The Data Management Plan with Dataverse

Mercè Crosas, Ph.D.
Director of Product Development

IQSS
The Institute for Quantitative Social Science
HARVARD UNIVERSITY

The Dataverse

The Data Management Plan

The Data Management Plan with Dataverse

The Dataverse

The Data Management Plan

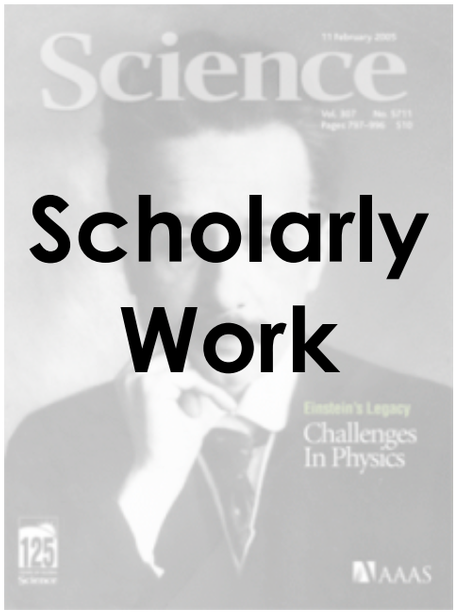
The Data Management Plan with Dataverse

The Intellectual Origin of the Dataverse Network

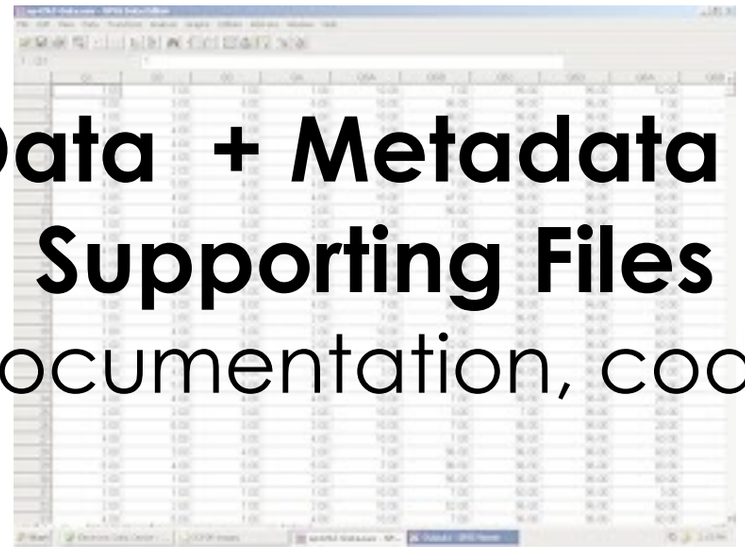
"The replication standard holds that sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party can replicate the results without any additional information from the author."

- 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 You Need to Make it Work

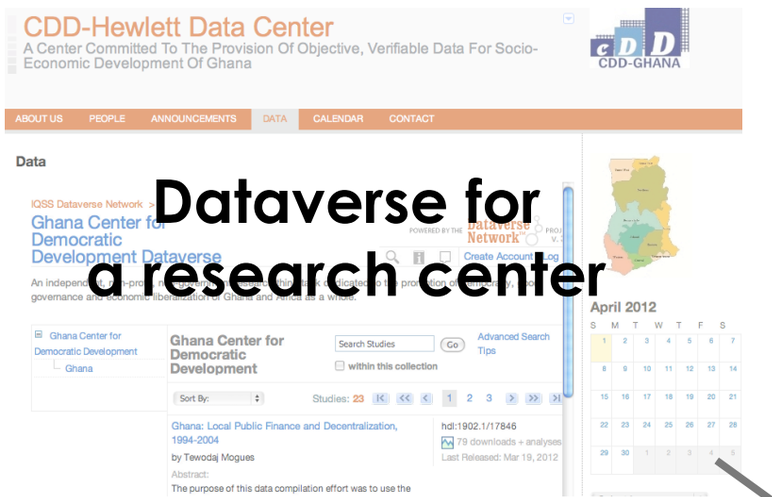
A repository for research data that takes care of long term preservation and good archival practices, while the researcher keeps control of and gets recognition for his data

Researcher

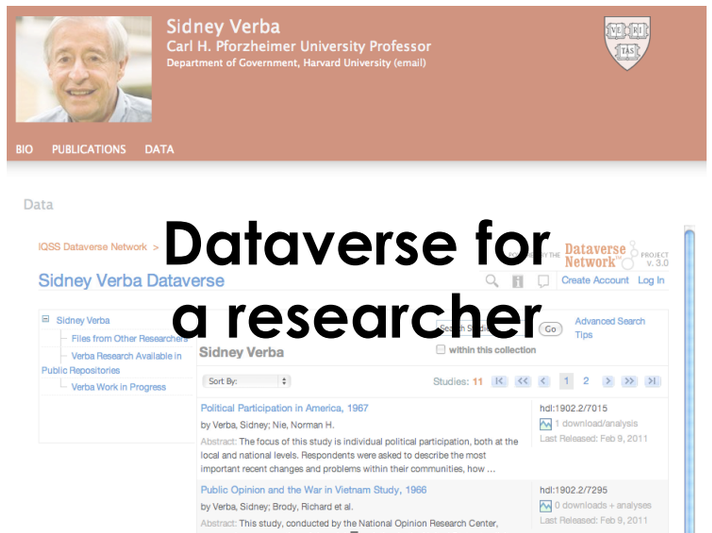
- ✓ Deposits data and enters metadata
- ✓ Gets data citation (handle, UNF)
- ✓ Displays data on own web site
- ✓ Manages data permissions
- ✓ Updates new versions

