JESLIB 2012; 1(3): 173-179
doi:10.7191/jeslib.2012.1020

Journal of eScience Librarianship
putting the pieces together: theory and practice

A Data Sharing Story

Mercè Crosas

Institute for Quantitative Social Science, Harvard University, Cambridge, MA, USA

Abstract

From the early days of modern science through this century of Big Data, data sharing has enabled some of the greatest advances in science. In the digital age, technology can facilitate more effective and efficient data sharing and preservation practices, and provide incentives for making data easily accessible among researchers. At the Institute for Quantitative Social Science at Harvard University, we have developed an open-source software to share, cite, preserve, discover, and analyze data, named the Dataverse Network. We share here the project's motivation, its growth and successes, and likely evolution.

The Beginning

Since the early days of modern science, the process of creating knowledge has been accelerated by reusing data from a new perspective. In the late sixteenth century, Tycho Brahe performed the most accurate measurements of Mars' motion before the invention of the telescope. In the last days of his life, Brahe appointed Johannes Kepler to analyze the data expecting this to prove his proposed Tychonic system, a hybrid between the Ptolemaic and Copernican systems. After Brahe's death, Kepler's analysis showed instead that the planets had an elliptical orbit about the sun, and deduced the three laws of planetary motion (Frautschi et al. 1986; Caspar 1959). Moving forward more than three centuries, Rosalind Franklin created high-resolution X-ray data of DNA, and proposed that DNA likely had a helical structure. It was not until Watson and Crick gained access to her data, that it was proven that the 3-D structure of DNA was an anti-parallel double helix (Watson 1980).

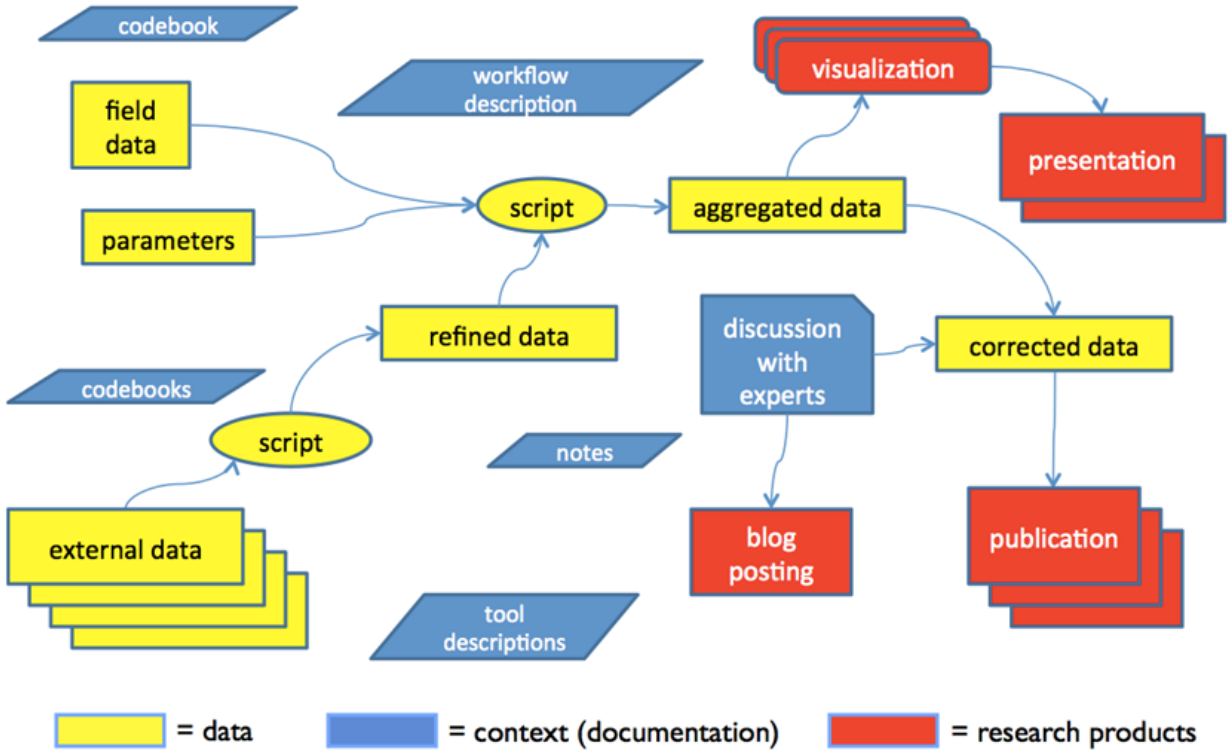
Unfortunately, in these examples, and others in the history of science, the data were not strictly "shared." In fact, Brahe kept most of the data away from Kepler during his life, but Kepler took it away after Brahe died. Brahe's consolatory last words to Kepler were, "Let me not seem to have lived in vain." In the quest for the structure of DNA, Wilkins, a colleague of Franklin's, showed the X-ray data to Watson and Crick without Franklin's knowledge or consent.

