INTRODUCTION TO THE DATAVERSE NETWORK™

JANUARY 7, 2015

Jonathan Crabtree
Assistant Director of Computing and Archival Research

THE ODUM INSTITUTE FOR RESEARCH IN SOCIAL SCIENCE
228 DAVIS LIBRARY, CB# 3355
UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL
CHAPEL HILL, NC  27599-3355
www.odum.unc.edu
DATA MANAGEMENT ACROSS THE RESEARCH LIFECYCLE

- What is Data Management?
- Why Manage Data?
- How Should I Manage My Data?
- Data Management Tools
  - Dataverse
ODUM INSTITUTE HELPS MANAGE DATA
ABOUT THE DATAVERSE NETWORK™

- Developed at Harvard University’s Institute for Quantitative Social Science (IQSS)
- Open source virtual archive software application
- Adopted by institutions worldwide
- Version 4.0 currently in development
- Learn more at: http://thedata.org
DATA COLLECTIONS

- Odum Dataverse Network
  - Over 33,739 studies
  - Over 690,000 files
- Partnership with Harvard IQSS Data Center
- Federated system
DATA MANAGEMENT SERVICES

- Odum provides three levels of service
  - Self-Archiving
  - Guided Archiving
  - Comprehensive Lifecycle Management
COME HEAR MORE AT THE DEMONSTRATION!
Archivists work alongside research team
Advising on data management issues
Supporting funders data sharing requirements
Promoting the project and the data generated
An example is MEASURE Evaluation
- Based at the Carolina Population Center
- http://vimeo.com/105755677
- http://www.cpc.unc.edu/measure
Branded To Promote Project

41 Studies

Detailed Usage Statistics
DATAVERSE NETWORK™ FEATURES

Data sharing and archiving with control and recognition for data producers

- Support for all file types
- Persistent data citations
- Data restrictions
- Customized branding
DATAVERSE NETWORK™ FEATURES

Rich data support for certain file formats

- Metadata extraction, subsetting, and R analysis for SPSS, Stata, and R
- Metadata extraction for FITS data
- Subsetting for social network data (GraphML)
DATAVERSE NETWORK™ FEATURES

Data management, standards, and archival best practices

- Data versioning
- General and domain-specific metadata following metadata standards
- Traffic and downloads tracking with Guestbook feature
- Permanent storage of data in preservation formats
- Geographically distributed preservation copies
ABOUT THE DATAVEVERSE NETWORK™

Odum Institute Dataverse Network – arc.irss.unc.edu/dvn

The Odum Institute Dataverse Network provides access to data collections curated by the Odum Institute as well as collections owned by other institutions and individual scholars. You can search across or browse any of these dataverses listed below. You may also create your own branded dataverse to manage and provide access to your data. Learn more about the Dataverse Network Project at http://www.thedata.org or view our video tutorials.

Many files have access restrictions. For more information, click Odum File Restrictions.

<table>
<thead>
<tr>
<th>Dataverses</th>
<th>Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>41 Dataverses</td>
<td>25,788 Studies, 117,322 Files, 505,910 Downloads</td>
</tr>
</tbody>
</table>

A Dataverse is a container for research data studies, customized and managed by its owner.

A study is a container for a research data set. It includes cataloging information, data files and complementary files.
GETTING STARTED

Create your user

CREATE ACCOUNT > CREATE YOUR OWN DATaverse

As a registered user, you can:
- Add studies to open and wiki dataverses, if available
- Contribute to existing studies in wiki dataverses, if available
- Add user comments to studies that have this option
- Create your own dataverse

Username
Password
Retype Password
First Name
Last Name
E-Mail
Institution
Position
Phone Number

Create Account
Cancel
GETTING STARTED

Agree to the terms of use
GETTING STARTED

Create a Dataverse

SUCCESS! > REGISTER WITH US

As a registered user, you can:

- Add studies to open and wiki dataverses, if available
- Contribute to existing studies in wiki dataverses, if available
- Add user comments to studies that have this option
- Create your own dataverse

Go to your account page to see your information. If you are interested in accessing restricted studies in this Dataverse Network, please check the data availability and terms of use of the study, or contact the dataverse administrator for more information.
CREATE YOUR OWN DATASURE

Create a dataverse to upload your own data sets and create collections of data. Archiving with a dataverse extends the useful life of data, preserving it in perpetuity against the failure of computers systems and the obsolescence of data formats.

