The Dataverse Network
An open-source Application for Sharing, Discovering and Preserving Data

Mercè Crosas, Ph.D.
Director of Product Development
IQSS, Harvard University
Problems with traditional ways to store research data

Data sets

Text

Images

Keep it in my laptop

Hard to share, not safe

Send it to an archive

Author doesn’t get enough credit

Post it in my web site

No professional archiving services (backups, recovering, cataloging, preservation, etc)
Solving the problems through technology

We’ve developed a system that:

- Gives ownership and recognition to author, distributor and publisher
- Generates a persistent data citation
- Converts data sets to a preservable and verifiable format
- Distributes data to the public, but also supports restricted access
- Indexes all metadata for quick data discovery
- Supports subsetting and analysis for (some) data files
- Makes it easy to use and maintain
- Provides you legal protection
- Inter-operates with other systems using standards
Your web site

Your dataverse

Gives ownership and recognition

Quick Links

Find Data

Deposit Data

Dissertation Grants

Home

The Henry A. Murray Research Archive is Harvard’s endowed, permanent repository for quantitative and qualitative research data at the Institute for Quantitative Social Science, and provides physical storage for the entire IDSS Dataverse Network. Our collection comprises over 10 terabytes of data, audio, and video. We preserve in perpetuity all types of data of interest to the research community, including numerical, text, audio, interview notes, and other data. We accept data deposits through our web site, which is powered by our Dataverse Network software.

The Henry A. Murray Archive, through its endowment, supports permanent bit-level preservation of all social science research studies directly deposited in the IDSS Dataverse Network. For details, see the Murray Archive Preservation policies.
Credit to the author

Credit to the distributor
A persistent identifier and url which never changes (hdl:1902.2/7768):

Universal Numerical Fingerprint to verify dataset:

UNF:3:upor0mnBeXa9fujiUKPqNNAA==

Journal article, book

Persistent Citation for Data

Formal data citation

Monday, June 21, 2010
Convert Data to a Preservable and Verifiable Format

1. Extract variable name, description and summary statistics
2. Convert data set to a preservation format, independent of the software package
3. Apply a cryptographic algorithm to canonical format
4. Get alphanumeric string based on semantic contents of the digital object:
   - uniquely summarizes the contents,
   - but does not convey its information

UNF:5:EKgHvTNfkkS86dNzABIhNw==

Change content → Changes UNF
Change format → Doesn’t change UNF
Everybody can explore the cataloging information (metadata), download data and documentation files, and subset and analyze data if available.
Distributes to the public, but also supports restricted access.

... or some files can be restricted ...

- Only authorized users or groups can access the files
- Groups can be based on IP address
... or the entire study can be restricted.

Restricted studies can be searched but user cannot access the study page unless has the right permissions
Indexes all metadata for quick data discovery

Quick and advanced search of all metadata fields:

- Title
- Authors
- Distributors
- Abstract
- keywords
- Years
- Geographic information
- Data collection
- Terms of use
- ...
- Variable labels
- Variable descriptions

Uses **Lucene** index server as the search engine.
Supports subsetting and analysis for (some) data files

For SPSS and STATA Files
A Dataverse Network holds multiple dataverses

Each dataverse can be created in two clicks and customized as you want

Your dataverse may include your data or links to data from other dataverses

The Dataverse Network = the software

A dataverse = a site for publishing your data
Agreement for Data Deposit

✓ If human subjects were studied in the collection of the Materials, you collected the Materials with IRB approval;
✓ The Materials do not contain high-risk confidential information such as social security numbers; credit card numbers; medical record numbers; health plan numbers; other account numbers; certificate/license numbers; or biometric identifiers (fingerprints, retina, voice print, DNA etc.).

The terms of Use in the Harvard Dataverse Network are betted by lawyers from the Harvard’s Office of the General Counsel
You can customize the terms of use data deposit and data downloads in your dataverse or entire Dataverse Network
Inter-operates with other systems

- Study and variable metadata is exported into XML (Dublin Core, Data Descriptive Initiative - DDI, and others) and MARC.
- Supports OAI-PMH, z39.50, LOCKKS
- Plan to add RDF support in the future
New Functionality

- **Versioning**: Never lose previous versions of studies
- **Deaccessioning**: Never lose track of previous holding
- **Open access options**: Enable contributors to collaboratively edit
- **User comments**: Collect user comments about data
- **UNF 5.0**: Advanced semantic fingerprints for your data
- **Social Network Data**: Deposit, format, analyze, and preserve social networks
- **LOCKSS compatibility**: Automatic replication for preservation
Thanks!

Next...
We invite you to participate in a workshop on Data Publication and Citation Principles at IQSS, Harvard University
Organizers: Micah Altman and Mercè Crosas

Contact me at mcrosas@hmdc.harvard.edu if you are interested.

http://thedata.org (The project site)
http://dvn.iq.harvard.edu (The Dataverse Network at Harvard)

Acknowledgements:
Gary King, Micah Altman, Gustavo Durand, Ellen Kraffmiller, Kevin Condon, Bob Treacy, Leonid Andrev, Michael Heppler, Steve Kraffmiller