Centralized Data Repository

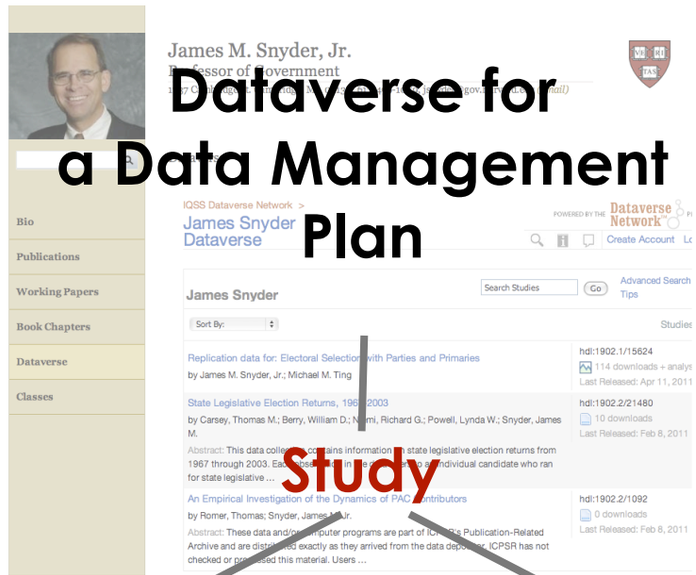
- ✓ Backups and replication of data in different locations (LOCKSS)
- ✓ Conversion to archival formats
- ✓ Extraction of Metadata from data sets
- ✓ Metadata standards (DDI, Dublin Core)
- ✓ Inter-operability (OAI, APIs)



Dataverse for a research center



Dataverse for a researcher



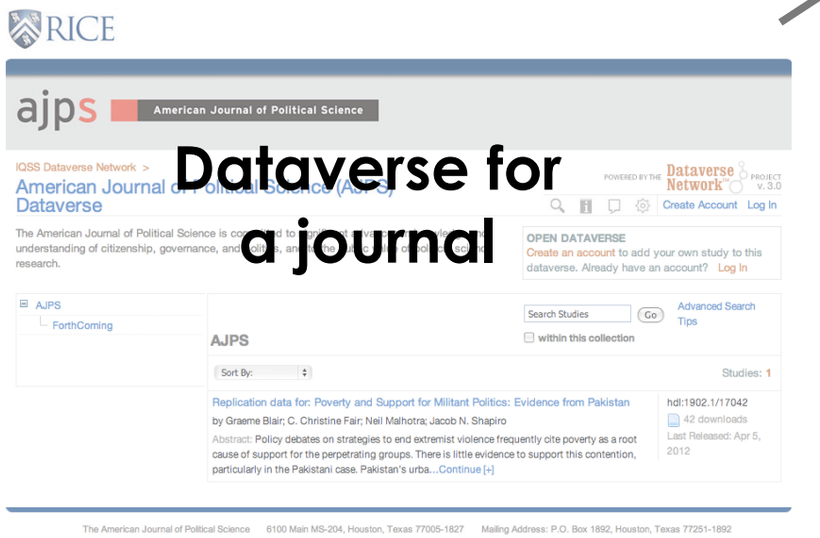
Dataverse for a Data Management Plan

Dataverse Network (centralized repository)

