



Mercè Crosas, Ph.D.

Twitter: @mercecrosas

Director of Data Science, IQSS

Harvard University



http://thedata.harvard.edu

IQSS The Institute for Quantitative Social Science HARVARD UNIVERSITY

HARVARD LIBRARY

research DATA collaborative

Share, Cite, Reuse, Archive Research Data
Scientific data for reproducible research

POWERED BY THE **Dataverse Network** PROJECT v. 3.4

Harvard Dataverse Network

Search this Dataverse Network [Advanced Search](#) [Tips](#)

The Harvard Dataverse Network is open to all scientific data from all disciplines worldwide. It includes the world's largest collection of social science research data. If you would like to upload your research data, first create a dataverse and then create a study. If you already have a dataverse, log in to add new studies. Learn more about the [Dataverse Network](#).

Dataverses

525 Dataverses

i A **Dataverse** is a container for research data studies, customized and managed by its owner.

RECENTLY RELEASED DATERVERSES

Thames, Frank	May 29, 2013
Mitts, Joshua	May 26, 2013
Damico, Anthony	May 22, 2013
Buntaine, Mark	May 22, 2013

Studies

51,966 Studies, **722,929** Files, **757,852** Downloads

i A **study** is a container for a research data set. It includes cataloging information, data files and complementary files.

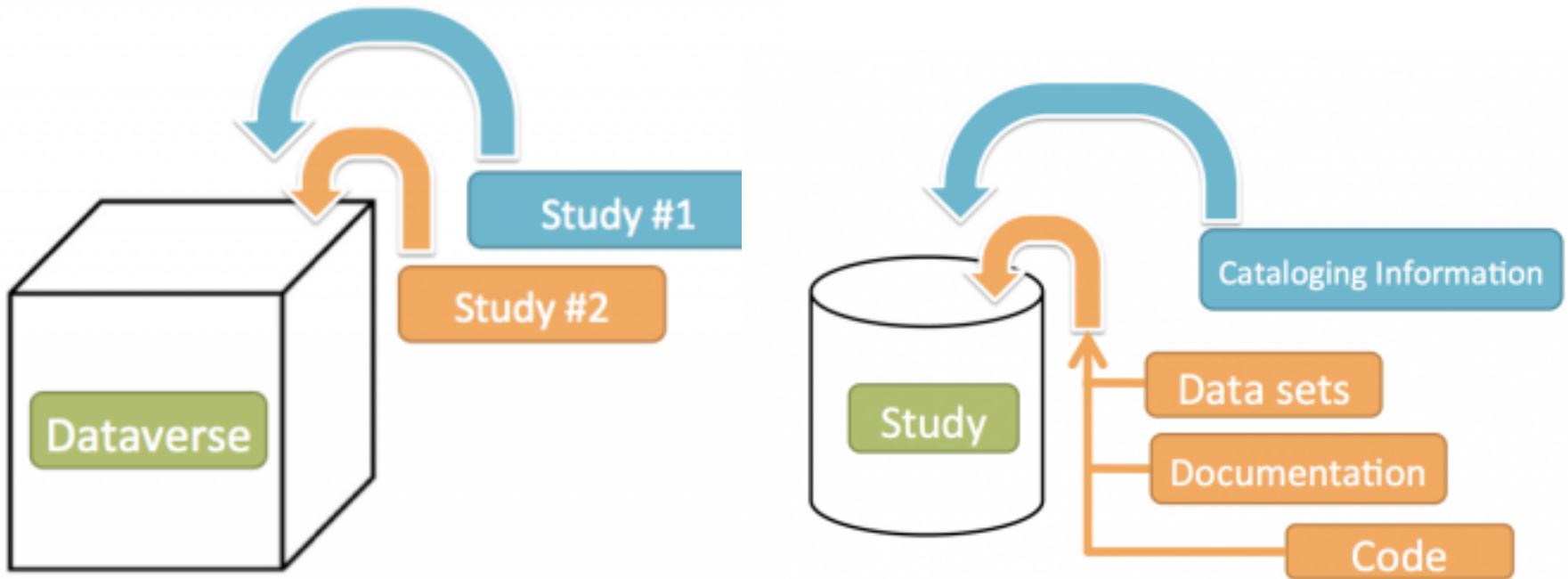
RECENTLY RELEASED STUDIES

Replication data for: Incentives for Personal Votes and Women's Representation in Legislatures by Thames, Frank; Williams, Margaret	May 29, 2013
Experimental Evidence in Electricity Behavior Research by Davis, Alex	May 29, 2013
Replication data for: "Long-Term Determinants of the Demographic Transition, 1870–2000" by Murtin, Fabrice	May 29, 2013

