

# Data Acquisition, Management, Security and Retention

Mercè Crosas

Director of Data Science

Institute for Quantitative Social Science

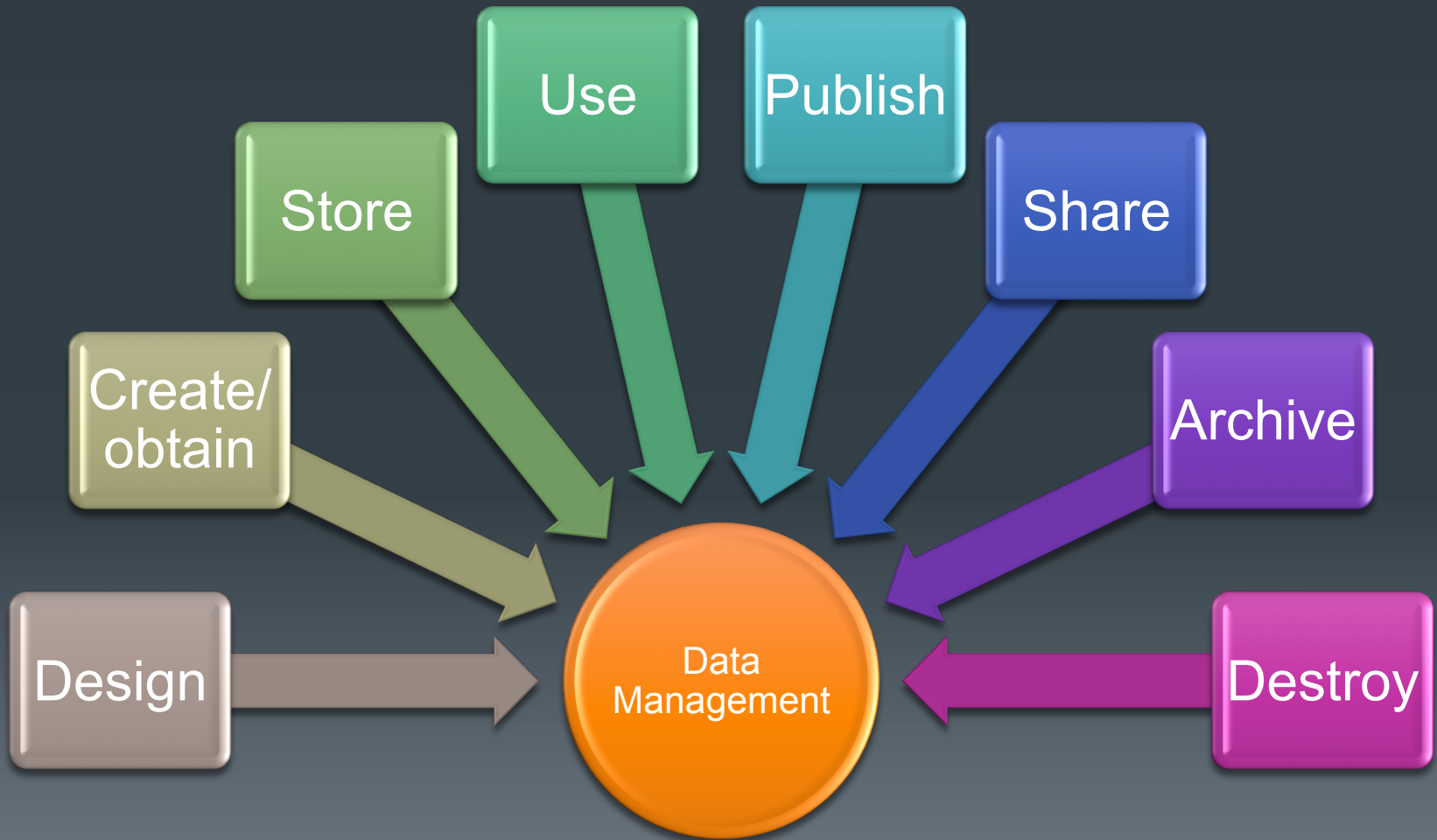
Kristen Bolt

Research Data Officer

Office of Vice Provost of Research

August 12, 2015

# Data Lifecycle





1. Data Acquisition, Management Security and Retention: What You Need to Know

2. Data Sharing: It's Good for You and for the World

# What you need to know for Harvard



## Harvard Policies and Practices:

- Human Subjects Data – the IRB
- Research with Animals – the IACUC
- Retention Policy: 7 years and exceptions
- Harvard Research Data Security Policy

# What you need to know for Funders

## Funding Requirements:

- NSF, NIH, require public access plans
- Foundations: Sloan, Gates (open data policy) – and public access plans
- Most plans must be created and submitted as part of the funding application process
- Plan ahead!

