

HARVARD
UNIVERSITY



Research Data Sharing: Dataverse at the Library

IT Summit 2012

Today's Agenda

**Research Data
Sharing:**

**Dataverse at the
Library**

Dataverse Pilot

How is Dataverse being Used?

Wendy Gogel – Manager of Content and Projects, HUIT - LTS

Dataverse Development and Features

Gustavo Durand – Manager of the Dataverse Network, IQSS

Dataverse Architecture and Infrastructure

Bill Horka – Backend Systems Software Developer, IQSS

Today's Agenda

**Research Data
Sharing:**

**Dataverse at the
Library**

Dataverse Pilot

How is Dataverse being Used?

Wendy Gogel – Manager of Content and Projects, HUIT - LTS

Dataverse Development and Features

Gustavo Durand – Manager of the Dataverse Network, IQSS

Dataverse Architecture and Infrastructure

Bill Horka – Backend Systems Software Developer, IQSS

Digital Scholarship Program

The program goal is to enable researchers to create, curate, and disseminate the digital objects related to research.

Digital Scholarship Advisory Group

Zak Kohane (HMS)

Stuart Shieber (OSC, FAS)

John Palfrey (HLS)

Peter Bol (FAS)

Amy Brand (OPP)

Jeffrey Schnapp (FAS,
metaLab, Berkman)

Gary King (IQSS)

Alyssa Goodman (CfA)

Anne Margulies (HUIT)

Jim Waldo (SEAS)

Gosia Stergios (HL)

Franziska Frey (HL)

Mary Lee Kennedy (HL)

Collaborative Pilot

Leads: Franziska Frey (HL), Mercé Crosas (IQSS)

To pursue offering Dataverse as a University-wide service



The Importance of Managing & Preserving Data

- **For future access & re-use**
 - Validate published results
 - Build-on the results for further research
 - Re-analyze data for new results
- **Recognition**
 - Credit for research
 - Publication citations
- **Requirements from funding agencies**

Data Management Plan Requirements

Home Funding Awards Discoveries News Publications Statistics About FastLane



National Science Foundation
Directorate for Engineering (ENG)

NSF Web Site



ENG Home

ENG Funding

ENG Awards

ENG Discoveries

ENG News

About ENG

Engineering

design element

NSF Data Management Plan Requirements

Beginning January 18, 2011, proposals submitted to NSF must include a supplementary document of no more than two pages labeled "Data Management Plan" (DMP). This supplementary document should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results. Proposals that do not include a DMP will not be able to be submitted. For more information about this new requirement, please see the [Grant Proposal Guide, Chapter II.C.2.j](#) and the [Data Management and Sharing Frequently Asked Questions\(FAQs\)](#).

Please note: the Engineering Directorate (ENG) has additional guidance for proposals submitted to ENG programs, http://nsf.gov/eng/general/ENG_DMP_Policy.pdf. Questions and/or suggestions about this new requirement may be addressed to Dr. Maria K. Burka at mburka@nsf.gov.

[ENG Home](#)

[About ENG](#)

[Funding Opportunities](#)

[Awards](#)

[News](#)

[Events](#)

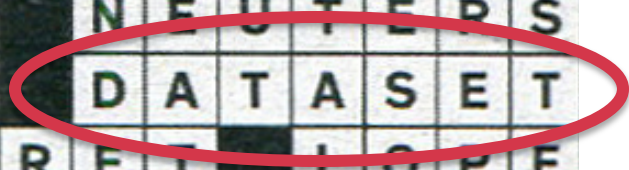
[Discoveries](#)

[Publications](#)

[Advisory Committee](#)

[Career Opportunities](#)

E	S	T	A	T	E		F	A	L	C	O	N		A	T	P	E	A	C	E	
M	O	O	R	E	D		O	N	E	D	G	E		N	E	U	T	E	R	S	
A	U	X	I	L	I	A	R	Y	V	E	R	B		D	A	T	A	S	E	T	
I	S	I	S		E	R	G	O		F	E	R	R	E	T		L	O	P	E	
L	E	N	T	O		M	A	N	E					E	A	R	P		P	E	R
			A	C	T	I	V	E	V	O	L	C	A	N	O	E	S				
P	B	A		T	R	E	E		I	D	I	O	M			E	L	A	N	D	
C	A	L	L	O	U	S		F	L	O	P	S		O	B	L	I	Q	U	E	
P	H	O	E	B	E		J	E	E	R	S		I	N	R	E	T	U	R	N	
		E	N	E		T	O	N	Y	S		O	N	E	A		A	M	T		
A	T	V	A	R	I	A	N	C	E		A	R	T	I	C	L	E	V	I	I	
V	E	E			C	R	E	E		S	N	A	R	L		O	R	I			
A	P	R	I	C	O	T	S		L	O	T	T	O		S	C	O	T	C	H	
S	E	A	S	O	N	S		M	I	L	L	E		S	A	U	S	A	G	E	
T	E	S	L	A			A	I	M	E	E		A	L	I	S		E	I	N	
			A	L	E	S	S	A	N	D	R	O	V	O	L	T	A				
W	A	H		S	R	A	S					S	N	A	G		S	T	R	A	D
A	D	E	E		A	L	T	T	A	B		E	T	A	S		M	E	D	O	
F	O	R	M	O	S	A		A	F	R	I	C	A	N	V	I	O	L	E	T	