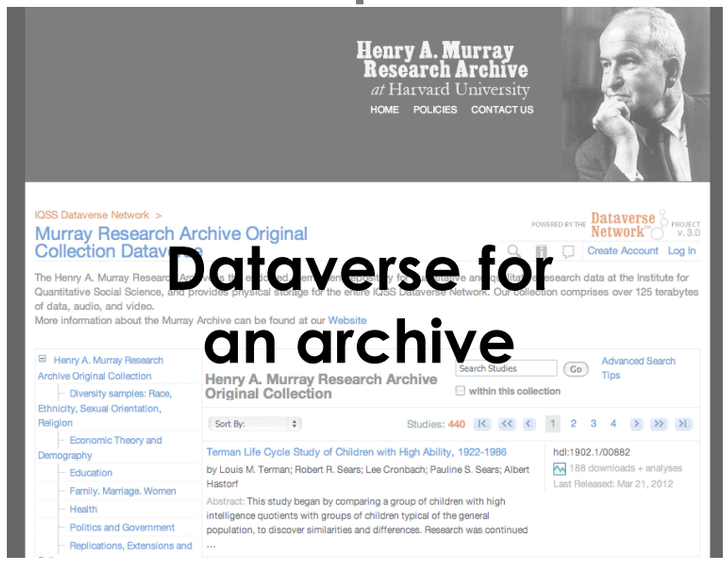
Study

Metadata

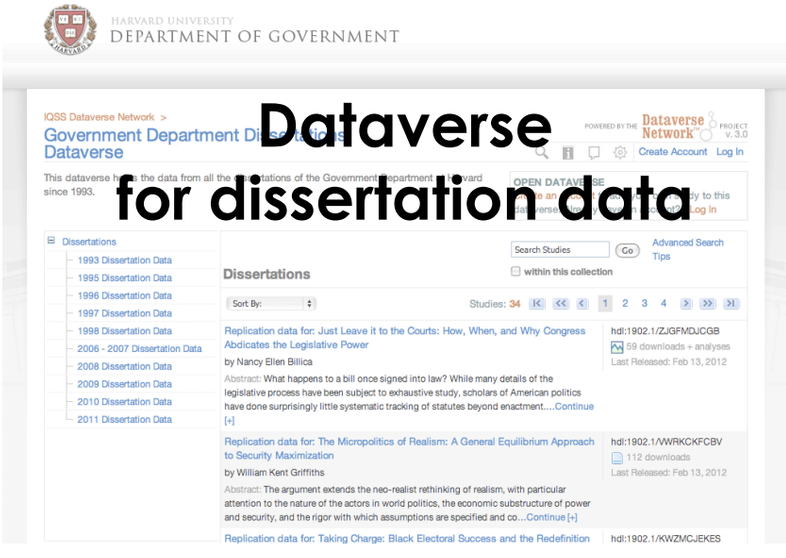
Data Files



Dataverse for a journal



Dataverse for an archive



Dataverse for dissertation data

The Dataverse

The Data Management Plan

The Data Management Plan with Dataverse

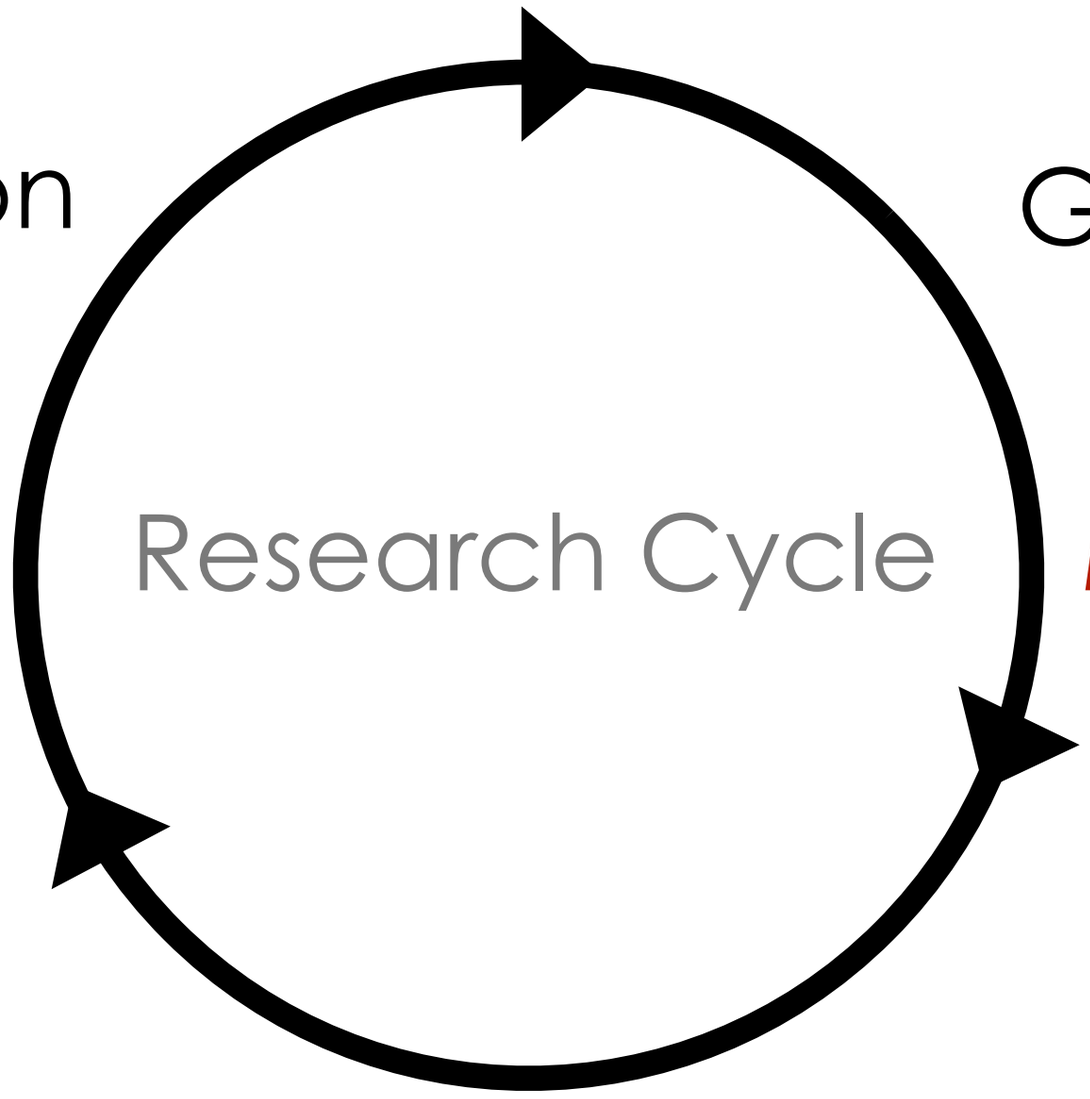
Project Idea
Grant Proposal

Data
Management
Plan!

Data
Collection
and Analysis

Publication

Data
Sharing
and
Archiving



Use Case: James Snyder's NSF proposal

Bio

Class

Publications

Working Papers

Book Chapters

Dataverse

James M. Snyder, Jr.

Professor of Government

1737 Cambridge St. Cambridge, MA 02138, 617-496-1089, jsnyder@gov.harvard.edu *(email)*

Professor Snyder's primary research and teaching interests are in American politics, with a focus on political representation. He has written on a variety of topics, including elections, campaign finance, legislative behavior and institutions, interest groups, direct democracy, the media, and corruption. He is a Research Associate at the National Bureau of Economic Research, and a Fellow of the American Academy of Arts and Sciences.

His articles have appeared in the American Political Science Review, the American Journal of Political Science, the Journal of Politics, the American Economic Review, the Journal of Political Economy, Econometrica, and many other other journals and edited volumes. He is co-author of *The End of Inequality: One Person, One Vote and the Transformation of American Politics*. Professor Snyder taught for six years in the Department of Economics at the University of Chicago, and for eighteen years in the Departments of Political Science and Economics at the Massachusetts Institute of Technology.

Admin

Checklist for generic NSF Data Management Plan

Data Description	What data will be generated? How will you create the data? (simulated, observed, experimental, software, physical collections)
Existing Data	Will you be using existing data? Relationship between the data you are collecting and existing data.
Audience	Who will potentially use the data?
Access and Sharing	How will data files be shared? How will others access them?
Formats	What data formats will you be creating?
Metadata and Documentation	What documentation will you provide to describe the data? Metadata formats and standards?