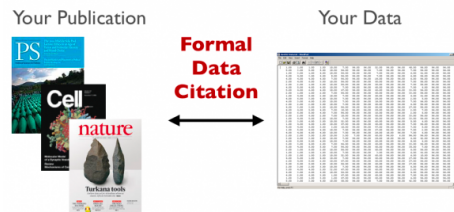
- The Harvard Dataverse Network is open to **all** research data from **all** domains (not exclusive to Harvard).
- The Dataverse Network software is **open-source** (in GitHub), installed in institutions across the world (<http://thedata.org>).

Dataverse: Container for your research studies

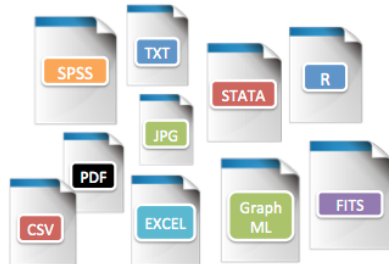
Study: Container for your data, documentation, and code



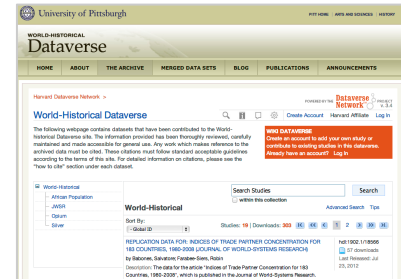
Data sharing and archiving with control and recognition for data authors, distributors



Persistent Data Citations
permanently linking your data to
Your publication



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



Customized Branding
or embed on your site



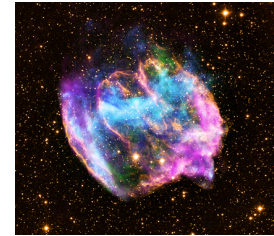
Data Restrictions
& terms of use options

Rich data support for some data formats



SPSS, Stata, R Data

metadata extraction, subsetting
& analysis (R, Zelig)



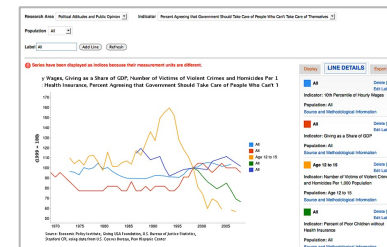
FITS Data

metadata extraction from file
header



Social Network Data (GraphML)

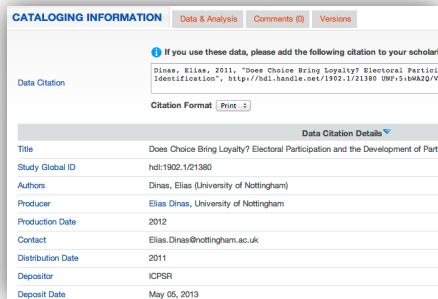
smart queries & subsetting



Data visualizations

for time series

Data management, standards and archival good practices



CATALOGING INFORMATION Data & Analysis Comments (3) Versions

Data Citation

Citation Format [Print](#)

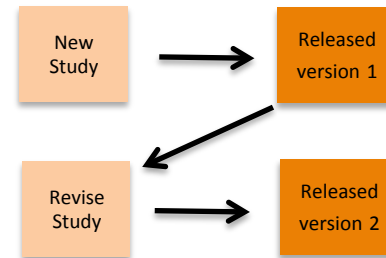
Data Citation Details	
Title	Does Choice Bring Loyalty? Electoral Participation and the Development of Party
Study Global ID	hdl:1902.1/21380
Authors	Dinas, Elias (University of Nottingham)
Producer	Elias Dinas, University of Nottingham
Production Date	2012
Contact	Elias.Dinas@nottingham.ac.uk
Distribution Date	2011
Depositor	ICPSR
Deposit Date	May 05, 2013