Pilot Goals

- Pursue the migration of Dataverse to centralized infrastructure maintained by HUIT
- Provide support and training through the libraries in collaboration with IQSS and CfA
- Assess outcome to see how the library would provide a similar service to other research data repositories at Harvard

FY13 Service Offerings

- **Introduction to the Concepts of Data Management**
 - Why?
 - Legal and policy considerations
 - Preparing data
- **How to use Dataverse**
 - For Astrophysicists
 - Including case studies and personal assistance
 - For Social Scientists
 - Including case studies and personal assistance

Potential Future Offerings

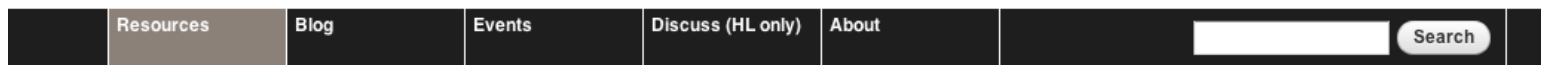
- **Sessions on**
 - Data Management Plans
 - Dataverse for multiple users and shared – contributions (e.g., journals, collaborative research projects)
 - Dissertations
- **Process for adding new Dataverse Networks in other domains**
- **Network of liaisons and contacts**

Digital Scholarship Web Site

Look for the announcement of our fall schedule for the new sessions and other information:

<http://ds.hul.harvard.edu/ds>

 DIGITAL SCHOLARSHIP @ HARVARD



[Home](#) » [Resources](#) » [All Categories](#)

Resources

> All Resources

Category

> Communities & Groups

> **Data Challenge**

> Open Access & Source

> Preservation Opportunities

> Research Impact Metrics

> Social Networks

> Curation & Annotation

> Library Services

> Research Lifecycle

Category: Data Challenge



DATA CHALLENGE

The Bookworm arXiv

Search for trends in hundreds of thousands of scientific articles on [arxiv.org](#)

Last Updated: May 20, 2012



DATA CHALLENGE

The Dataverse Network

The Dataverse Network offers you the option to create a dataverse - *your own data archive* - for uploading and sharing.

Last Updated: May 20, 2012



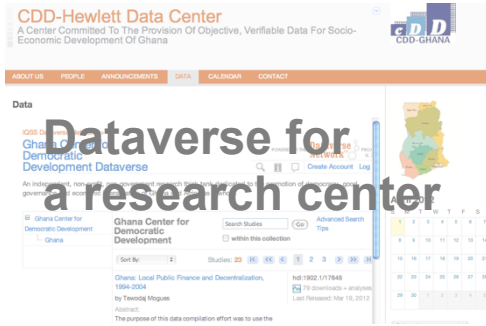
DATA CHALLENGE

The Astronomy Dataverse Network

The Astronomy data repository for Harvard affiliates allows you to...

Last Updated: May 19, 2012

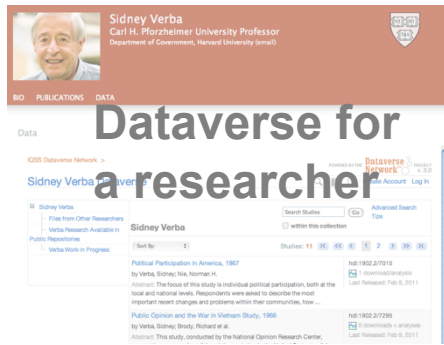
How is Dataverse being used?



