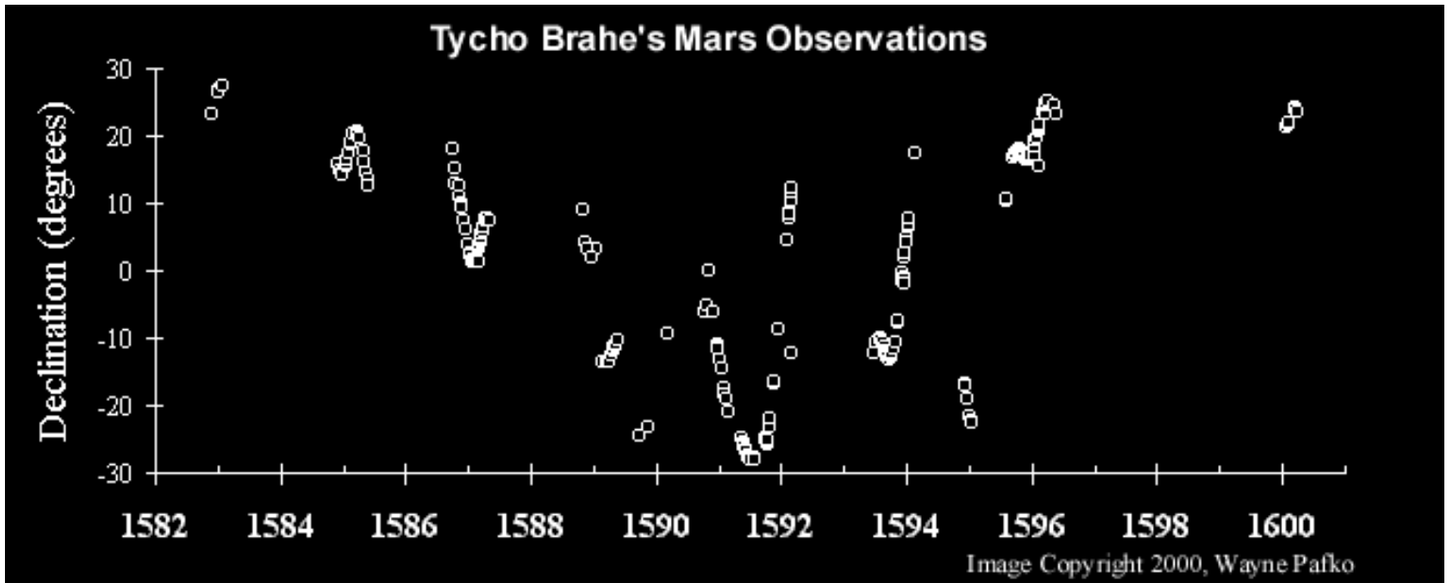
The Dataverse Network

Mercè Crosas, Ph.D.
Director of Product Development

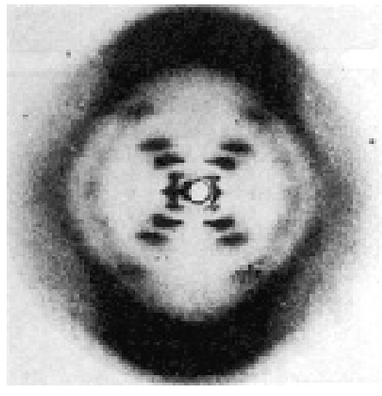
IQSS
The Institute for Quantitative Social Science
HARVARD UNIVERSITY

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.



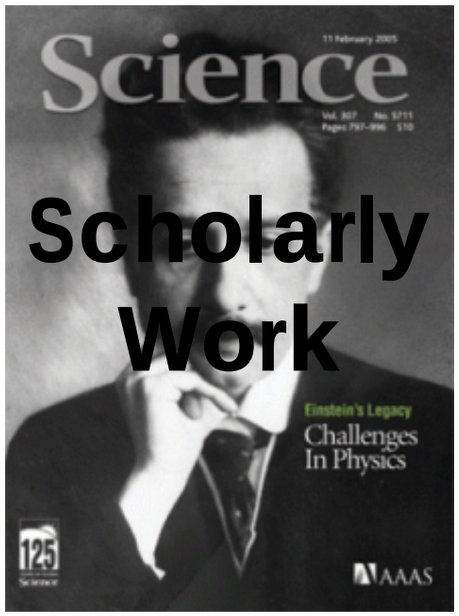
(Unfortunately, in these cases the data were "stolen" rather than "shared" - we can do better!)

