Dataverse: Research Transparency through Data Sharing

Mercè Crosas, Director of Data Science
Institute for Quantitative Social Science
Harvard University

Summer Institute
June 2014
Data sharing is good for science

Making your research data accessible is important:

• To reproduce research
• To make public assets available to the public
• To leverage investments in research data
• To advance research and innovation

Borgman, Oct 2013, “Why you should care about open data” Open Access Week Talk
10,555 studies that created gene expression microarray data:

• Studies that share data received 9% more citations

• Authors published most papers using their own data within 2 years

• Data reuse papers by third-party investigators continued for 6 years

But data sharing must include long-term accessibility.

Pepe, Goodman, Muench, Crosas, Erdmann, 2014 “Sharing, Archiving and Citing Data in Astronomy” Forthcoming
We hosted a workshop at Harvard University to address issues about data sharing and reuse, and the result was: “10 Rules”

Editorial

Ten Simple Rules for the Care and Feeding of Scientific Data

Alyssa Goodman¹, Alberto Pepe¹*, Alexander W. Blocker¹, Christine L. Borgman², Kyle Cranmer³, Merce Crosas¹, Rosanne Di Stefano¹, Yolanda Gil⁴, Paul Groth⁵, Margaret Hedstrom⁶, David W. Hogg³, Vinay Kashyap¹, Ashish Mahabal⁷, Aneta Siemiginowska¹, Aleksandra Slavkovic⁸

¹ Harvard University, Cambridge, Massachusetts, United States of America, ² University of California, Los Angeles, Los Angeles, California, United States of America, ³ New York University, New York, New York, United States of America, ⁴ University of Southern California, Los Angeles, Los Angeles, California, United States of America, ⁵ Vrije Universiteit Amsterdam, Amsterdam, The Netherlands, ⁶ University of Michigan, Ann Arbor, Michigan, United States of America, ⁷ California Institute of Technology, Pasadena, California, United States of America, ⁸ Pennsylvania State University, State College, Pennsylvania, United States of America
Rule 1: Love your data, and let others love them too

• If you make your data easily available to others, others are more likely to do the same—eventually.

• Or at least take solace in the fact that you'll be able to find and reuse your own data if you treat them well.
Rule 2: Share your data online, with a permanent identifier

- Your personal web site is unlikely to be a good option for long-term data storage.

- Publish your data in a general or a domain-specific data repository that guarantees long-term access, and assigns a persistent identifier to the data (DOI, HDL, PURL).
Rule 3: Conduct science with data reuse in mind

- The higher the quality of provenance information, the higher the chance of enabling data reuse.

- Keep: 1) data, 2) metadata, and 3) information about the process of generating those data, such as code.
Rule 4: Publish workflow as context

Publish a description of your processing steps to offer essential context for interpreting and re-using data.
Rule 5: Link your data to your publications as early as possible

- Many journals now offer standard ways to contribute data to their archives or trusted data repositories and link it to your paper.

- Use a formal data citation in the publication’s reference list.
Rule 6: Publish your code

Same best practices in relation to data and workflow also apply to software materials.
Rule 7: Say how you want to get credit for your data

- Simply describe your expectations on how you would like to be acknowledged.

- You can also release your data under a license, but making it simple for others to reuse it, when possible (Creative Commons, Open Data Commons, COS Open Data Badges).
Rule 8: Foster and use data repositories

Seek help from librarians, archivists or research communities on domain-based repositories and generic repositories available.
Rule 9: Reward colleagues who share their data properly

• Praise those following good practices.

• Follow good scientific practice and give credit to those whose data you use.
Rule 10: Help establish data science and data scientists as vital

- Advocate for hiring data specialists and for the overall support of institutional programs that improve data sharing.

- Teach whole courses, or mini-courses, related to caring for data and software, or incorporate the ideas into existing courses.
The Dataverse repository as a solution for data sharing

- The Dataverse hosted at Harvard is open and free to all researchers worldwide.
- Serves as a solution to help you follow the 10 Rules.
- Contains already > 53,000 data sets, the largest general-purpose data repositories in the world.
- The Dataverse open-source software is developed at Harvard’s IQSS, by our data science team plus contributors.
Find or publish data at: http://thedata.harvard.edu
Data Publishing Steps

1. Create a **dataverse**: your own virtual data repository
2. Add a **study** (or dataset): the data unit you want to publish
3. Enter study **metadata** (or cataloging fields)
4. Upload **Files**
5. **Release** when everything is ready
Benefits of publishing data with Dataverse

What you contribute

• Sufficient information accompanying the data
• Data files with rich metadata

What Dataverse gives you

• Credit for your data through data citation
• Control on how to share your data
• Data exploration and analysis for tabular data
• Long-term data preservation
Sufficient information with the data

The replication standard holds that:

Sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party could replicate the results without any additional information from the author.

“sufficient information”?

How were the respondents selected? Who did the interviewing? What was the question order? How did you decide which informants to interview or villages to visit? How long did you spend in each community? Did you speak to people in their language or through an interpreter? Which version of the ICPSR file did you extract information from? How knowledgeable were the coders? How frequently did the coders agree? Exactly what codes were originally generated and what were all the recodes performed? Precisely which measure of unemployment was used? What were the exact rules used for conducting the content analysis? When did the time series begin and end? What countries were included in your study and how were they chosen? What statistical procedures were used? What method of numerical optimization did you choose? Which computer program was used? How did you fill in or delete missing data?

Consider using the following files for tabular data sets:

- **R Data**: R is open-source, with a growing community
- **SPSS, STATA**: Also commonly used in social sciences
- Add full variable metadata
- Indicate properly missing data
Dataverse generates a **data citation** with a persistent identifier, which you and others can use to reference your data set in an article or book.
Importance of data citation

Dataverse data citation is compliant with the Joint Declaration of Data Citation Principles, which states that:

Sound, reproducible scholarship rests upon a foundation of robust, accessible data. For this to be so in practice as well as theory, data must be accorded due importance in the practice of scholarship and in the enduring scholarly record. In other words, data should be considered legitimate, citable products of research. Data citation, like the citation of other evidence and sources, is good research practice and is part of the scholarly ecosystem supporting data reuse.

To learn more and endorse the principles: https://www.force11.org/datacitation
Dataverse processes tabular data files and provides summary statistics and access to data analysis.
Data Analysis with Zelig

Dataverse integrates with Zelig:

- Zelig is an R package that provides a common interface to a large set of statistical models
- It is also developed at Harvard’s IQSS, by our data science team plus contributors
- An enhanced version (Zelig 5) will be available this summer
- More information at:
  
  http://datascience.iq.harvard.edu/zelig

June 2014

BITSS Summer Institute
Additional Dataverse Features

Dataverse also allows you to:

• Link your data set to the original publication(s)
• Publish multiple versions of your datasets
• Set terms of use for your data
• Restrict data files, while metadata and documentation can be kept public (but we encourage open data, when possible)
• Brand your dataverse banner with your logo, image or colors
• Track downloads for your data, and enable a guestbook
• List data sets from other dataverses in your dataverse
Dataverse 4.0 comes this summer with a full new user interface and many new features!

To test our Beta version and give us feedback: http://dataverse-demo.iq.harvard.edu
Dataverse 4.0 will include a new interactive data exploration and analysis tool, TwoRavens, which integrates with Zelig.
Learn more about upcoming research tools at: http://datascience.iq.harvard.edu

THANKS – mcrosas@iq.harvard.edu Twitter:@mercecrosas