Storage, backup, replication, versioning	Are the data files backed up regularly? Are there replicas in different locations? Are older versions of the data kept?
Security	Are the system and storage that will be used secure?
Budget	Any costs for preparing the data? Costs for storage and long-term access?
Privacy, Intellectual Property	Does the data contain private or confidential information? Any copyrights?
Archiving, Preservation, Long-term Access	What plans do you have to archive the data and other research products? Will it have long-term accessibility?
Adherence	How will you check for adherence of this plan?

Data Description	**What data will be generated? How will you create the data? (simulated, observed, experimental, software, physical collections)**
Existing Data	**Will you be using existing data? Relationship between the data you are collecting and existing data.**
Audience	**Who will potentially use the data?**
Access and Sharing	How will the data be shared? How will others access them?
Formats	What data formats will you be creating?
Metadata and Documentation	What documentation will you provide to describe the data? Metadata formats and standards?
Storage, backup, replication, versioning	Are the data files backed up regularly? Are there replicas in different locations? Are older versions of the data kept?
Security	Are the system and storage that will be used secure?
Budget	Any costs for preparing the data? Costs for storage and long-term access?
Privacy, Intellectual Property	Does the data contain private or confidential information? Any copyrights?
Archiving, Preservation, Long-term Access	What plans do you have to archive the data and other research products? Will it have long-term accessibility?
Adherence	How will you check for adherence of this plan?

The Political Economy of U.S. State Courts:

The Influence of Media and Selection Systems

James M. Snyder, Jr., Harvard University

Claire S. H. Lim, *Stanford University*

Grant Proposal

Data Collected by this project will include:

1) collection of detailed national scale data set on press coverage of the U.S. state courts,

2) collection of data on election of judges in trial, appellate, and supreme courts in all states, and

3) construction of data set on county composition of judicial districts of trial courts in all states.