The Intellectual Origin of the Dataverse Network

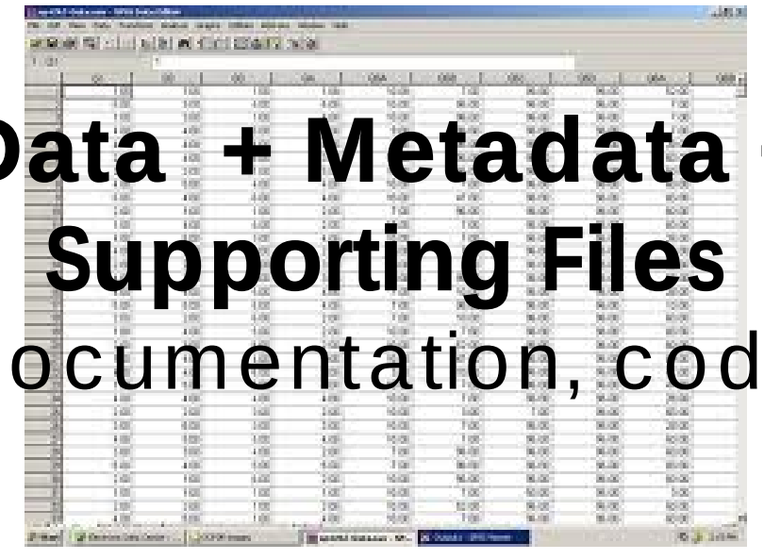
"The replication standard holds that sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party can replicate the results without any additional information from the author."

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

A Basic Principle



+



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, enhance and advance science

What Do You Need to Make it Work?

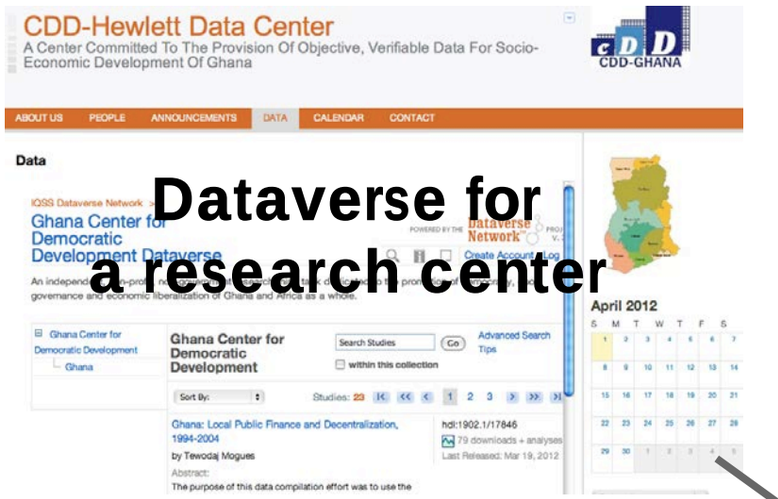
A repository for research data that takes care of long term preservation and good archival practices, while the researcher keeps control of and gets recognition for his data

Researcher

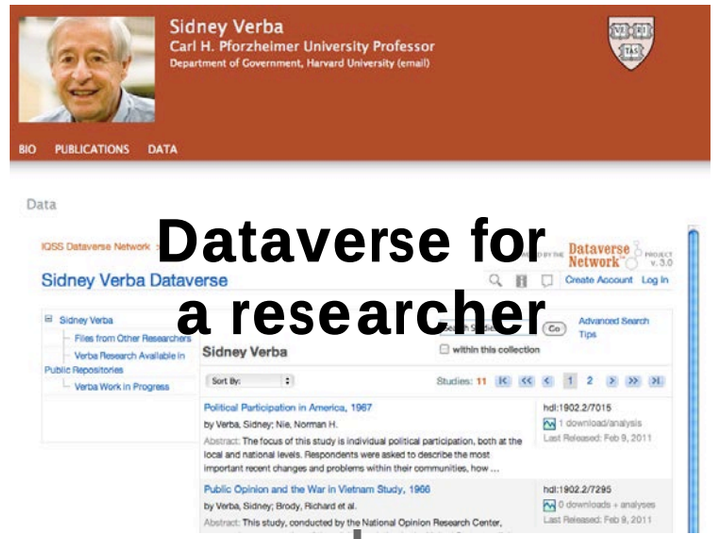
- ✓ Deposits data and enters metadata
- ✓ Gets data citation (handle, UNF)
- ✓ Displays data on own web site
- ✓ Manages data permissions
- ✓ Updates new versions