CDD-Hewlett Data Center
A Center Committed To The Provision Of Objective, Verifiable Data For Socio-Economic Development Of Ghana

Dataverse for a research center

Search for: Ghana Center for Democratic Development
Studies: 23



Dataverse for a researcher

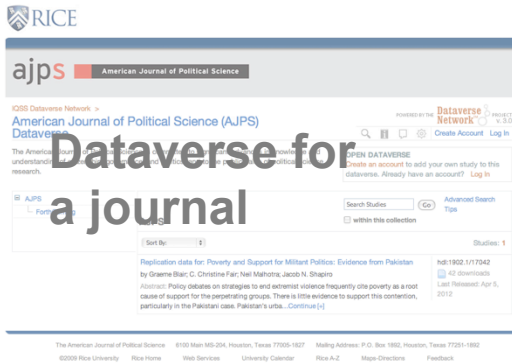
Search for: Sidney Verba
Studies: 11



Dataverse for a Data Management Plan

Search for: James M. Snyder, Jr.
Studies: 1

Dataverse Network



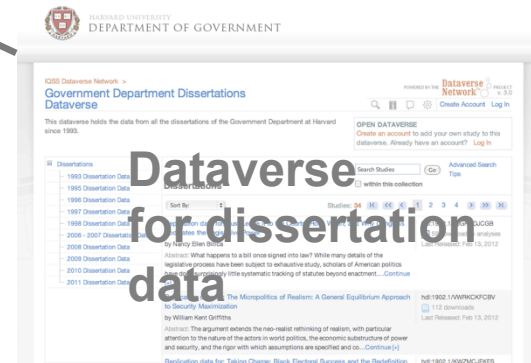
Dataverse for a journal

Search for: American Journal of Political Science (AJPS)
Studies: 1



Dataverse for an archive

Search for: Murray Research Archive Original Collection
Studies: 1



Dataverse for dissertation data

Search for: Government Department Dissertations
Studies: 34

How is Dataverse being used?

An individual researcher

- Andrew Beath, Department of Government, Harvard University

Randomized Institutional Isomorphism - Evidence from Afghanistan

A collaborative program

- Program in Quantitative Genomics, Harvard School of Public Health

Responding to “increasing interdisciplinary needs, especially quantitative needs, in handling massive data in genetics and genomics in the HSPH and the Longwood area.”



Dr. Alicia Margarita Soderberg

Assistant Professor, Harvard Astronomy Department



Harvard-Smithsonian Center for Astrophysics, Institute for Theory and Computation

60 Garden Street, MS-51, Cambridge, MA 02138

617-496-7919 (office) 617-335-0939 (mobile)

[\(email\)](#)

Biography

Supernova Forensics

Astronomy Courses

Group Members

Publications

Dataverse

Press Releases

[All Astronomy Dataverses >](#)

Supernova Forensics Dataverse

POWERED BY THE **Dataverse Network**™ PROJECT v. 2.2.5

[Search](#) [User Guides](#) [Report Issue](#)

[Log In](#) [Create Account](#)

This page is dedicated to the data products published by the Harvard Supernova Forensics research team led by Professor Soderberg. Please feel free to download and use these data tables (and in some cases, reduced data products) and cite using the handle. For any questions, please contact: asoderberg@cfa.harvard.edu

Alicia Soderberg

[Advanced Search](#)
[Tips](#)

Sort By:

Studies: **1**

[Replication data for: Panchromatic Observations of SN2011dh](#)

by Alicia Soderberg

Abstract: We report the discovery and detailed monitoring of X-ray emission associated with the Type IIb SN 2011dh using data from the Swift and Chandra satellites, placing it among the best studied X-ray supernovae ...

hdl:10904/10126

19 downloads

Last Released: Jan 30, 2012

Today's Agenda

Research Data Sharing:

Dataverse at the Library

Dataverse Pilot

How is Dataverse being Used?