Data Description	What data will be generated? How will you create the data? (simulated, observed, experimental, software, physical collections)
Existing Data	Will you be using existing data? Relationship between the data you are collecting and existing data.
Audience	Who will potentially use the data?
Access and Sharing	**How will data files be shared? How will others access them?**
Formats	What data formats will you be creating?
Metadata and Documentation	What documentation will you provide to describe the data? Metadata formats and standards?
Storage, backup, replication, versioning	Are the data files backed up regularly? Are there replicas in different locations? Are older versions of the data kept?
Security	Are the system and storage that will be used secure?
Budget	Any costs for preparing the data? Costs for storage and long-term access?
Privacy, Intellectual Property	Does the data contain private or confidential information? Any copyrights?
Archiving, Preservation, Long-term Access	What plans do you have to archive the data and other research products? Will it have long-term accessibility?
Adherence	How will you check for adherence of this plan?

Bio

Publications

James M. Snyder, Jr.

Professor of Government

1737 Cambridge St. Cambridge, MA 02138, 617-496-1089, jsnyder@gov.harvard.edu *(email)*

IQSS Dataverse Network >

James Snyder Dataverse

POWERED BY THE Dataverse Network™

Create Account Lo

James Snyder

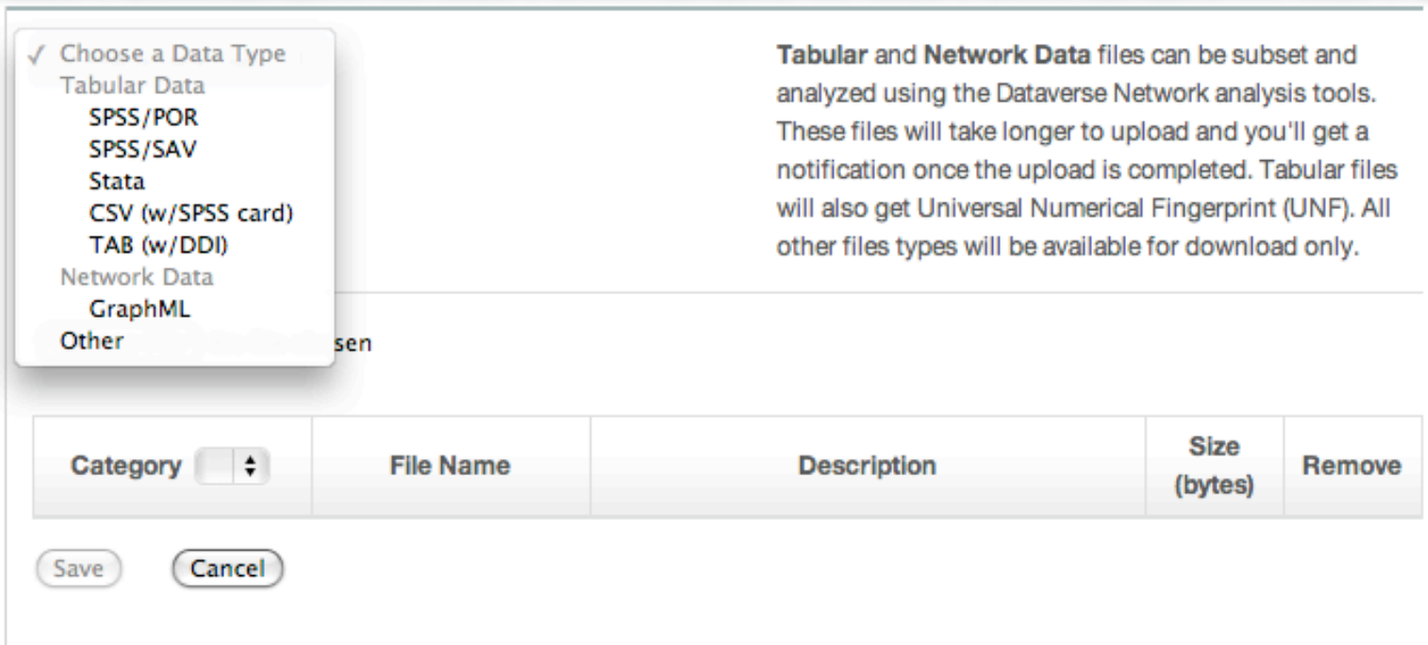
Search Studies Go Advanced Search Tips

- All data collected or generated will be deposited in the researcher's Dataverse.
- The Dataverse allows researchers to deposit data in an organized, well curated and citable network… ultimately facilitating access and sharing.

Data Description	What data will be generated? How will you create the data? (simulated, observed, experimental, software, physical collections)
Existing Data	Will you be using existing data? Relationship between the data you are collecting and existing data.
Audience	Who will potentially use the data?
Access and Sharing	How will data files be shared? How will others access them?
Formats	**What data formats will you be creating?**
Metadata and Documentation	What documentation will you provide to describe the data? Metadata formats and standards?
Storage, backup, replication, versioning	Are the data files backed up regularly? Are there replicas in different locations? Are older versions of the data kept?
Security	Are the system and storage that will be used secure?
Budget	Any costs for preparing the data? Costs for storage and long-term access?
Privacy, Intellectual Property	Does the data contain private or confidential information? Any copyrights?
Archiving, Preservation, Long-term Access	What plans do you have to archive the data and other research products? Will it have long-term accessibility?
Adherence	How will you check for adherence of this plan?