# What you need to know Fed & State

## Regulatory Restrictions:

- HIPAA (18+ identifiers – alone or in combination data sets)
  - Informed Consent (IRB, secondary use of data – HIPAA waivers)
- FERPA (education information and special protections)
- MA residence state law
- Stem Cell data and Genomics data must be published in approved repository ,but also must be de-identified.

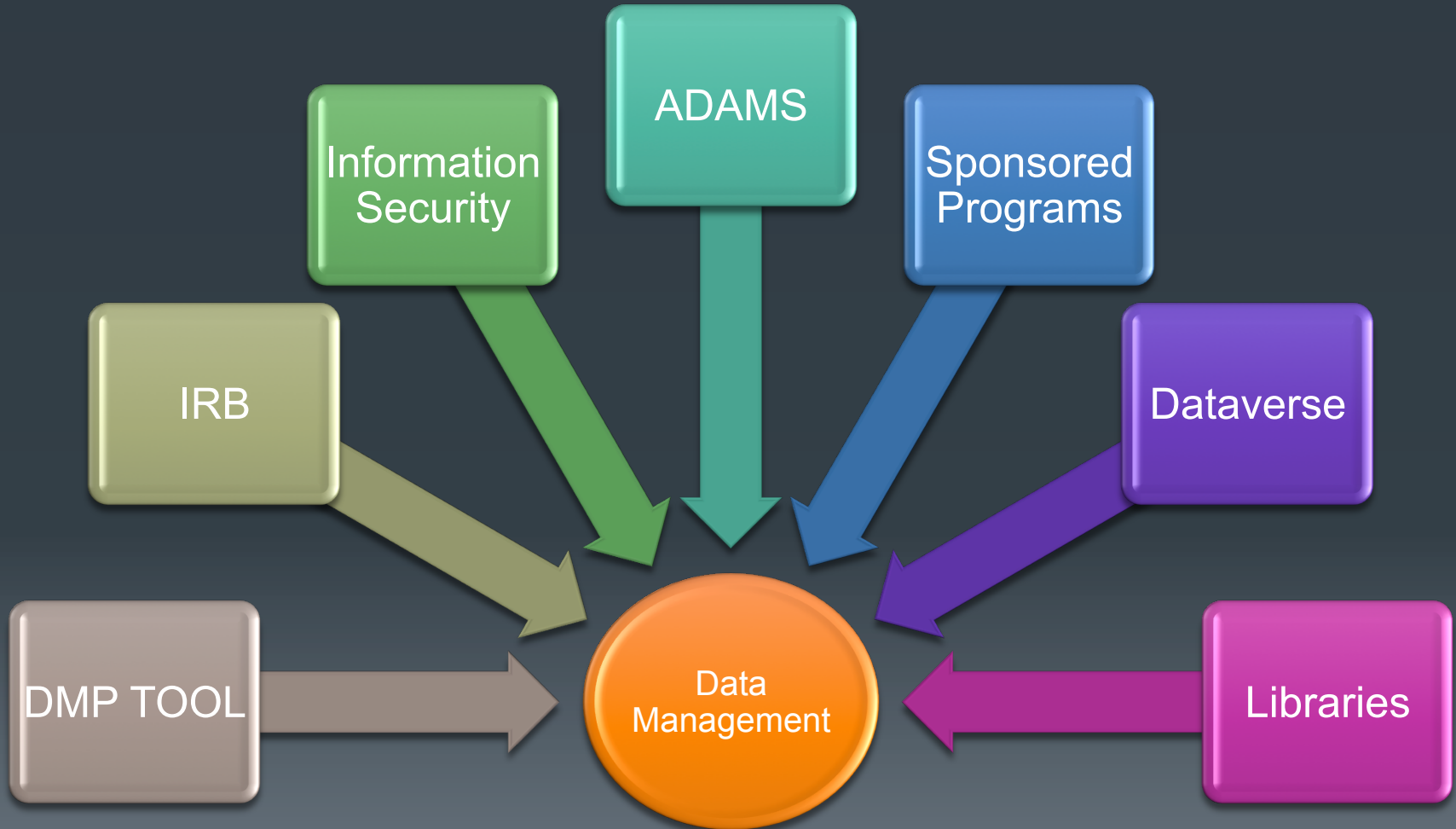
# What you need to know for 3<sup>rd</sup> Party Data