Wendy Gogel – Manager of Content and Projects, HUIT - LTS

Dataverse Development and Features

Gustavo Durand – Manager of the Dataverse Network, IQSS

Dataverse Architecture and Infrastructure

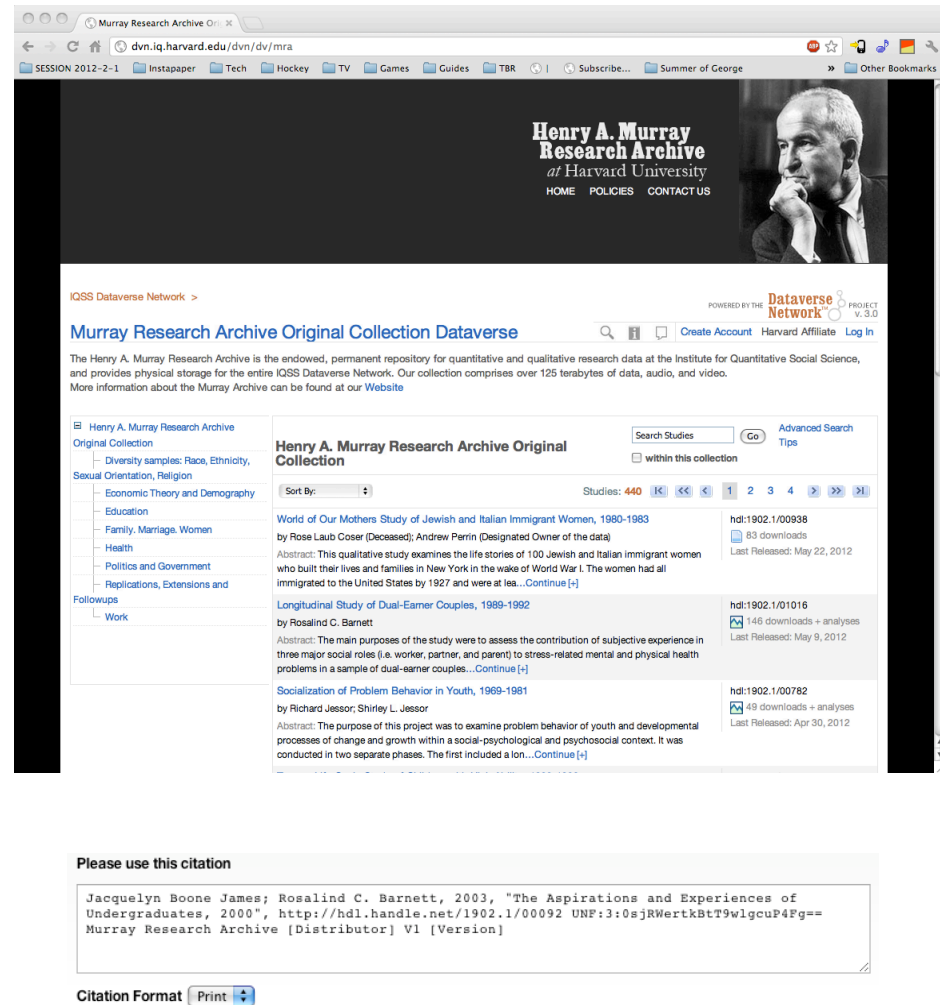
Bill Horka – Backend Systems Software Developer, IQSS

The Dataverse Network

- An open-source Application for Sharing, Discovering and Preserving Data
 - It enables data archiving and preservation through reformatting, standards and exchange protocols
 - It provides control and recognition for researchers through data management, branding and formal data citation

Basic Features

- Seamless integration with your website (Custom Branding / Integration with Open Scholar)
- Version Management
- Permission control at the dataverse, study, or file level
- Collections for grouping studies (static, dynamic, linked)
- Data Conversion to a Preservable and Verifiable Format
- Persistent Citation for Data
- Search all metadata fields



The screenshot displays the Murray Research Archive website. At the top, there is a navigation bar with the site's name and a portrait of Henry A. Murray. Below this, the main content area features a search bar and a list of studies. The first study listed is "World of Our Mothers Study of Jewish and Italian Immigrant Women, 1980-1983" by Rose Laub Coser (Deceased) and Andrew Perrin. The second study is "Longitudinal Study of Dual-Earner Couples, 1989-1992" by Rosalind C. Barnett. The third study is "Socialization of Problem Behavior in Youth, 1969-1981" by Richard Jessor and Shirley L. Jessor. Each study entry includes a brief abstract and a download icon. At the bottom of the page, there is a section for "Please use this citation" with a pre-formatted citation for the first study and a "Citation Format" button.

Henry A. Murray Research Archive at Harvard University

HOME POLICIES CONTACT US

IQSS Dataverse Network

POWERED BY THE Dataverse Network PROJECT v. 3.0

Murray Research Archive Original Collection Dataverse

Create Account Harvard Affiliate Log In

The Henry A. Murray Research Archive is the endowed, permanent repository for quantitative and qualitative research data at the Institute for Quantitative Social Science, and provides physical storage for the entire IQSS Dataverse Network. Our collection comprises over 125 terabytes of data, audio, and video. More information about the Murray Archive can be found at our Website

Henry A. Murray Research Archive Original Collection

Search Studies Go Advanced Search Tips

within this collection

Sort By: Studies: 440

World of Our Mothers Study of Jewish and Italian Immigrant Women, 1980-1983
by Rose Laub Coser (Deceased); Andrew Perrin (Designated Owner of the data)
Abstract: This qualitative study examines the life stories of 100 Jewish and Italian immigrant women who built their lives and families in New York in the wake of World War I. The women had all immigrated to the United States by 1927 and were at lea...Continue [+]
hdi:1902.1/00938
83 downloads
Last Released: May 22, 2012