James M. Snyder, Jr.
Professor of Government
1737 Cambridge St. Cambridge, MA 02138, 617-496-1089, jsnyder@gov.harvard.edu *(email)*

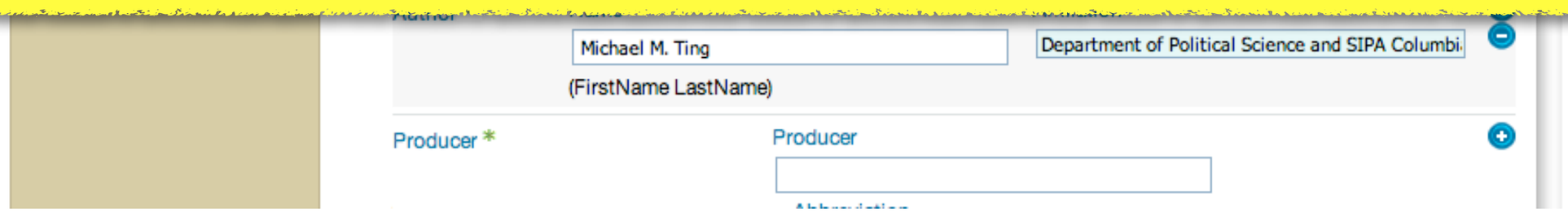
- Quantitative data will be deposited either in SPSS, Stata, CSV, Tab delimited, or GraphML.
- Images in JPEG-2000 or TIFF.[good practice]
- Audio in MP3 or WAVE.[good practice]
- Dataverse accepts all data formats to accommodate the flexibility researchers need.



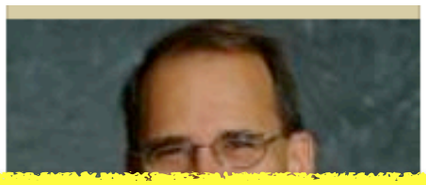
Data Description	What data will be generated? How will you create the data? (simulated, observed, experimental, software, physical collections)
Existing Data	Will you be using existing data? Relationship between the data you are collecting and existing data.
Audience	Who will potentially use the data?
Access and Sharing	How will data files be shared? How will others access them?
Formats	What data formats will you be creating?
Metadata and Documentation	**What documentation will you provide to describe the data? Metadata formats and standards?**
Storage, backup, replication, versioning	Are the data files backed up regularly? Are there replicas in different locations? Are older versions of the data kept?
Security	Are the system and storage that will be used secure?
Budget	Any costs for preparing the data? Costs for storage and long-term access?
Privacy, Intellectual Property	Does the data contain private or confidential information? Any copyrights?
Archiving, Preservation, Long-term Access	What plans do you have to archive the data and other research products? Will it have long-term accessibility?
Adherence	How will you check for adherence of this plan?

James M. Snyder, Jr.
Professor of Government
1737 Cambridge St. Cambridge, MA 02138, 617-496-1089, jsnyder@gov.harvard.edu *(email)*

- Study metadata will be entered in the Dataverse Cataloging Information page which provides ~ 100 fields to choose from, plus custom fields.
- Basic metadata fields are: Title, identifier, year, author, abstract, keywords.
- Additional documentation will be uploaded in PDF or plain text formats. Code can be uploaded too.
- A formal Data Citation will be generated automatically.
- Metadata will be exported into XML (DDI, DC).



Data Description	What data will be generated? How will you create the data? (simulated, observed, experimental, software, physical collections)
Existing Data	Will you be using existing data? Relationship between the data you are collecting and existing data.
Audience	Who will potentially use the data?
Access and Sharing	How will data files be shared? How will others access them?
Formats	What data formats will you be creating?
Metadata and Documentation	What documentation will you provide to describe the data? Metadata formats and standards?
Storage, backup, replication, versioning	**Are the data files backed up regularly? Are there replicas in different locations? Are older versions of the data kept?**
Security	Are the system and storage that will be used secure?
Budget	Any costs for preparing the data? Costs for storage and long-term access?
Privacy, Intellectual Property	Does the data contain private or confidential information? Any copyrights?
Archiving, Preservation, Long-term Access	What plans do you have to archive the data and other research products? Will it have long-term accessibility?
Adherence	How will you check for adherence of this plan?



James M. Snyder, Jr.

- The Dataverse will keep multiple versions of the data.
- Deposited data will never be destroyed (unless legally required).
- In addition, the Dataverse Network at IQSS provides system backups in a daily basis.