Data Cataloging

custom metadata templates for easy discovery (DDI, Dublin Core)

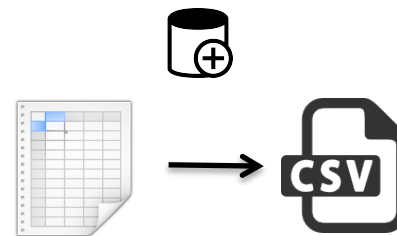


Log traffic & downloads to your dataset with Guestbook



Data Versioning

preserve & cite previous versions



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

Enabling Replication of published work

REPLICATION DATA FOR: WHAT TO DO ABOUT MISSING DATA IN TIME-SERIES CROSS-SECTIONAL DATA

[< View Previous Study Listing](#)

hdl:1902.1/14316UNF:5:RzZmkys+IaJKkDMAeQBObQ==

Version: 3 – Released: Mon Mar 14 16:45:20 EDT 2011

CATALOGING INFORMATION

[Data & Analysis](#)

[Comments \(0\)](#)

[Versions](#)

i If you use these data, please add the following citation to your scholarly references. [Why cite?](#)

Data Citation

```
James Honaker; Gary King, 2010, "Replication data for: What To Do about Missing Data in Time-Series Cross-Sectional Data", http://hdl.handle.net/1902.1/14316 UNF:5:RzZmkys+IaJKkDMAeQBObQ== IQSS Dataverse Network [Distributor] V3 [Version]
```

Citation Format

i Results found in this publication can be replicated using these data.










Original Publication

James Honaker and Gary King. 2010. "What To Do about Missing Data in Time-Series Cross-Sectional Data." *American Journal of Political Science* 54 (2): 561-81. [article available here](#)

Data Citation Details











Title	Replication data for: What To Do about Missing Data in Time-Series Cross-Sectional Data
Study Global ID	hdl:1902.1/14316
Authors	James Honaker (The Pennsylvania State University); Gary King (Harvard University)
Production Date	2010

Documentation, code and data

<input type="checkbox"/> 1. Documentation		
<input type="checkbox"/> Honaker&King.pdf Adobe PDF - 1 MB - 68 downloads	 Download	Published article
<input type="checkbox"/> Readme.txt Plain Text - 984 bytes - 62 downloads	 Download	Detailed information on the files in this study
<input type="checkbox"/> 2. Figure 6		
<input type="checkbox"/> F6 - aaReadme.txt Plain Text - 1 KB - 47 downloads	 Download	The files in this folder replicate the results presented in figure 6 comparing the coefficients of first differences of changes in trade dependence on the level of violence, from Burgoon (2006)
<input type="checkbox"/> F6 - analyzeburg.r Plain Text - 13 KB - 44 downloads	 Download	replicates the original Burgoon results, and then reruns these models with the imputed data. All quantities of interest can be found here. The code also creates the comparison of the... Continue (+)
<input type="checkbox"/> F6 - burgoonsubset.RData R Data - 338 KB - 43 downloads	 Download	raw data in an R readable format
<input type="checkbox"/> F6 - impburg.r Plain Text - 298 bytes - 43 downloads	 Download	reads in the data and runs the imputation in Amelia. Running this code will create one hundred imputed datasets named "p3ny\$\$.csv". Those used to create the graph in the paper are in... Continue (+)
<input type="checkbox"/> figure6.zipx application/octet-stream - 1 MB - 40 downloads	 Download	imputed datasets named "p3ny\$\$.csv"
<input type="checkbox"/> 3. Figure7a		
<input type="checkbox"/> F7a-aaReadme.txt Plain Text - 844 bytes - 37 downloads	 Download	The files in this folder replicate the first (uppermost) row of graphs in figure 7
<input type="checkbox"/> F7a - impburg.r Plain Text - 298 bytes - 38 downloads	 Download	reads in the data and runs the imputation in Amelia. Running this code will create one hundred imputed datasets named

Multiple formats, subsets and data analysis (Zelig)