## Removing Research Restrictions:

- Third-party data (you did not collect the data; you are accessing existing data that was gathered for other purposes)
- Data user agreements (any contract, even an invoice can have terms that can restrict your IP or publication rights)
- Licensing agreements – can also include restrictions

# Resources for Harvard





# Resources for Harvard

- **Data Management Plans**
  - [DMPTOOL.ORG](http://DMPTOOL.ORG)
  - Funder templates
  - Hundreds of examples
- **IRB**
  - Collection plans
  - Data Levels
- [RDSAP@harvard.edu](mailto:RDSAP@harvard.edu)
- **Information Security**
  - [itsec-ec@harvard.edu](mailto:itsec-ec@harvard.edu)
  - School Officers
  - Secure access
- [ADAMS.harvard.edu](http://ADAMS.harvard.edu)
  - GSE, Econ
  - DUAs
  - Electronic workflow
- **Sponsored Programs**
  - Negotiate contracts
  - Institutional Signature
  - Publication restriction review
- **Publishing Data and Data Retention**
  - Dataverse
  - Libraries
  - Genomic Data
- <http://vpr.harvard.edu>
- Links to all resources
- DMP tool guidance, data security level examples and worksheets



1. Data Acquisition, Management Security and Retention: What You Need to Know

2. Data Sharing: It's Good for You and for the World

“Ideally, research protocols should be registered in advance and monitored in virtual notebooks.”

“Where possible, trial data also should be open for other researchers to inspect and test.”

## Problems with scientific research

# How science goes wrong

Scientific research has changed the world. Now it needs to change itself

Oct 19th 2013 | From the print edition

f Like

15k

Tweet

1,120



Trust, but VERIFY

A SIMPLE idea underpins science: “trust, but verify”. Results should always be subject to challenge from experiment. That simple but powerful idea has generated a vast body of knowledge. Since its birth in the 17th century, modern science has changed the world beyond recognition, and overwhelmingly for the better.



“Instances in which scientists detect and address flaws in work constitute evidence of success, not failure.”

“Ensuring that the integrity of science is protected is the responsibility of many stakeholders.”



## SCIENTIFIC INTEGRITY

# Self-correction in science at work

Improve incentives to support research integrity

By Bruce Alberts,<sup>1</sup> Ralph J. Cicerone,<sup>2</sup> Stephen E. Fienberg,<sup>3</sup> Alexander Kamb,<sup>4</sup> Marcia McNutt,<sup>5\*</sup> Robert M. Nerem,<sup>6</sup> Randy Schekman,<sup>7</sup> Richard Shiffrin,<sup>8</sup> Victoria Stodden,<sup>9</sup> Subra Suresh,<sup>10</sup> Maria T. Zuber,<sup>11</sup> Barbara Kline Pope,<sup>12</sup> Kathleen Hall Jamieson<sup>13,14</sup>

**W**EEK after week, news outlets carry word of new scientific discoveries, but the media sometimes give suspect science equal play with substantive discoveries. Careful qualifications about what is known are lost in categorical headlines. Rare instances of misconduct or instances of irreproducibility are translated into concerns that science is broken. The Octo-

ber 2013 *Economist* headline proclaimed “Trouble at the lab: Scientists like to think of science as self-correcting. To an alarming degree, it is not” (1). Yet, that article is also rich with instances of science both policing itself, which is how the problems came to *The Economist’s* attention in the first place, and addressing discovered lapses and irreproducibility concerns. In light of such issues and efforts, the U.S. National Academy of Sciences (NAS) and the Annenberg Retreat at Sunnyslans convened our group to examine ways to remove some of the current disincentives to high standards of integrity in science.

Like all human endeavors, science is imperfect. However, as Robert Merton noted more than half a century ago “the

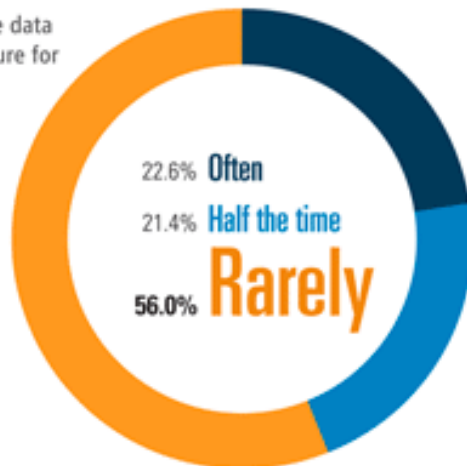
activities of scientists are subject to rigorous policing, to a degree perhaps unparalleled in any other field of activity” (2). As a result, as Popper argued, “science is one of the very few human activities—perhaps the only one—in which errors are systematically criticized and fairly often, in time, corrected” (3). Instances in which scientists detect and address flaws in work constitute evidence of success, not failure, because they demonstrate the underlying protective mechanisms of science at work.