Dataverse

Cataloging Information | Documentation, Data and Analysis | User Comments | Versions

Version History

Difference	Version	Status	Comments	Released	Contributors
☐	3	Released		Mon Apr 11 15:12:13 EDT 2011	IQSS Admin
☐	2	Archived		Tue Apr 05 18:29:06 EDT 2011	IQSS Admin
☐	1	Archived		Tue Mar 08 09:11:00 EST 2011	Helen Lewis

Data Description	What data will be generated? How will you create the data? (simulated, observed, experimental, software, physical collections)
Existing Data	Will you be using existing data? Relationship between the data you are collecting and existing data.
Audience	Who will potentially use the data?
Access and Sharing	How will data files be shared? How will others access them?
Formats	What data formats will you be creating?
Metadata and Documentation	What documentation will you provide to describe the data? Metadata formats and standards?
Storage, backup, replication, versioning	Are the data files backed up regularly? Are there replicas in different locations? Are older versions of the data kept?
Security	**Are the system and storage that will be used secure?**
Budget	Any costs for preparing the data? Costs for storage and long-term access?
Privacy, Intellectual Property	Does the data contain private or confidential information? Any copyrights?
Archiving, Preservation, Long-term Access	What plans do you have to archive the data and other research products? Will it have long-term accessibility?
Adherence	How will you check for adherence of this plan?

- The Dataverse Network at IQSS follows "good computer use practices" set by the Security & Privacy group at Harvard.

Data Description	What data will be generated? How will you create the data? (simulated, observed, experimental, software, physical collections)
Existing Data	Will you be using existing data? Relationship between the data you are collecting and existing data.
Audience	Who will potentially use the data?
Access and Sharing	How will data files be shared? How will others access them?
Formats	What data formats will you be creating?
Metadata and Documentation	What documentation will you provide to describe the data? Metadata formats and standards?
Storage, backup, replication, versioning	Are the data files backed up regularly? Are there replicas in different locations? Are older versions of the data kept?
Security	Are the system and storage that will be used secure?
Budget	**Any costs for preparing the data? Costs for storage and long-term access?**
Privacy, Intellectual Property	Does the data contain private or confidential information? Any copyrights?
Archiving, Preservation, Long-term Access	What plans do you have to archive the data and other research products? Will it have long-term accessibility?
Adherence	How will you check for adherence of this plan?

Harvard's *home for social science research.*
Creating, sharing, and preserving scientific knowledge about human societies.

POWERED BY THE Dataverse Network™ PROJECT v. 3.0

IQSS Dataverse Network

Create Account Log In

Access the world's largest collection of social science research data here by searching across or browsing through one of the virtual data archives - called "dataverses" - listed

CREATE A DATAVERSE
Create a Dataverse to upload your own data sets

- The Dataverse Network at IQSS is free and open to all social science research data.
- Restrictions are 2GB per file, with no limit in the number of files.

  (In the future, a fee might apply to archive very large collections - currently under review)

Research Projects Scholars	University of the Thai Chamber of Commerce View Info [+]	University	Apr 9, 2012	
	American Journal of Political Science (AJPS)	Rice University	Apr 4, 2012	

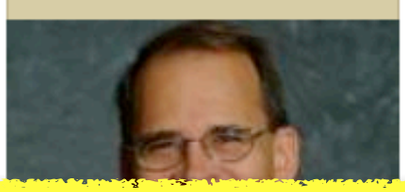
Data Description	What data will be generated? How will you create the data? (simulated, observed, experimental, software, physical collections)
Existing Data	Will you be using existing data? Relationship between the data you are collecting and existing data.
Audience	Who will potentially use the data?
Access and Sharing	How will data files be shared? How will others access them?
Formats	What data formats will you be creating?
Metadata and Documentation	What documentation will you provide to describe the data? Metadata formats and standards?
Storage, backup, replication, versioning	Are the data files backed up regularly? Are there replicas in different locations? Are older versions of the data kept?
Security	Are the system and storage that will be used secure?
Budget	Any costs for preparing the data? Costs for storage and long-term access?
Privacy, Intellectual Property	**Does the data contain private or confidential information? Any copyrights?**
Archiving, Preservation, Long-term Access	What plans do you have to archive the data and other research products? Will it have long-term accessibility?
Adherence	How will you check for adherence of this plan?

James M. Snyder, Jr.

- Data uploaded to the Dataverse cannot contain confidential information.
- Researcher agrees to the terms of use upon uploading the data, which states that:
  - You give permission and any required licenses to IQSS and the Archive to store and backup the materials
  - The Materials do not infringe upon the copyrights or other intellectual property rights, ..
  - 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 ...
  - You give permission and any required licenses to IQSS to make the Content available for archiving, preservation and access, within the Data Preservation Alliance for the Social Sciences ("Data-PASS")