In the twenty-first century, we can, must, and already do better (http://www.ncbi.nlm.nih.gov/genbank, http://thedata.org, http://datadryad.org, http://www.pangaea.de). The digital age has brought a deluge of research data which must be archived so it is not lost, and it must be shared in order to be reused. Two important standards define the intellectual beginning of data sharing in the current age: replication and data citation.

Correspondence to Mercè Crosas: mcrosas@iq.harvard.edu
Keywords: data sharing, data management, preservation, data citation, Dataverse Network

JESLIB 2012; 1(3): 173-179
doi:10.7191/jeslib.2012.1020

Figure 1: Both the data workflow and documentation are necessary to fully understand and reproduce the results claimed in the research products (acknowledgment: Andrea Goethals).

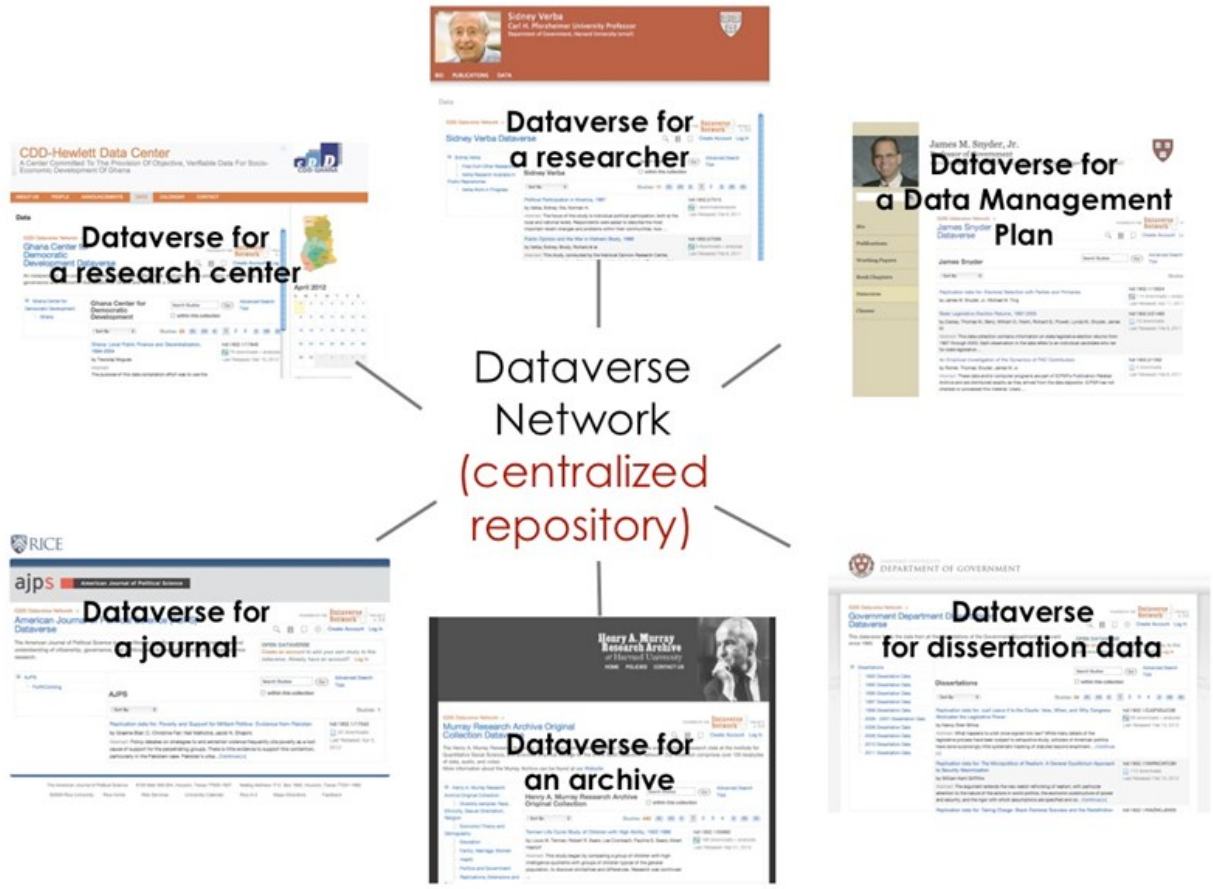


1) Replication:
In defining a necessary standard for replication (King 1995), King asserts that "sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party can replicate the results without any additional information from the author." Although this may be a difficult standard to achieve due to undocumented assumptions and processing idiosyncrasies, it is important to make achieving this standard a goal. Without an adequate description, the original data can rapidly become meaningless. Description of the data and their analysis may include metadata about the study and its data sets, scripts or code used in the analysis, documentation, and additional supplementary files. Even when all supporting files are provided, the original results might still be difficult to replicate depending on the data and the type of study – for example, conditions might be impossible to reproduce, software might be obsolete, or formats might have changed.

2) Data Citation:
A formal data citation is also necessary to provide proper data sharing. It establishes a persistent reference from the scholarly work to the underlying data used to develop any new claims. Altman and King (Altman and King, 2007) propose a data citation standard which includes attribution to the authors and distributors, the year the data are available, the title of the data set or study, and more importantly, a persistent URL (e.g., http://handle.net/ or http://www.doi.org/), and a Universal Numerical Fingerprint (UNF). The UNF is a hash applied to the contents of the data set, independent of the file format (Altman, Gill, and McDonald 2003).