5. Table 1

<input type="checkbox"/> baum.do Stata Syntax - 8 KB - 38 downloads	 Download	a Stata file that contains the commands to replicate the results presented in Table 1. It requires the h\$\$\$.csv datasets and generates the hnew\$\$\$.dta datasets. Results are displayed a... Continue [+]			
<input type="checkbox"/> baumlog.log Plain Text - 6 KB - 36 downloads	 Download	Results generated by baum.do			
<input type="checkbox"/> baummerge2.csv Plain Text - 526 KB - 40 downloads	 Download	the raw data for use in R			
<input type="checkbox"/> baummergenona.csv Plain Text - 511 KB - 35 downloads	 Download	the raw data in a slightly more convenient form for reading into Stata, with Stata standard missing values			
<input type="checkbox"/> hnew10_1.tab Tab Delimited - 878 KB - 44 downloads + analyses	 <div style="border: 1px solid gray; padding: 2px;"><input checked="" type="checkbox"/> Download as... Tab Delimited Saved original (Stata Binary) Splus (generated) R (generated)</div>  Access Analysis + Subsetting	copies of h1.csv through h10.csv saved in Stata format, with some processing of the variables, such as "tsset"-ing the dataset to set the cross-section and time-series internal ident... Continue [+]  View Data Citation [+]			
<table border="1" data-bbox="125 963 656 1006"><tr><td>TABULAR DATA</td><td>6492 Cases</td><td>22 Variables</td></tr></table>	TABULAR DATA	6492 Cases	22 Variables		
TABULAR DATA	6492 Cases	22 Variables			
<input type="checkbox"/> hnew1_1.tab Tab Delimited - 878 KB - 36 downloads + analyses	 Download as...	copies of h1.csv through h10.csv saved in Stata format, with some processing of the variables, such as "tsset"-ing the dataset to set the cross-section and time-series internal ident... Continue [+]			
<table border="1" data-bbox="125 1228 656 1270"><tr><td>TABULAR DATA</td><td>6492 Cases</td><td>22 Variables</td></tr></table>	TABULAR DATA	6492 Cases	22 Variables	 Access Analysis + Subsetting	 View Data Citation [+]
TABULAR DATA	6492 Cases	22 Variables			



Current projects (I)

- Integrate PKP's Open Journal System(OJS) with Dataverse:
 - Build a Plugin to submit data from OJS seamlessly to Dataverse
 - Establish persistent link between publication and data
 - Work with publishing workflows (data as part of submissions; data review and approval)
 - Pilot phase: 6 publishers, ~ 80 journals (total Journals using OJS: ~ 5,000)

<http://projects.iq.harvard.edu/ojs-dvn/>

Current projects (II)



- Extend data sharing to sensitive data:
 - Define (5) Privacy Tags to assess privacy risk and terms of use of a data set
 - Support secure transfer, storage and authentication for sensitive data
 - Integrate with privacy methods to access and analyze data

<http://privacytools.seas.harvard.edu/>

Current projects (III)



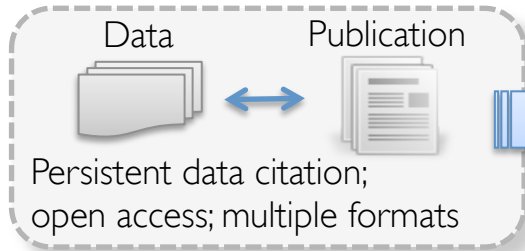
- Generate socio-metrics for data:
 - Connect data through metadata
 - Connect data through data usage
 - Connect data through methods