Centralized Data Repository

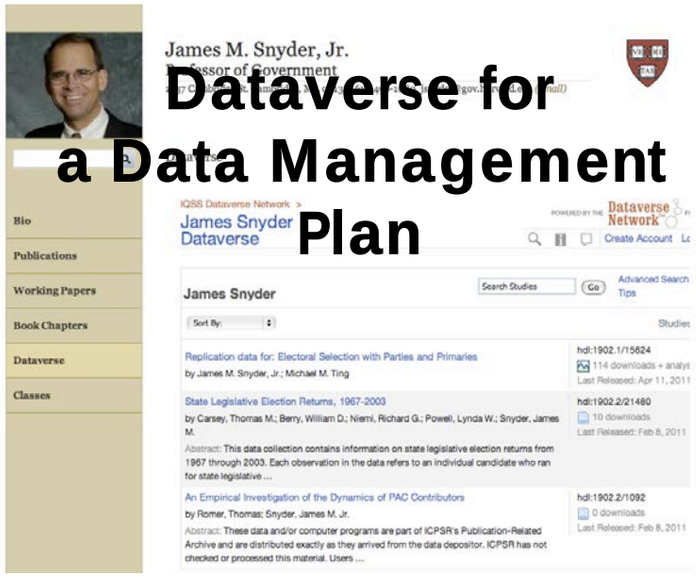
- ✓ Backups and replication of data in different locations (LOCKSS)
- ✓ Conversion to archival formats
- ✓ Extraction of Metadata from data sets
- ✓ Metadata standards (DDI, Dublin Core)
- ✓ Inter-operability (OAI, APIs)