1. CREATE ACCOUNT
Register with our network, by creating an account...

2. AGREE TO TERMS
Read and agree to the terms of use...

3. NAME DATASURE
Enter the title, type and affiliation... Click save...

✓ SUCCESS!
You have created your own dataverse; now you are ready to start adding data.

CONTINUE >>
Complete the fields on this page to create your dataverse. You can select more than one classification and add your dataverse to those groups.
**RELEASE YOUR DATAVERSE**

Introduction to Data Curation Dataverse Network >

**ddd Dataverse**

![Image of Dataverse settings interface]

**General** | **Advanced** | **Customization** | **Promote Your Dataverse**
---|---|---|---

- **Dataverse Release Settings**: Not Released
- **Dataverse Name**: ddd
  - Name used to refer to this dataverse in network homepage and other pages.
- **Dataverse Alias**: ddd
  - Short name used to build the URL for this dataverse, e.g., http://.../dv/alias'. It is case sensitive.
DATA ARCHIVING WITH DATaverse NETWORK™

**STEP 1: Prepare dataset files**

- **Dataset**
  - SPSS
  - Stata
  - R

- **Documentation**
  - Instrument
  - Data Deposit Form
  - Methodology brief/notes
  - Codebooks
  - Related publications
  - Other related files
STEP 2: Create a study
THE EARLY PREGNANCY STUDY

If you use these data, please add the following citation to your scholarly references. Why cite?

Allen Wilcox, 2013-09-29, "The Early Pregnancy Study", http://hdl.handle.net/1902.28/10487 UNF:1srx90Xm9xg7i59a2b95/087904922809

Citation Details

Study Global ID: hdl:1902.28/10487
Authors: Allen Wilcox (NIH-NIEHS-Epidemiology Branch)
Producer: Bob McConnaughey (DHHS, WastewaterNEISIHS Epidemiology Support)
Distributor: Bob McConnaughey, WastewaterNEISIHS Epidemiology Support

Provenance: Bob McConnaughey Dataverse

Description and Scope
DATA ARCHIVING WITH DATAVERSE NETWORK™

STEP 3: Upload the study files

Introduction to Data Curation Dataverse Network >

Charleston Conference Materials Dataverse

T
hdl:TEST/10011

DRAFT Study Version: 1 – No Released Version

- Edit Cataloging Information
- Add File(s)
- Edit Study Version Notes
- Release

Permissions
- Create Study Template
- Delete Draft Version

Cataloging Information  DATA & ANALYSIS  Comments (0)  Versions

No files have been uploaded to this study. To add a file, use the upload files utility on the Add File(s) page.
**UPLOAD FILES**

Select the Data Type from the drop-down list. Then click the Browse to select the file to add to this study.

SPSS (.sav or .por), STATA (.dta), CSV (character-separated values), TAB (tab-delimited), or GraphML (.xml) files are processed as subsettable data files, which can be analyzed online by using the Dataverse Network analysis tools. When selecting a CSV (character-separated values) data type, an SPSS Control Card file is first required. When selecting a TAB (tab-delimited) data type, a DDI Control Card file is first required.

You can use the file Category to define and group the types of file that you upload, such as Documentation or Data files.

After you have selected the file(s) that you would like to add to the study, click Save.

<table>
<thead>
<tr>
<th>Category</th>
<th>File Name</th>
<th>Description</th>
<th>Size (bytes)</th>
<th>Remove</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>harris_s24784_spss.tab</td>
<td>Harris Study Datafile SPSS</td>
<td>3056228</td>
<td></td>
</tr>
</tbody>
</table>

NOTE: Subsettable files might take a long time to upload and be processed. We will send you an e-mail notification when the upload process starts and when it completes.