JESLIB 2012; 1(3): 173-179
doi:10.7191/jeslib.2012.1020

Figure 2: A Dataverse Network offers a centralized data repository. Each dataverse hosted is owned and customized by the data owner or provider.



Employing these two standards, a third party can reuse the data that build scientific knowledge, thus reexamining and validating, but most importantly, advancing science.

A Solution

Once we understand the fundamental elements of data sharing, we then need to facilitate its practice. Two distinct goals that conflict with each other must be reconciled and consolidated – namely, author ownership and data preservation.

The Data Author Perspective

The data author wants to keep ownership of and gain recognition for his data (King 2007). A centralized data repository, however, must ensure long term preservation and high-quality archival practices to allow reusing the data in 50 or 100 years from now (http://www.crl.edu/archiving-preservation/digital-archives). At the Institute for Quantitative Social Sciences (IQSS) at Harvard University, we have built the Dataverse Network (http://thedata.org), an open-source software to share, cite, preserve, discover, and analyze research data (King 2007; Crosas 2011). A Dataverse Network consists of multiple dataverses, where each dataverse is a virtual archive storing data generated by a researcher, or replication data for a journal, or the archive of an entire research project, or all the data created by the dissertations of an institution.

Each dataverse contains studies, which can be organized in collections of related studies, and each study contains the metadata or cataloging fields as well as the data files and supporting files (documentation, codebooks, code, auxiliary files).

A data author creates a dataverse and within that dataverse creates a study with a research data set and cataloging information that describes the data. The cataloging fields follow a metadata template based on the Data Documentation Initiative (DDI) schema (http://www.ddialliance.org/). Some of the fields in the metadata template, specifically those related to data collection approaches and methodologies, can be customized for each discipline or domain.