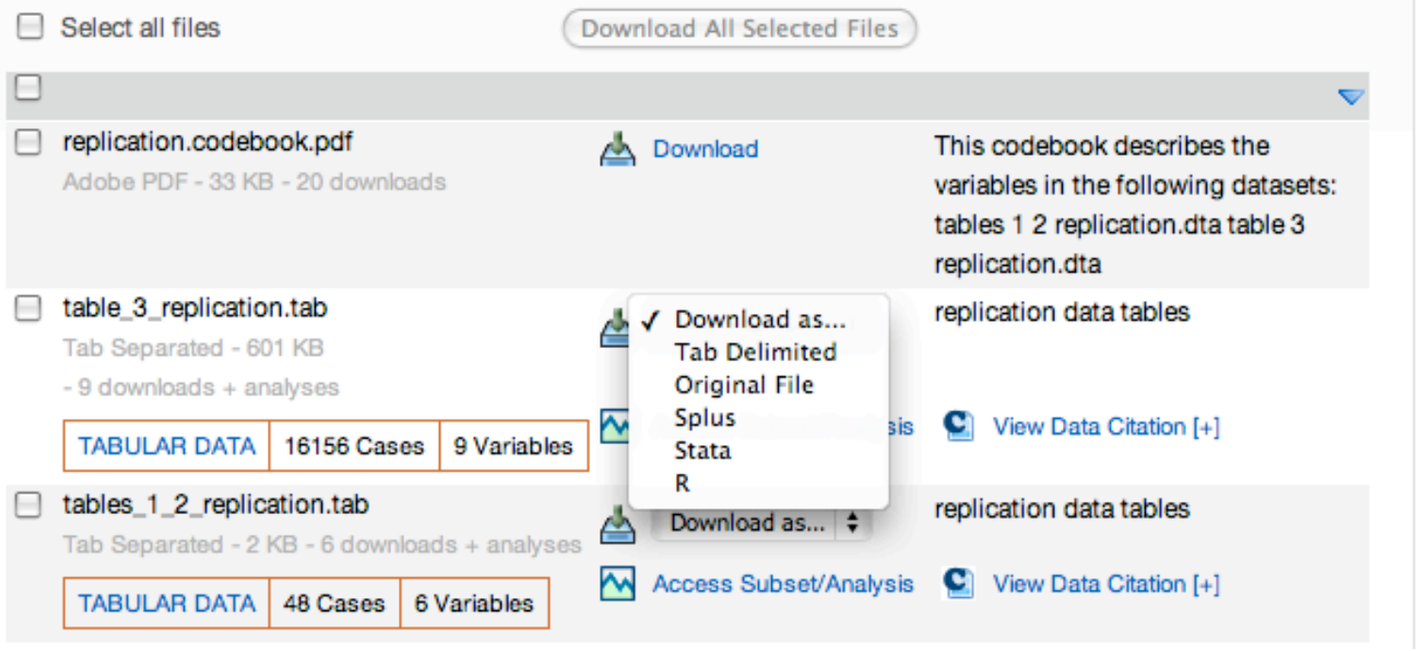
Continue

Data Description	What data will be generated? How will you create the data? (simulated, observed, experimental, software, physical collections)
Existing Data	Will you be using existing data? Relationship between the data you are collecting and existing data.
Audience	Who will potentially use the data?
Access and Sharing	How will data files be shared? How will others access them?
Formats	What data formats will you be creating?
Metadata and Documentation	What documentation will you provide to describe the data? Metadata formats and standards?
Storage, backup, replication, versioning	Are the data files backed up regularly? Are there replicas in different locations? Are older versions of the data kept?
Security	Are the system and storage that will be used secure?
Budget	Any costs for preparing the data? Costs for storage and long-term access?
Privacy, Intellectual Property	Does the data contain private or confidential information? Any copyrights?
Archiving, Preservation, Long-term Access	**What plans do you have to archive the data and other research products? Will it have long-term accessibility?**
Adherence	How will you check for adherence of this plan?



James M. Snyder, Jr.
Professor of Government

- The IQSS Dataverse Network commits to good archival practices:
  - Metadata is exported to XML
  - Data files are re-formatted for long term acces
  - All versions are kept
  - Metadata and data are replicated to multiple locations through LOCKSS



Data Description	What data will be generated? How will you create the data? (simulated, observed, experimental, software, physical collections)
Existing Data	Will you be using existing data? Relationship between the data you are collecting and existing data.
Audience	Who will potentially use the data?
Access and Sharing	How will data files be shared? How will others access them?
Formats	What data formats will you be creating?
Metadata and Documentation	What documentation will you provide to describe the data? Metadata formats and standards?
Storage, backup, replication, versioning	Are the data files backed up regularly? Are there replicas in different locations? Are older versions of the data kept?
Security	Are the system and storage that will be used secure?
Budget	Any costs for preparing the data? Costs for storage and long-term access?
Privacy, Intellectual Property	Does the data contain private or confidential information? Any copyrights?
Archiving, Preservation, Long-term Access	What plans do you have to archive the data and other research products? Will it have long-term accessibility?
Adherence	How will you check for adherence of this plan?

http://thedata.org

Dataverse Networks at Harvard (collaboration between IQSS, Harvard Library and University IT):

- IQSS Dataverse Network: Open to all Social Science research data
- Astronomy Dataverse Network: Open to the Harvard-Smithsonian Center for Astrophysics

Dataverse Networks in other institutions:

- The software is open-source and it's free to install in any institutions.

Thanks