Still, as in any human venture, science writ large does not always live up to its ideals. Although attempts to replicate the 1998 Wakefield study alleging an association between autism and the MMR (measles,

# Science Survey

How often do you access or use data sets from the published literature for your original research papers?

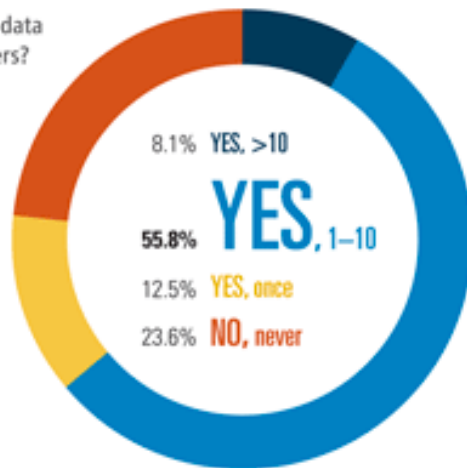
From archival databases?



Access data from published work

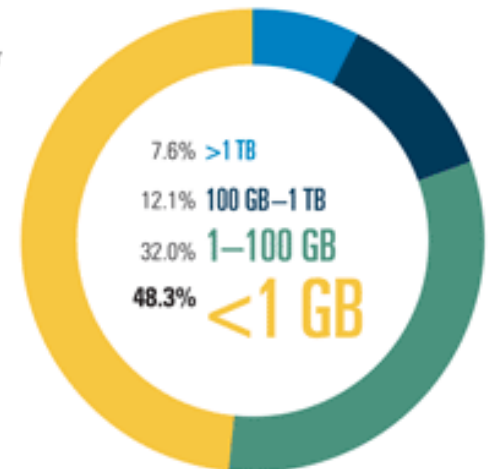
Have you asked colleagues for data related to their published papers?

If you answered yes, have the appropriate data been provided?



Ask colleagues for data

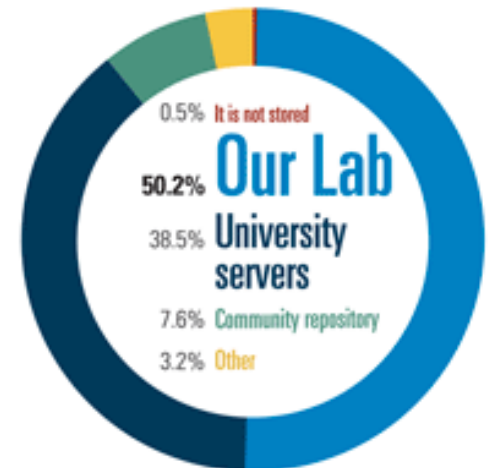
What is the size of the largest data set that you have used or generated in your research?



Size of data

Where do you archive most of the data generated in your lab or for your research?

“ Even within a single institution there are no standards for storing data, so each lab, or often each fellow, uses ad hoc approaches. ”



Archival location

# Why Share Data in Public Repositories?

- Good for the world:
  - To reproduce research
  - To make public assets available to the public
  - To leverage investments in research data
  - To advance research and innovation
- Data in your computer or web site is likely not to be accessible or reusable after 10 years
- When you publish your data, you get credit and more citations



# The Care and Feeding of Scientific Data: 10 Simple Rules



- Rule 1: Love your data, and let others love it too
- Rule 2: Share your data online, with a permanent identifier
- Rule 3: Conduct science with reuse in mind
- Rule 4: Publish workflow as context
- Rule 5: Link your data to your publications as early as possible
- Rule 6: Publish your code
- Rule 7: Say how you want to get credit for your data (and software)
- Rule 8: Foster and use data repositories
- Rule 9: Reward colleagues who share their data properly
- Rule 10: Help establish data science and data scientist as vital




Harvard Dataverse

A collaboration with Harvard Library, Harvard University IT, and IQSS




Share, publish, and archive your data. Find and cite data across all research fields.