Once the study that holds the data is released to others, the author gets a data citation generated by the Dataverse Network. This data citation follows the standard proposed by Altman and King 2007, including a Handle as the persistent identifier (in the form of "http://hdl.handle.net/prefix/ID"), a UNF, and a version number. The Handle is automatically registered in the Handle.net service (http://handle.net), and it resolves to the study in the Dataverse Network repository. The current UNF Version 5 (http://thedata.org/book/unf-version-5) only applies to tabular data sets (Altman 2008). In addition, these quantitative tabular data sets are automatically converted to preservation formats. This means that commonly used statistical formats, such as SPSS and STATA files, are reformatted to tab delimited data files; and those files can then be downloaded in multiple formats, including the original one. These data files are further processed to facilitate downloads of subsets and statistical analysis through the Zelig statistical software (http://projects.iq.harvard.edu/zelig).

The data author can customize his dataverse or embed the dataverse in his own web site. He maintains ownership of his data by setting the appropriate access rules to the data files - terms of use, public or restricted access - by conducting updates, and by releasing new study versions.

The Archivist Perspective

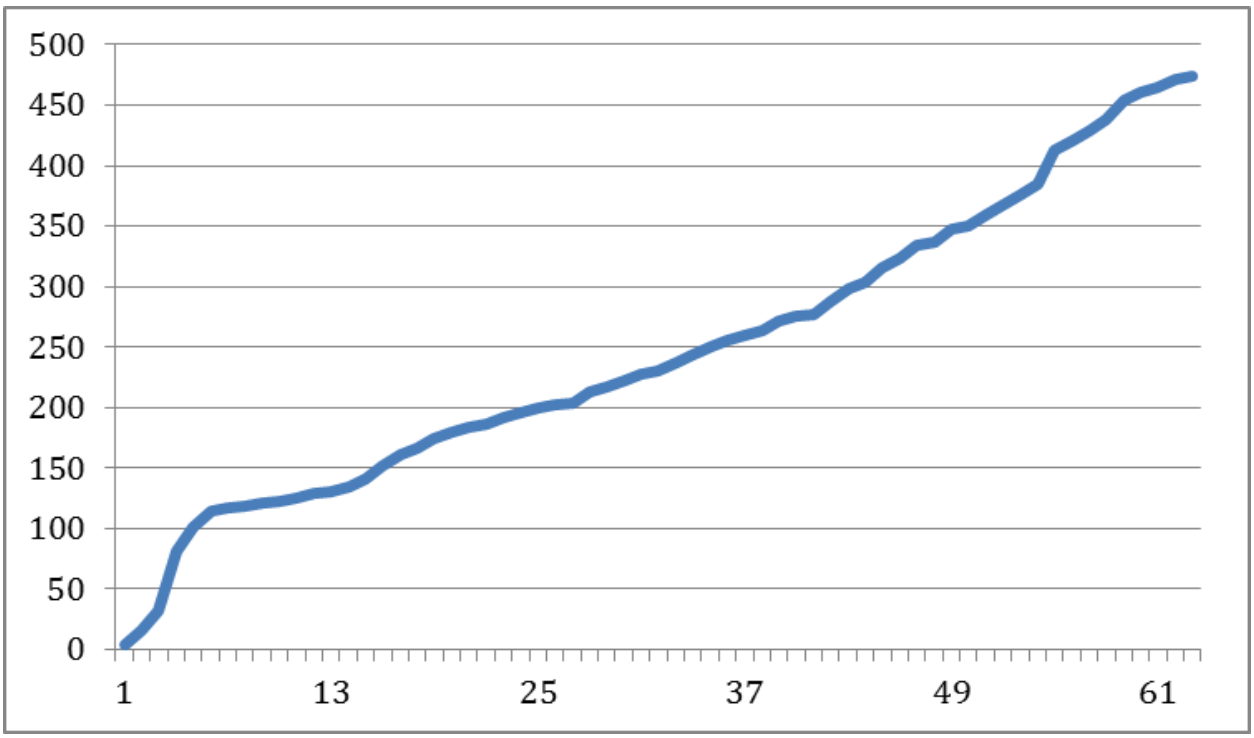
Behind the scenes, and without being a barrier for the author, the centralized repository provides the necessary archival support in a scalable configuration: 1) it keeps redundant copies of the data in multiple locations to ensure greater data safety using an open-source replication system, LOCKSS (Lots of Copies Keep Stuff Safe) (Reich and Rosenthal 2001); 2) it converts the data sets to an archival format which does not depend on a specific software package so the data can be easily converted to other formats; 3) it exports the cataloging information that describes the data into standard metadata formats (such as Data Documentation Initiative and Dublin Core) for preservation and discoverability; and 4) it inter-operates with other systems through standard metadata harvest protocols (such as the Open Archives Initiative protocol OAI-PMH) and through REST-style web services to search and get data from other web applications.

