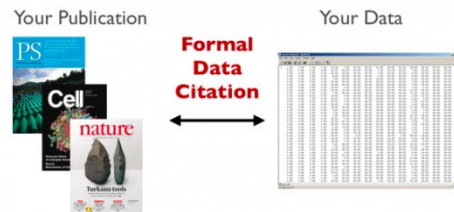
- The **Harvard Dataverse Network** is a data sharing repository open to **all** research data from **all** domains.
- The Dataverse Network software is **open-source**, installed in institutions across the world (**http://thedata.org**)

Dataverse: Container for research studies

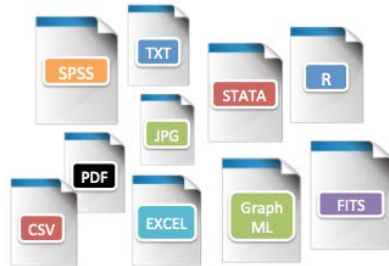
Study: Container for data, documentation, and code



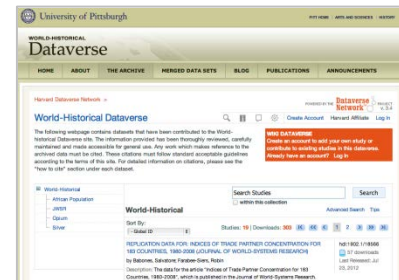
Data sharing and archiving with control and recognition for data authors, distributors



Persistent Data Citations
permanently linking your data to
your publication (Altman, King, 2007)



Support for all file types
any format, max 2 GB per file



Customized Branding
or embed on your site



Data Restrictions
& terms of use options, although
encouraging **Open Data**

Rich data support for some data formats



SPSS, Stata, R Data

metadata extraction, subsetting
& analysis (R, Zelig)



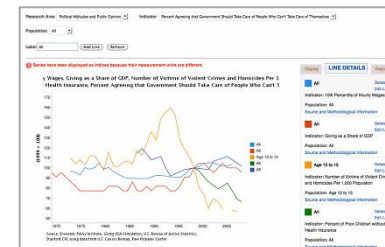
FITS Data

metadata extraction from file
header



Social Network Data (GraphML)

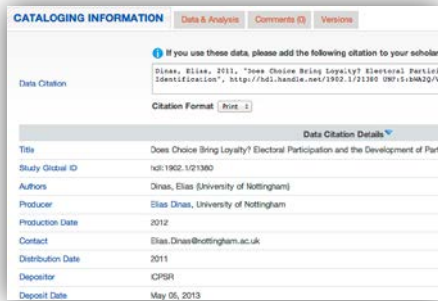
smart queries & subsetting



Data visualizations

for time series

Data management, standards and archival good practices



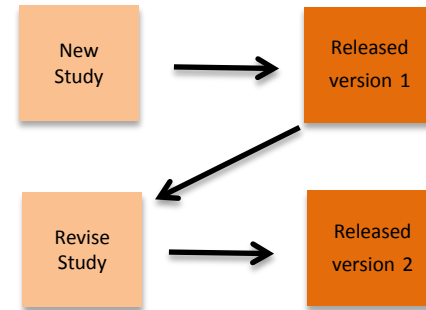
CATALOGING INFORMATION	
Data Citation	
If you use these data, please add the following citation to your scholarship: Dinas, Elias, 2011, "Does Choice Bring Loyalty? Electoral Participation and the Development of Part...	
Citation Format <input type="text" value="print"/>	
Data Citation Details	
Title	Does Choice Bring Loyalty? Electoral Participation and the Development of Part...
Study Global ID	doi:10.21203/1.1300
Authors	Dinas, Elias (University of Nottingham)
Producer	Elias Dinas, University of Nottingham
Production Date	2012
Contact	Elias.Dinas@nottingham.ac.uk
Distribution Date	2011
Depositor	CPDR
Deposit Date	May 05, 2013

Data Cataloging

self-curated, with custom metadata templates (DDI, Dublin Core)

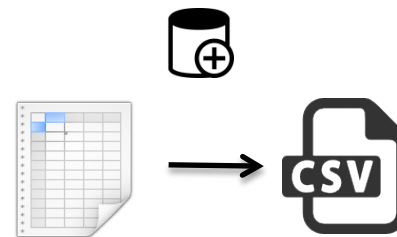


Log traffic & downloads to your dataset with Guestbook



Data Versioning

preserve & cite previous versions



Permanent storage

preservation format with w/copies in multiple locations (OAI-PMH, LOCKSS)

Backup slides

Record	505825338
Hospital	162: Sacred Heart Medical Center in Providence
Admit Type	1: Emergency
Type of Stay	1: Inpatient
Length of Stay	6 days
Discharge Date	Oct-2011
Discharge Status	6: <u>Dsch/Trfn</u> to home under the care of an health service organization
Charges	\$71708.47
Payers	1: Medicare 6: Commercial insurance 625: Other government sponsored patients
Emergency Codes	E8162: motor vehicle traffic accident due to loss of control; loss control mv- <u>mocycl</u>
Diagnosis Codes	80843: closed fracture of other specified part of pelvis 51851: pulmonary insufficiency following trauma & surgery

	86500: injury to spleen without mention of open wound into cavity
	80705: closed fracture of rib(s); fracture five ribs-close
	5849: acute renal failure; unspecified
	8052: closed fracture of dorsal [thoracic] vertebra without mention of spinal cord injury
	2761: <u>hyposmolality</u> &/or <u>hyponatremia</u>
	78057: tachycardia
	2851: acute <u>posthemorrhagic anemia</u>
Age in Years	60
Age in Months	725
Gender	Male
ZIP	98851
State Reside	WA
Race/Ethnicity	White, Non-Hispanic

MAN, 60, THROWN FROM MOTORCYCLE

A 60-year-old Soap Lake man was hospitalized Saturday afternoon after he was thrown from his motorcycle. Ronald Jameson was riding his 2003 Harley-Davidson north on Highway 25, when he failed to negotiate a curve to the left. His motorcycle became airborne before landing in a wooded area. Jameson was thrown from the bike; he was wearing a helmet during the 12:24 p.m. incident. He was taken to Sacred Heart Hospital. The police cited speed as the cause of the crash.
[News Review 10/18/2011]

Washington State Health Database