<




World Agroforestry Centre  
ICRAF Dataverse



Population Services International  
(PSI) Dataverse



International Food Policy  
Research Institute (IFPRI)  
Dataverse



Murray Research Archive  
Dataverse

>

Q Find
Advanced Search

+ Add Data

- Dataverses (1,196)**
- Datasets (58,999)**
- Files (278,969)**

**Dataverse Category**

- Research Project (128)
- Organization or Institution (112)
- Researcher (81)
- Journal (46)
- Teaching Course (7)

---

**Publication Date**

- 2015 (14,521)
- 2011 (10,131)
- 2007 (9,586)
- 2012 (8,670)
- 2009 (6,251)

More...

---

**Subject**

- Social Sciences (4,696)
- Earth and Environmental Sciences

1 to 10 of 60,195 Results Sort « < Previous 1 2 3 4 5 Next > »

**Replication Data For: Policy Rhetoric Can Have Economic Consequences: Presidential Rhetorical Liberalism and Economic Policy Uncertainty**

Aug 11, 2015 - Christopher Olds Dataverse

Olds, Christopher, 2015, "Replication Data For: Policy Rhetoric Can Have Economic Consequences: Presidential Rhetorical Liberalism and Economic Policy Uncertainty", <http://dx.doi.org/10.7910/DVN/IAUUQJ>, Harvard Dataverse, V1

The response of economic variables to changes in the policy ideology expressed through presidential rhetoric is an area in need of extensive empirical exploration. An increase in liberal policy rhetoric can serve as an indicator of an executive branch that will, in general, attem...

**Replication Data for Attention&SA**

Aug 11, 2015

Cao, Liyu, 2015, "Replication Data for Attention&SA", <http://dx.doi.org/10.7910/DVN/GWMBFJ>, Harvard Dataverse, V1

This dataset is described in manuscript titled 'Attention wins over sensory attenuation in a sound detection task'.

**Replication Data For: The Negative Effect of Economic Policy Uncertainty on Presidential Rhetorical Optimism About the Economy in the United States**

Aug 11, 2015 - Christopher Olds Dataverse

Olds, Christopher, 2015, "Replication Data For: The Negative Effect of Economic Policy Uncertainty on Presidential Rhetorical Optimism About the Economy in the United States", <http://dx.doi.org/10.7910/DVN/5JHRZO>, Harvard Dataverse, V1



# Resources: Harvard Dataverse

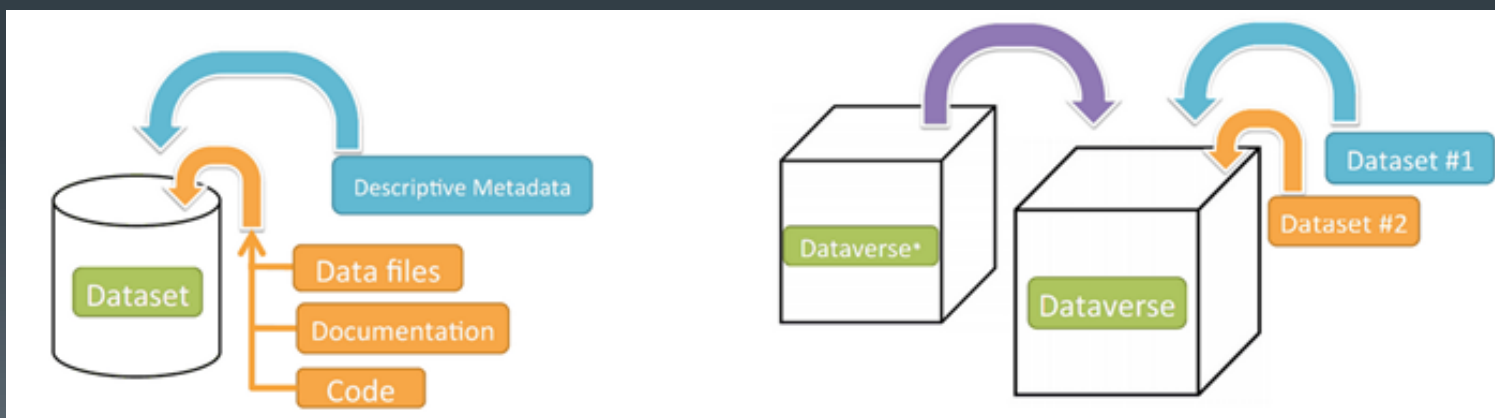


## Why we built the Dataverse?

- Provide a repository to help Harvard researchers find, share, archive, cite, and explore research data
- Satisfy Harvard researchers' insatiable need for data
- Give credit and control to data authors and distributors
- Fulfill funding agencies' data management plan requirements for Harvard researchers
- Help Harvard researchers comply with journals' data policies
- Open source these benefits to researchers worldwide