Dataverse for a research center



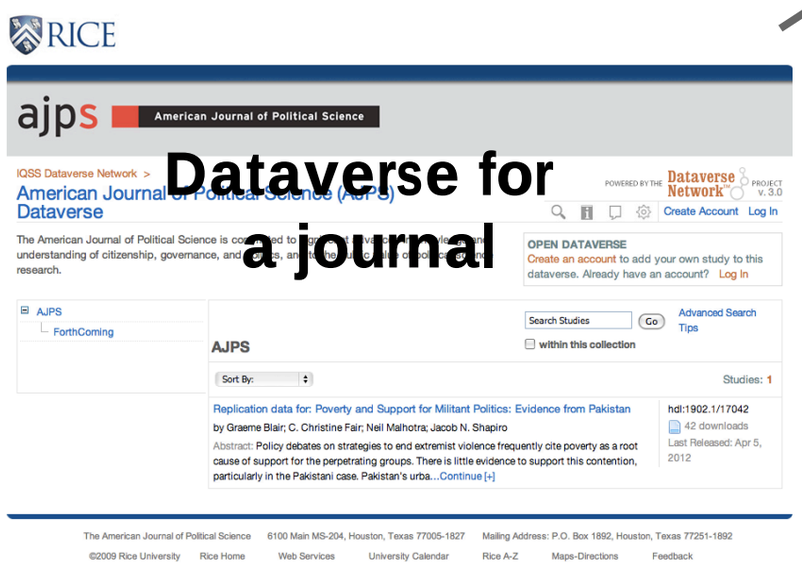
Dataverse for a researcher



Dataverse for a Data Management Plan

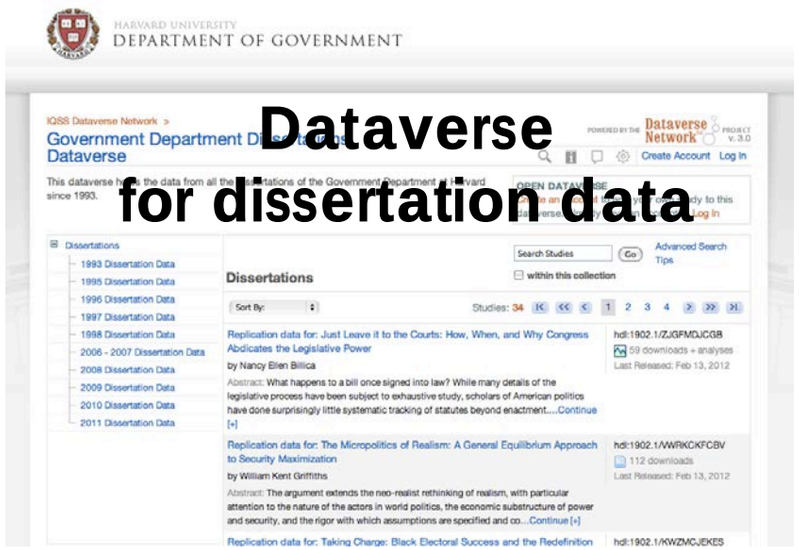
Dataverse Network

(centralized repository)

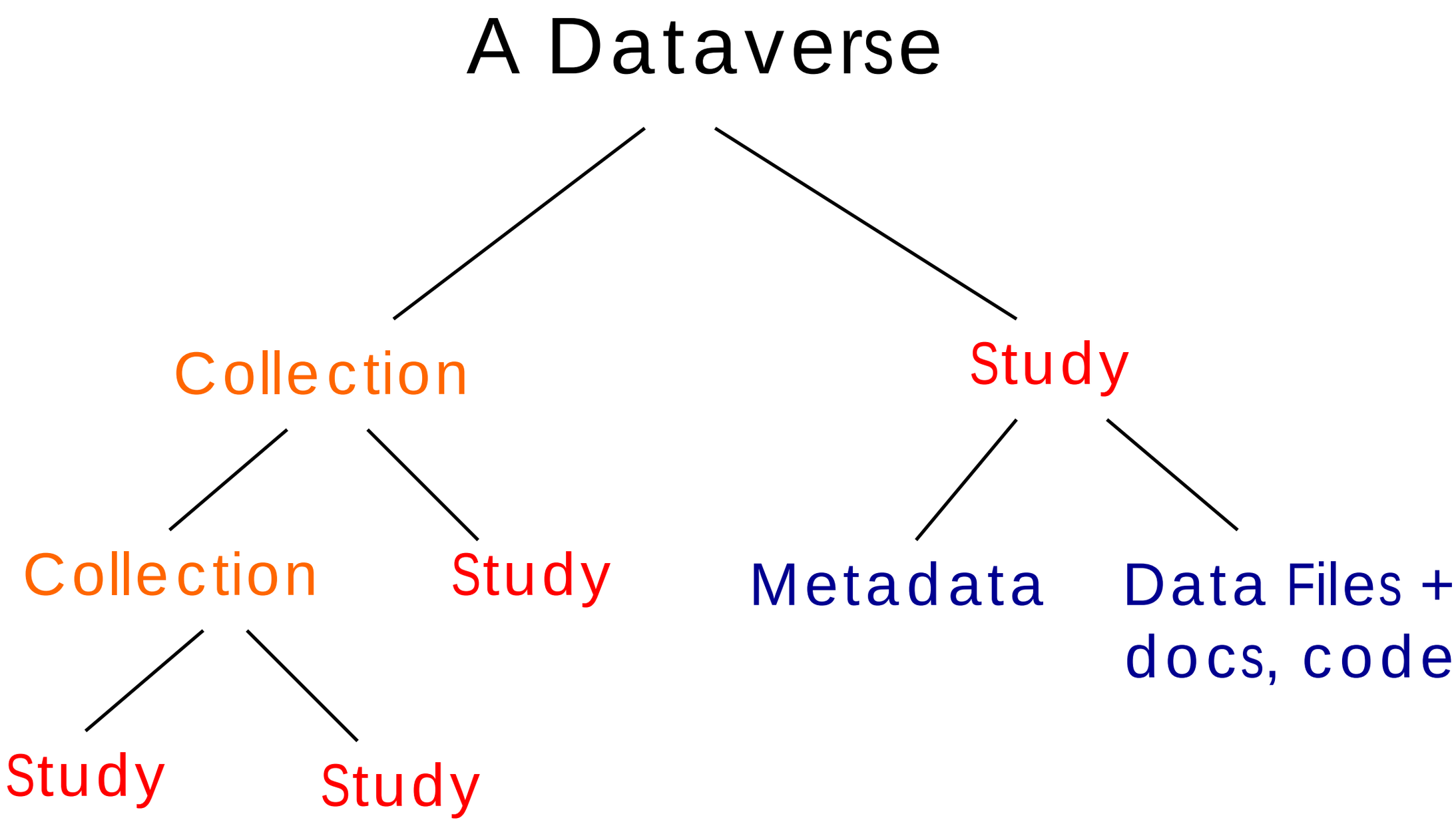


Dataverse for a journal

Dataverse for an archive



Dataverse for dissertation data


A Dataverse
Collection
Study
Collection
Study
Study
Study
Metadata
Data Files +
docs, code

