Mercè Crosas, Ph.D.
Twitter: @mercecrosas
Director of Data Science, IQSS
Harvard University
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

Customized Branding
or embed on your site

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

Data Restrictions
& terms of use options

Your Publication  \[\text{Formal Data Citation}\]  Your Data

Persistent Data Citations
permanently linking your data to
Your publication
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

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

Data Versioning
- preserve & cite previous versions

Log traffic & downloads
- to your dataset with Guestbook

Permanent storage
- preservation format with w/copies in multiple locations (OAI-PMH, LOCKSS)
Enabling Replication of published work
## Documentation, code and data

### 1. Documentation
- **Honaker&King.pdf**
  - Adobe PDF - 1 MB - 68 downloads
  - Published article
- **Readme.txt**
  - Plain Text - 984 bytes - 62 downloads
  - Detailed information on the files in this study

### 2. Figure 6
- **F6 - aaReadme.txt**
  - Plain Text - 1 KB - 47 downloads
  - 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)
- **F6 - analyzeburg.r**
  - Plain Text - 13 KB - 44 downloads
  - 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...
- **F6 - burgoonsubset.RData**
  - R Data - 338 KB - 43 downloads
  - raw data in an R readable format
- **F6 - impburg.r**
  - Plain Text - 298 bytes - 43 downloads
  - 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...

### 3. Figure 7a
- **F7a-aaReadme.txt**
  - Plain Text - 844 bytes - 37 downloads
  - The files in this folder replicate the first (uppermost) row of graphs in figure 7
- **F7a - impburg.r**
  - Plain Text - 298 bytes - 38 downloads
  - 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)
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/
NOW
Large # of small data sets

Downloadable, fixed-size data files (KB – GB)

Largest collection of social science data (>700,000 files); astronomy data

Persistent data citation; open access; multiple formats

Publication

Data

COMING SOON
Small # of very large data sets

Citation of a subset

Provenance

Real-time Explore

Combine Data

Crowdsourcing

Big Data (TB - EB)

Frequently Updated

Real-time Streaming

Privacy Sensitive

Examples:

Connectome

Pan-STARRS

Health

Education

Social Media

thedata.org

Crosas, Sweeney, Waldo, Seltzer, Goodman, Dominici, King, Pfister, Lewis Mostak, Knowles-Barley, Jones, Bar-Sinai

The Dataverse Network Project

Find, Share, Cite, Reuse, Reproduce Research

OPEN SCIENCE FOR BIG DATA
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.