# Dataverse: Organization

- **Dataset:** contains data files, documentation, code, and metadata
- **Dataverse:** a full featured (virtual) archive, containing datasets and other dataverses, appearing on your website, branded as yours



# Dataverse: Features

- Standard, persistent data citation
- Branding for each dataverse
- Standard, extensible metadata:
  - citation metadata
  - domain-specific metadata
  - file-level metadata
- Faceted search for all metadata
- Multiple levels of access control
  - CC0/ terms of use/ restricted/ guestbook
- Multiple roles and permissions
- Versioning
- Re-formatting of tabular data files
- Separate metadata from data upon data file upload
- APIs to integrate with journals, data visualizations and analysis



# HARVARD ELECTION DATA ARCHIVE

Sharing and Improving Election Data



Election Data Archive Dataverse (Harvard University) [Home Page](#)

[Harvard Dataverse](#) > **Election Data Archive Dataverse**



The Harvard Election Data Archive is a space to share and improve election data. You can create an account and either correct the cataloging information for the studies in this dataverse or upload new data files. Thank you to our colleagues Michael McDonald and David Bradlee for their help with data collection.

The Harvard Election Data Archive by [Stephen Ansolabehere](#) and [Jonathan Rodden](#) is licensed under a [Creative Commons Attribution 3.0 Unported License](#).

[Advanced Search](#)

[Dataverses \(0\)](#)

[Datasets \(45\)](#)

[Files \(584\)](#)

**Publication Date**

2011 (43)

2014 (1)

2015 (1)

1 to 10 of 45 Results

Sort ▾

« < Previous **1** 2 3 4 5 Next > »

[Virginia State Legislative Data Files](#)

Jul 13, 2015



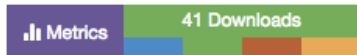
Ansolabehere, Stephen, 2015, "Virginia State Legislative Data Files", <http://dx.doi.org/10.7910/DVN/ERZWUK>, Harvard Dataverse, V1

Shapefile and Election Results

Election Data Archive, a collaboration to share and Improve election data (Steve Ansolabehere, Department of Government at Harvard)

## Ebola Kenema Dataverse (Harvard University)

Harvard Dataverse > Ebola Kenema Dataverse > **Clinical Illness and Outcomes in Patients with Ebola in Sierra Leone**



### Clinical Illness and Outcomes in Patients with Ebola in Sierra Leone

Schieffelin, John; Shaffer, Jeffrey; Goba, Augustine; Gbakie, Michael; Gire, Stephen; Colubri, Andres; Sealfon, Rachel; Kanneh, Lansana; Moigboi, Alex; Momoh, Mambu; Fullah, Mohammed; Moses, Lina; Brown, Bethany; Andersen, Kristian; Winnicki, Sarah; Schaffner, Stephen; Park, Daniel; Yozwiak, Nathan; Jiang, Pan-Pan; Kargbo, David; Jalloh, Simbirie; Fonnies, Mbalu; Sinnah, Vand; French, Issa; Kovoma, Alice; Kamara, Fatima; Tucker, Veronica; Konuwa, Edwin; Sellu, Josephine; Mustapha, Ibrahim; Foday, Momoh; Yillah, Mohamed; Kanneh, Franklyn; Saffa, Sidiki; Massally, James; Boisen, Matt; Branco, Luis; Vand, Mohamed; Grant, Donald; Happi, Christian; Gevao, Sahr; Fletcher, Thomas; Fowler, Robert; Bausch, Daniel; Sabeti, Pardis; Khan, Humarr; Garry, Robert, 2015, "Clinical Illness and Outcomes in Patients with Ebola in Sierra Leone", <http://dx.doi.org/10.7910/DVN/29296>, Harvard Dataverse, V1

Download Citation ▾

If you use these data, please add this citation to your scholarly resources. [Learn about Data Citation Standards.](#)

#### Description

This data comprises a total of 213 cases evaluated for Ebola virus infection at the Kenema Government Hospital in Sierra Leone between May 25 and June 18, 2014. Outcome data was available for 87 of 106 EBOV positive cases. Metabolic panels were performed on 98 Ebola virus disease and non-Ebola virus disease illness patients with adequate samples volumes. Ebola virus load was determined in 63 cases with adequate samples volumes by quantitative polymerase chain reaction (qPCR) at Harvard University. Sign and symptom data was obtained on 44 patients with a clinical chart that were admitted to Kenema Hospital. The metabolic panels were obtained from serum samples analyzed with a Piccolo Blood Chemistry Analyzer and Comprehensive Metabolic Reagent Discs (Abaxis).