E-Mail Address: Jonathan_Ostroe@unc.edu

Separate multiple e-mail addresses with a comma (and no spaces).
The Early Pregnancy Study:
Allen J Wilcox, Donna D. Baird, Clarice Weinberg

<table>
<thead>
<tr>
<th>Variable Information Table</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type</strong></td>
</tr>
<tr>
<td>Discrete</td>
</tr>
<tr>
<td>Discrete</td>
</tr>
<tr>
<td>Discrete</td>
</tr>
<tr>
<td>Discrete</td>
</tr>
<tr>
<td>Discrete</td>
</tr>
<tr>
<td>Discrete</td>
</tr>
<tr>
<td>Discrete</td>
</tr>
<tr>
<td>Discrete</td>
</tr>
<tr>
<td>Discrete</td>
</tr>
<tr>
<td>Character</td>
</tr>
<tr>
<td>Character</td>
</tr>
<tr>
<td>Character</td>
</tr>
<tr>
<td>Character</td>
</tr>
<tr>
<td>Character</td>
</tr>
<tr>
<td>Character</td>
</tr>
<tr>
<td>Character</td>
</tr>
<tr>
<td>Character</td>
</tr>
<tr>
<td>Character</td>
</tr>
<tr>
<td>Character</td>
</tr>
<tr>
<td>Character</td>
</tr>
</tbody>
</table>

**Value (Label)**
- 0 (No)
- L (Late Void)
- X (Yes, Not Noted on DRC)
- Y (Yes - Void Collected)

**Frequency**
- 214.0
- 195.0
- 526.0

**UNF**:
- UNF:5:48L/1yuGz7Wpfxz7h1f95d6==

*Note: The image contains a screenshot of a data analysis page from the Odum Institute Dataverse Network.*
MATERIALS FOR CHARLESTON CONFERENCE INTRODUCTION TO DATA CURATION

Download Subsets: Piccoza & Case-Subset, Descriptive Statistics, Advanced Statistical Analysis

Selected Variables: Choose File Format to download selected variables:
- Text
- R Data
- Stata
- Data

Select variables from table below (selected variables will be displayed above)

<table>
<thead>
<tr>
<th>Variable Information Table</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>Name</td>
<td>Label</td>
<td>Summary</td>
<td></td>
</tr>
<tr>
<td>Discrete</td>
<td>ID</td>
<td>Survey Craft ID</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discrete</td>
<td>NEW_WT</td>
<td>Weight Factor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discrete</td>
<td>WT</td>
<td>Weight Factor</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discrete</td>
<td>GEST_DAY</td>
<td>[label missing]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Character (Date Value)</td>
<td>BDAY1</td>
<td>[label missing]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Character (Date Value)</td>
<td>DDAY1</td>
<td>[label missing]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discrete</td>
<td>BMI_BEGI</td>
<td>BMI Before Pregnancy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discrete</td>
<td>BMI_BIRT</td>
<td>BMI at Birth</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discrete</td>
<td>BMI_NOW</td>
<td>BMI Now</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discrete</td>
<td>BMI_BE_1</td>
<td>BMI Before Pregnancy by Category</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Value (Label)          | Frequency  
1 (Underweight)       | 54.0        
2 (Normal Weight)      | 676.0       
3 (Overweight)         | 375.0       
4 (Obese)              | 373.0       
UNF                     |             

Statistical analysis powered by Zelig
HARRIS 2006 LISTENING TO MOTHERS II, STUDY NO. 24784


Citation for full data set:

Citation for subset analysis:
ORGANIZING DATA WITH DATaverse NETWORK™
FUTURE GOALS

- Policy based data management
  - iRODS
  - Virtual Institute for Social Research (VISR)
- Secure research workspace
  - Collaboration with secure data
INTRODUCTION TO THE DATAVERSE NETWORK™

Odum Institute Data Archive
odumarchive@unc.edu

Thu-Mai Christian
Archival Research Associate
tlchristian@unc.edu

Jonathan Crabtree
Assistant Director for Computing and Archival Research
Jonathan_Crabtree@unc.edu

Sophia Lafferty-Hess
Research Data Manager
slaffer@email.unc.edu
The Odum Institute for Research in Social Science was founded in 1924 by Howard Washington Odum, making it the oldest university-based interdisciplinary social science research institute in the United States. We provide education, training, data collection, and archive services to researchers both within and beyond the University of North Carolina at Chapel Hill. Our mission is to facilitate scientifically rigorous social science research that contributes to better lives for the citizens of North Carolina and the world.
THE H. W. ODUM INSTITUTE
FOR RESEARCH IN SOCIAL SCIENCE

90 years

Lets give it a test drive!