<http://databridge.web.unc.edu/>

Find, Share, Cite, Reuse, Reproduce Research

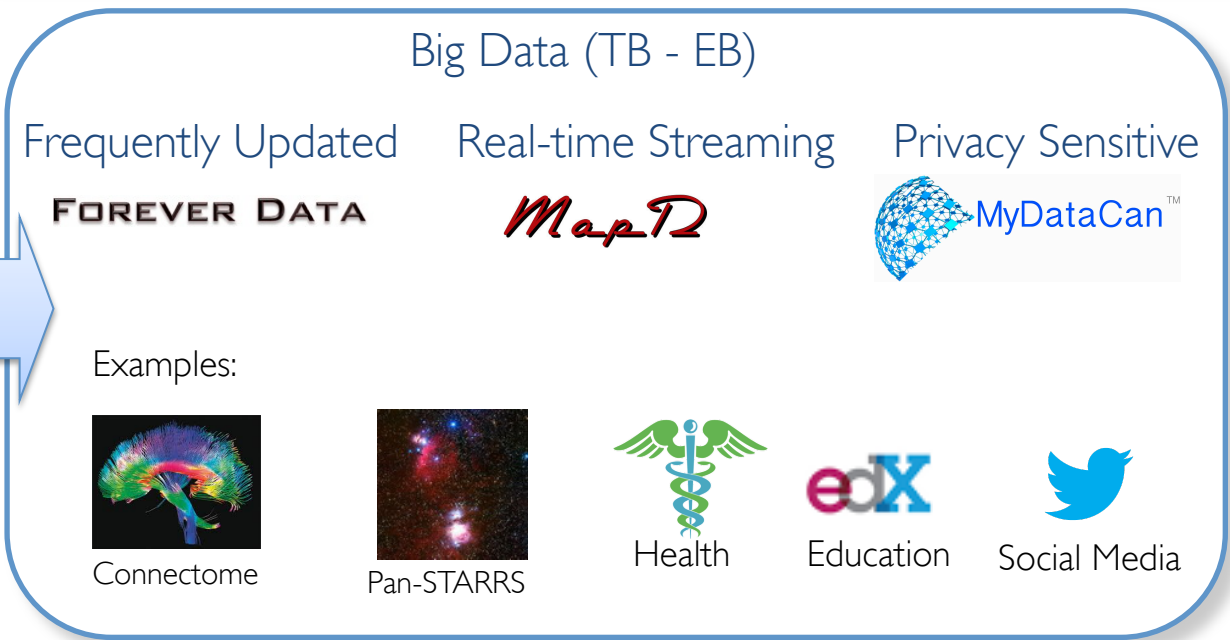
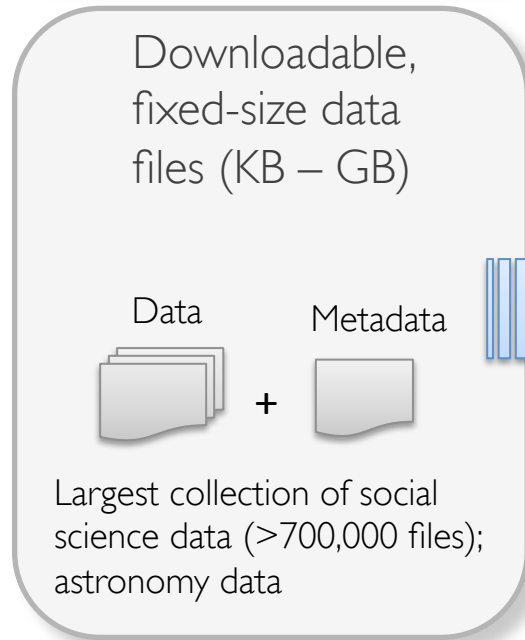
NOW

Large # of small data sets



COMING SOON

Small # of very large data sets



Formal Data (and Code) Citation in Published Work



the Future of Research Communications and e-Scholarship

[Contact](#) | [RSS Feed](#) | [Login](#) | [Join](#)



[About](#) | [Target Areas](#) | [Discussions](#) | [Tools and Resources](#) | [Publications](#) | [Blogs](#) | [Events](#) | [Members](#)

search this site



[Publications](#) >

Amsterdam Manifesto

Endorse

Comment

The Amsterdam Manifesto on Data Citation Principles

Preface:

We wish to promote best practices in data citation to facilitate access to data sets and to enable attribution and reward for those who publish data. Through formal data citation, the contributions to science by those that share their data will be recognized and potentially rewarded. To that end, we propose that:

1. Data should be considered citable products of research.
2. Such data should be held in persistent public repositories.
3. If a publication is based on data not included with the article, those data should be cited in the publication.
4. A data citation in a publication should resemble a bibliographic citation and be located in the publication's reference list.
5. Such a data citation should include a unique persistent identifier (a DataCite DOI recommended, or other persistent identifiers already in use within the community).
6. The identifier should resolve to a page that either provides direct access to the data or information concerning its accessibility. Ideally, that landing page should be machine-actionable to promote interoperability of the data.
7. If the data are available in different versions, the identifier should provide a method to access the previous or related versions.
8. Data citation should facilitate attribution of credit to all contributors