A Data Sharing Story

Mercè Crosas, Ph.D.
Director of Product Development
Institute for Quantitative Social Science (IQSS)
Harvard University
How it all started

Success stories from social science

Where we are heading
Sharing of Data Enables the Maximum Advance of Science

In the Early days of modern science, Tycho Brahe’s data on Mars allowed Kepler to define the laws of planetary motion.

In the 20th century, Rosalind Franklin’s X-Ray data of DNA allowed Watson and Crick to discover the double helix model of the structure of DNA.

(Unfortunately, in these cases the data were “stolen” rather than “shared” - we can do better!)
"The replication standard holds 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."

- King, Gary. 1995 “Replication, Replication”
- Altman, Micah, King, Gary. 2007 “A Proposed Standard for the Scholarly Citation of Quantitative Data”
A Basic Principle

Scholarly Work + Data + Metadata +
Supporting Files
(documentation, code)

Formal Data Citation:
Authors, Year, Title, **Persistent Identifier (handle)**, **Universal Numerical Fingerprint (UNF)**, Distributor, Version, [+ Optional Fields]

== A third party can replicate and reuse, thus validate and advance science
What You Need to Make it Work

A repository for research data that takes care of long term preservation and good archival practices, while the researcher keeps control of and gets recognition for his data.

Researcher

- Deposits data and enters metadata
- Gets data citation
- Displays data on own web site
- Manages access control
- Updates new versions

Centralized Data Repository

- Backups and replication of data in different locations (LOCKSS)
- Re-format for long term accessibility
- Extract Metadata from data sets
- Metadata standards (DDI, Dublin Core)
- Inter-operability (OAI, APIs)
How it all started

Success stories from social science

Where we are heading
A project to share and improve election data:
In less than one year, 43 states, ~ 5000 downloads.
A Data Management Plan for NSF
Replication Data: when authors submit a paper, they deposit the data in the journal’s Dataverse.
How it all started

Success stories from social science

Where we are heading
Reinforcing the Link between Scholarly Work and Research Data

Alfred Sloan Foundation grant

Seamless integration between the two systems

Open Journal System

The Dataverse Network
Working with the Harvard Library to offer Dataverse as a Data Management Plan

- Provide services on research data management:
  - Data preservation and curation
  - Using the Dataverse
- Centralized hosting for Dataverse Networks in multiple domains
The AstroData Project

- Started as a pilot project at the Harvard-Smithsonian Center for Astrophysics, and it was launched on Monday!
- Does not substitute observatory-specific repositories
- Encourages Dataverse expansion:
  - Support Metadata for multiple domains (v 3.0)
  - Deposit larger data files
  - Explore and visualize large Images
  - Extract metadata from FITS files
Thanks
The Dataverse Network Project: thedata.org
Social Science: dvn.iq.harvard.edu
Astronomy: theastrodata.org
mcrosas@iq.harvard.edu