The Dataverse Network software is open-source under the Apache 2 license and can be installed for free by any institution (http://guides.thedata.org). It is a multi-tier Java application that requires setting up an application server (Glassfish), a database server to store the metadata (PostgreSQL), a file system for the data files, and a server to run the statistical analysis (RServe).

Success Stories from Social Science

In the six years since the Dataverse Network was born, we have observed a sharp growth in the amount of data archived and shared at the IQSS Dataverse Network for social science data hosted at Harvard University. There are now more than 450 dataverses shared with the world, hosting a total of 50,000 studies and 700,000 files. Of these dataverses, 98 were released in 2011, and

Figure 3: Number of "released" (shared) dataverses each month (i.e., dataverses published and searchable which contained public or restricted data) versus time (in months), from 2007 until October 2012 (acknowledgment: Gustavo Durand).



86 have already been shared in the first half of 2012.

Of the 50,000 studies, about 40,000 are harvested from external Dataverse Network installations and other repositories, and the other 10,000 have been deposited directly into the IQSS Dataverse Network. More than half of the studies are shared only among approved requestors or researchers from an authorized institution, while the rest are completely public, without any access restrictions (specifically, 40% of the 40,000 harvested studies are public and 60% of the 10,000 directly deposited studies are public).

We describe here a few of the many success stories from this journey that illustrate different usages and purposes of a dataverse.

An early success story is the Harvard Election Data Archive (http://projects.iq.harvard.edu/eda/data), a project led by Stephen Ansolabehere, a political science professor at Harvard University, and Jonathan Rodden, a political science professor at Stanford University, to help share and improve election data. Within a year of its existence, and using a dataverse, the project has collected and improved the usability of election data sets from 43 states in common, shareable formats. There have been more than 5,000 downloads to access these sources. Prior to this, state-level data and metadata were either printed on paper or distributed across state-specific databases of limited accessibility. The dataverse has enabled this research team to easily allow contributors from an array of institutions to participate in improving the usability of the data sets - assembling original data, transforming and standardizing them, enhancing the metadata, and constantly updating the archive.

A second story is the dataverse for James

Snyder, a professor at the Government Department of Harvard University (http://scholar.harvard.edu/jsnyder/data). Snyder has used his dataverse as a solution to the Data Management Plan required of most grant recipients by the National Science Foundation. An NSF Data Management Plan must describe the methods by which the anticipated data will be stored for long-term access, what formats will be supported, what metadata standards will be provided, how the data will be backed up, how others will access the data, and how the data will be secured (National Science Foundation 2011). Being compliant with a data management and preservation plan is often a challenge for an individual researcher, but a necessary requirement of many grant applications. Depositing and sharing the data through a dataverse has provided Snyder a ready-made and reusable solution that fulfills all Data Management requirements.

Finally, a third successful story is the dataverse for the Review of Economics and Statistics (http://dvn.iq.harvard.edu/dvn/dv/restat), a peer-reviewed scientific journal edited by Harvard University's Kennedy School of Government and published by MIT Press. The journal uses its dataverse to share replication data for each article published in the journal. Each publication can then be validated which adds credibility to the research and results of the scholarly work, and the simple presence of replication data often leads to more citation and notice of the original work (Piwowar, Day, and Fridsma 2007). In the two years since the dataverse for this Review was released, authors have deposited the data associated with their article, have obtained a standardized data citation to reference the data sets from the original publication, and have gained direct recognition for producing or enhancing the primary data.

An Expanding Approach to Sharing

Data sharing with the Dataverse Network continues to expand, strengthening the connection between scholarly work and data, and supporting additional scientific disciplines.