Metadata

- Based on DDI 2.0:
  - Data Citation
  - Description and Scope
  - Data Collection and Methodology
  - Data Availability
  - Terms of Use

Data Files

Tabular Data:
STATA
SPSS
CSV + control card
Tab delimited + DDI

Social Network Data:
GraphML

+ complementary files
in **any** format

- ✓ Convert to preservation format (primary data + variable metadata)
- ✓ Calculate Universal Numerical Fingerprint (UNF)
- ✓ Define Terms of Use and Permissions
- ✓ Download in multiple formats
- ✓ Download a subset of the data
- ✓ Run R statistical models (Zelig)
- ✓ Visualize time series

Metadata Templates

- Templates allow:
  - Setting recommended, required and hidden fields
  - Pre-populating fields with default values
  - Adding controlled vocabularies
  - Adding custom fields under Data Collection and Methodology *(to support domains outside social science, such as astronomy and biomedical)*

Standards and Preservation

- Metadata is exported into DDI, Dublin Core and MARC formats for preservation and exchange protocols
- OAI protocol for harvesting metadata
- RESTful API for searching and getting metadata
- LOCKSS for replicating metadata and data in multiple locations

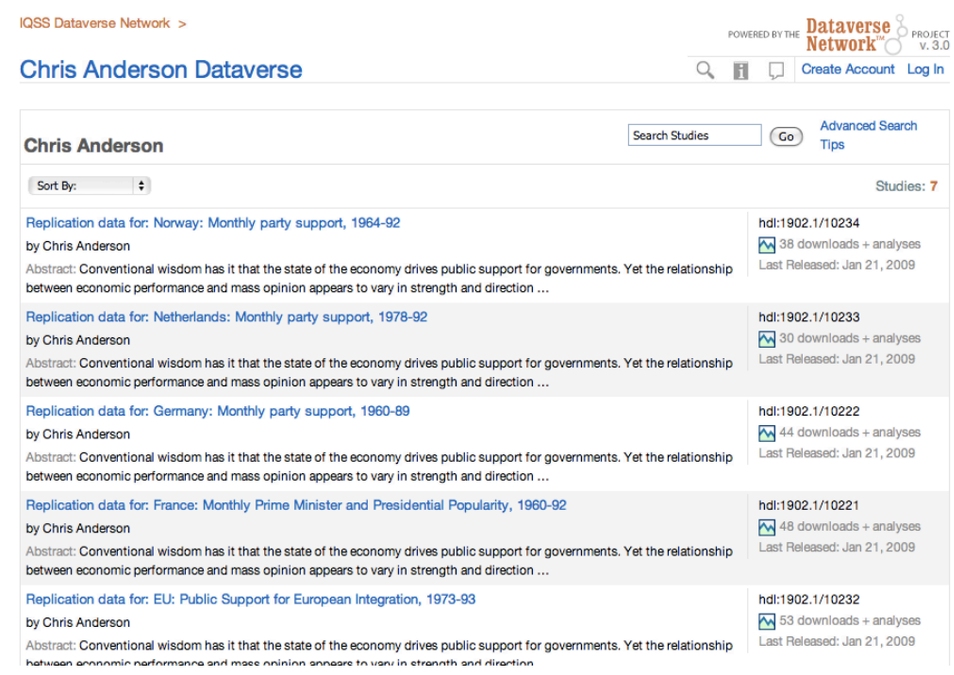
Upcoming Work

Reinforcing the Link between Scholarly Work and Research Data

(funded by Alfred P. Sloan Foundation grant)

Open Journal System

Seamless integration between the two systems



The Dataverse Network

Thanks

For more information:
The Dataverse Network Project:
http://thedata.org

The IQSS Dataverse Network
dvn.iq.harvard.edu
(Open to all social science research data, in collaboration with Harvard Library)

mcrosas@iq.harvard.edu