#### Keyword

Ebola, clinical data, laboratory data, outbreak, fatality rate

#### Related Publication

Clinical Illness and Outcomes in Patients with Ebola in Sierra Leone. John S. Schieffelin, et al. N Engl J Med 2014; 371:2092-2100 November 27, 2014 DOI: 10.1056/NEJMoa1411680 [doi: 10.1056/NEJMoa1411680](https://doi.org/10.1056/NEJMoa1411680)

Data set on Ebola from article published in N England Journal of Medicine (Pardis Sabeti, Department of Organismic and Evolutionary Biology at Harvard)



**Robert J. Sampson Dataverse** (Harvard University, Department of Sociology)

[Home Page](#)

[Harvard Dataverse](#) > **Robert J. Sampson Dataverse**




[Q Find](#)

[Advanced Search](#)

[Dataverses \(0\)](#)

[Datasets \(3\)](#)

[Files \(28\)](#)

**Publication Date**

2009 (3)

**Author Name**

[Felton J. Earls \(3\)](#)

[Jeanne Brooks-Gunn \(1\)](#)

[Robert J. Sampson \(1\)](#)

[Stephen W. Raudenbush \(1\)](#)

1 to 3 of 3 Results

Sort ▾

[Project on Human Development in Chicago Neighborhoods: Systematic Social Observation, 1994-1998](#)

Mar 18, 2009 - [Murray Research Archive Dataverse](#)



Felton J. Earls, 2009, "Project on Human Development in Chicago Neighborhoods: Systematic Social Observation, 1994-1998", <http://hdl.handle.net/1902.1/01952>, Harvard Dataverse, V1

The purpose of the Project on Human Development in Chicago Neighborhoods was to understand how families, schools, and neighborhoods affect child and adolescent development. This included an extensive undertaking on understanding the causes and the pathways of juvenile delinquency...

[Project on Human Development in Chicago Neighborhoods: Longitudinal Cohort Study, 1994-2001](#)

Mar 18, 2009 - [Murray Research Archive Dataverse](#)



Felton J. Earls, 2009, "Project on Human Development in Chicago Neighborhoods: Longitudinal Cohort Study, 1994-2001", <http://hdl.handle.net/1902.1/01953>, Harvard Dataverse, V1

The purpose of the Project on Human Development in Chicago Neighborhoods (PHDCN) was to examine how families, schools, and neighborhoods affect child and adolescent development. This included understanding the causes and the pathways of juvenile delinquency, adult crime, substanc...

[Project on Human Development in Chicago Neighborhoods: Community Survey, 1994 - 1995](#)

Dataverse for Robert Sampson's data (Department of Sociology at Harvard)

Technology Science Dataverse (Harvard University)


Harvard Dataverse > Technology Science Dataverse



This dataverse has datasets published in Technology Science at techscience.org

 Find

[Advanced Search](#)

 Add Data

 **Dataverses (0)**

 **Datasets (4)**

 **Files (59)**

**Publication Date**

2015 (4)

**Subject**

Computer and Information Science

(3)

Social Sciences (3)

**Author Name**

Gee, Grace (1)


Khanna, Aran (1)

Rahman, Mohammed (1)

Rose, Michael (1)

Shore, Jennifer (1)

1 to 4 of 4 Results

 Sort ▾

**Replication Data for: Facebook's Privacy Incident Response, a study of geolocation sharing on Facebook Messenger** 

Aug 10, 2015



Khanna, Aran, 2015, "Replication Data for: Facebook's Privacy Incident Response, a study of geolocation sharing on Facebook Messenger", <http://dx.doi.org/10.7910/DVN/D2SNRI>, Harvard Dataverse, V1 [UNF:6:hiXa2O0z0wPt9CL8yBGHDA==]

This dataset was used for this paper published on 8/11/2015 on Technology Science <http://techscience.org/a/2015081104/>

**Replication Data for: Did You Really Agree to That? The Evolution of Facebook's Privacy Policy** 

Aug 6, 2015



Steinman, Jill; Shore, Jennifer, 2015, "Replication Data for: Did You Really Agree to That? The Evolution of Facebook's Privacy Policy", <http://dx.doi.org/10.7910/DVN/JROUKG>, Harvard Dataverse, V1

This dataset was used for this paper published on 8/11/2015 on Technology Science <http://techscience.org/a/2015081102/>