In particular, in a project funded by the Alfred P. Sloan Foundation, we are integrating the Dataverse Network with the Open Journal System, an open-source software for the management of academic journals created by the Public Knowledge Project (Willinsky 2005). This integration will allow authors to deposit data seamlessly to a dataverse when they submit an article. The publication will become automatically linked to the data set using the persistent data citation generated by the dataverse. This project will not only contribute to establishing a formal linkage between scholarly publications and the underlying research data, but will also build awareness of the importance of data sharing to enable the maximum advance of science.

We have also launched a Dataverse Network for Astronomy data, through collaboration between IQSS, the Harvard Library, and the Harvard-Smithsonian Center for Astrophysics. As part of this effort, the Dataverse software was made sufficiently flexible and extensible to support data from additional scientific disciplines. It now provides custom metadata fields through metadata templates, which can be tailored for each discipline, and allows deposit of any data file format. Further development will enable automatic metadata extraction and re-formatting of discipline-specific data types beyond those in social science (e.g., FITS data format in Astronomy), support massive, complex data files, and aid the exploration and visualization of vast data sets.

Conclusion

Interdisciplinary collaboration has proven significant to both successful technology development and the overall evolution of knowledge in the age of Big Data. Collaborations with research groups outside the social sciences, with journals, and with libraries, while still in the early-stage of our ef-

forts, support further rapid technology development in the interest of data sharing within and across a broader range of domains. Working with journals offers an opportunity to develop application programming interfaces (API) to deposit data from publishers' sites to an archival data repository; working with the library offers an opportunity to develop additional preservation services; and working with other disciplines such as astronomy offers an opportunity to expand preservation, visualization, and analysis for large image files. In social sciences, with an average of 70 data downloads per study so far, the original data are already starting to be reused and we should soon see the impacts in validating studies and advancing knowledge.

References

Altman, Micah. "A Fingerprint Method for Verification of Scientific Data." Presentation at the *International Conference on Systems, Computing, Sciences and Software Engineering*, 2007. http://thedata.org/publications/fingerprint-method-verification-scientific-data.

Altman, Micah and Gary King. "A proposed Standard for the Scholarly Citation of Quantitative Data." *D-Lib Magazine* 13, no. 3/4 (2007), http://dx.doi.org/10.1045/march2007-altman

Altman, Micah, Jeff Gill, and Michael McDonald. *Numerical Issues In Statistical Computing for the Social Scientist.* New York: Wiley-Interscience, 2003.

Caspar, Max. *Kepler.* London: Abelard Schuman, 1959.

Crosas, Mercè. The Dataverse Network: An Open-Source Application for Sharing Discovering and Preserving Data. *D-Lib Magazine,* 17, no. 1/2 (2011), http://dx.doi.org/10.1045/january2011-crosas

Frautschi, Steven, Richard Olenick, Tom Apostol, and David Goodstein. *The Mechanical Universe: Mechanics and Heat.* Cambridge University Press, 1986.

King, Gary. "An Introduction to the Dataverse Network as an Infrastructure for Data Sharing." *Sociological Methods and Research 36 (2007)*: 173-199.

King, Gary. "Replication, Replication." *PS: Political Science and Politics* 28 (1995): 443-449, http://dx.doi.org/10.1177/0049124107306660

NSF Data Management Plan Requirements," National Science Foundation, accessed November 8, 2012, http://www.nsf.gov/eng/general/dmp.jsp.

Piwowar, Heather A., Roger S. Day, and Douglas B. Fridsma. "Sharing Research Data is Associated With Increased Citation Rate." *PLoS ONE* 2, no. 3 (2007): e308, http://dx.doi.org/10.1371/journal.pone.0000308

Reich, Vicky and David Rosenthal. "LOCKSS: A Permanent Web Publishing and Access System." *D-Lib Magazine* 7, no. 6 (2001), http://dx.doi.org/10.1045/june2001-reich

Watson, James. *The Double Helix: A Personal Account of the Discovery of the Structure of DNA.* New York: Atheneum, 1968.

Willinsky, John. "Open Journal Systems: An Example of Open Source Software for Journal Management and Publishing." *Library Hi-Tech* 23, no. 4 (2005): 504-519, http://dx.doi.org/10.1108/07378830510636300

Disclosure: The author reports no conflicts of interest.

ISSN 2161-3974