**Replication Data for: "Who's Paying More to Tour These United States? Price Differences in International Travel Bookings"** 

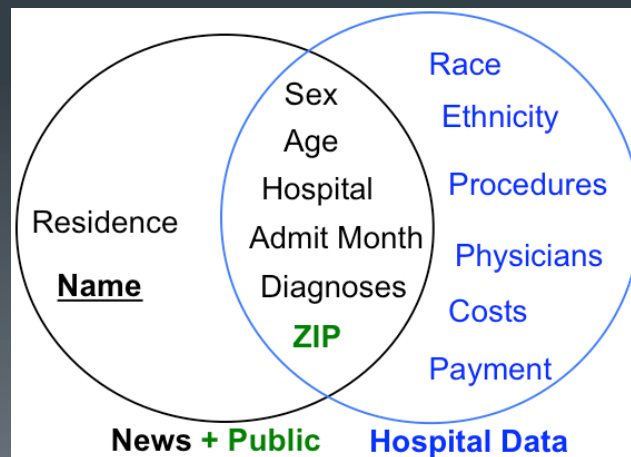
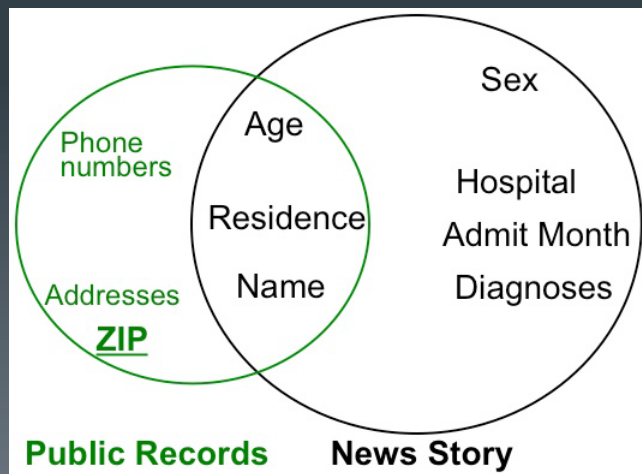
Aug 6, 2015

Dataverse for replication data from the new Journal of Technology Science published at Harvard



# Sharing Sensitive Data

- Even when data sets are anonymized, re-identification is increasingly possible:
  - Sweeney (2000) showed that 87 % of all Americans could be uniquely identified using only three bits of information: ZIP code, birthdate, and sex.
  - In 2013, Sweeney also showed that combining the Washington State Health Database with news (accidents, hospitalized people) could re-identify 43% of the records.





# DataTags and Dataverse

- DataTags helps researchers share sensitive data with confidence.
- DataTags will integrate with Dataverse in 2016
- A collaboration with Harvard SEAS, IQSS, Data Privacy Lab, Berkman Center, and MIT Information Sciences.

Tag Type	Description	Storage & Transit	Access
Blue	Non-confidential information, stored and shared freely.	Clear	Open
Green	Not harmful personal information, shared with some access control.	Clear	Email, OAuth verified registration
Yellow	Potentially harmful personal information, shared with loosely verified and/or approved recipients.	Encrypted	Password, Registered , Approval click-through DUA
Orange	Sensitive personal information, shared with verified and/or approved recipients under agreement.	Encrypted	Password, Registered, Approval, signed DUA
Red	Very sensitive personal information, shared with strong verification of approved recipients under signed agreement.	Encrypted	Two-factor Auth, Registered, Approval, signed DUA
Crimson	Maximum sensitive, explicit permission for each transaction, strong verification of approved recipients under signed agreement.	Double Encrypted	Two-factor Auth, Registered, Approval, signed DUA

# Data (and code) Sharing Resources and References

- Harvard Dataverse: <https://dataverse.harvard.edu/>
- Open Data Assistance Program @Harvard, open office and training for Dataverse: <http://projects.iq.harvard.edu/odap>
- Data Privacy Lab: <http://dataprivacylab.org/>
- DataTags: <http://datatags.org>
- Open Science Framework by the Center for Open Science: Manage and organize your entire research project
- IPython Notebook: Keep track of your research workflow and notes
- GitHub: Share openly your code, with version control
- *Alberts, Cicerone, Fienberg et al, 2015. Self-correction in Science at Work: <http://www.sciencemag.org/content/348/6242/1420.full.pdf>*
- *The Economist, 2013. How Science Goes Wrong*
- *Science 11 February 2011: Vol. 331 no. 6018 pp. 692-69*



Thank you