Longitudinal Study of Dual-Earner Couples, 1989-1992
by Rosalind C. Barnett
Abstract: The main purposes of the study were to assess the contribution of subjective experience in three major social roles (i.e. worker, partner, and parent) to stress-related mental and physical health problems in a sample of dual-earner couples...Continue [+]
hdi:1902.1/01016
146 downloads + analyses
Last Released: May 9, 2012

Socialization of Problem Behavior in Youth, 1969-1981
by Richard Jessor; Shirley L. Jessor
Abstract: The purpose of this project was to examine problem behavior of youth and developmental processes of change and growth within a social-psychological and psychosocial context. It was conducted in two separate phases. The first included a lon...Continue [+]
hdi:1902.1/00782
49 downloads + analyses
Last Released: Apr 30, 2012

Please use this citation

Jacquelyn Boone James; Rosalind C. Barnett, 2003, "The Aspirations and Experiences of Undergraduates, 2000", <http://hdl.handle.net/1902.1/00092> UNF:3:0sJRWerktk8tT9wlgcuP4Fg== Murray Research Archive [Distributor] V1 [Version]

Citation Format Print

Defining a Study

- Simple web form to enter cataloging field
- More than 100 (searchable) fields to describe a study
 - authors, abstract, geospatial information, data collection and methodologies, etc.
- Ability to create custom fields and define controlled vocabularies
- Ability to create templates
 - Require, recommend or hide fields
 - Provide default values or select controlled vocabulary

The image shows a web form titled "Citation Information" with several sections. The "Citation Information" section includes fields for Title, Study ID (with a value of hdl:1902.1/12678), Author (Name and Affiliation), and Producer (Producer, Affiliation, URL, and Logo URL). Below this are fields for Production Date, Distributor (Distributor, Affiliation, URL, and Logo URL), Contact (Contact and Affiliation), and E-mail. There are also fields for Distribution Date, Deposit Date (with a value of 2009-05-08), and Replication For. The bottom section is titled "Abstract and Scope" and contains an "Abstract" field. A note at the bottom states: "Note: Copying and pasting from a Word document can create errors when you save this page."

Uploading Files

- Upload any file type (data files, documents, code, images, etc)
- Tabular data (in recognized formats) automatically get additional services
- Network data uploaded in GraphML automatically get subsetting and some analysis options
- File size limit is 2 GB, but the number of files in a study is unlimited
- Group files by categories

The screenshot displays a file management interface with a list of files organized into three study categories. At the top, there is a checkbox for "Select all files" and a button for "Download All Selected Files".

1. Political Participation and Equality in Seven Nations, 1966-1971

- 07768-0001-Codebook.pdf (Adobe PDF - Unknown file size - 1 download) with a "Download" button and "Codebook" label.
- 07768-0001-Data.txt (Fixed Field Text Data - Unknown file size - 0 downloads + analyses) with a "Download as..." dropdown, "SUBSETTING" button, "1769 CASES" and "414 VARIABLES" indicators, "Access Subset/Analysis" button, and "View Data Citation [+]" button.
- 07768-0001-Setup.sas (SAS Syntax - Unknown file size - 0 downloads) with a "Download" button and "Setup" label.
- 07768-0001-Setup.sps (SPSS Syntax - Unknown file size - 0 downloads) with a "Download" button and "Setup" label.

2.2: India

- 07768-0002-Codebook.pdf (Adobe PDF - Unknown file size - 0 downloads) with a "Download" button and "Codebook" label.
- 07768-0002-Data.txt (Fixed Field Text Data - Unknown file size - 0 downloads + analyses) with a "Download as..." dropdown, "SUBSETTING" button, "2637 CASES" and "384 VARIABLES" indicators, "Access Subset/Analysis" button, and "View Data Citation [+]" button.
- 07768-0002-Setup.sas (SAS Syntax - Unknown file size - 0 downloads) with a "Download" button and "Setup" label.
- 07768-0002-Setup.sps (SPSS Syntax - Unknown file size - 0 downloads) with a "Download" button and "Setup" label.

3.3: Japan

- 07768-0003-Codebook.pdf (Adobe PDF - Unknown file size - 0 downloads) with a "Download" button and "Codebook" label.
- 07768-0003-Data.txt (Fixed Field Text Data - Unknown file size - 0 downloads + analyses) with a "Download as..." dropdown, "SUBSETTING" button, "2657 CASES" and "272 VARIABLES" indicators, "Access Subset/Analysis" button, and "View Data Citation [+]" button.
- 07768-0003-Setup.sas (SAS Syntax - Unknown file size - 0 downloads) with a "Download" button and "Setup" label.
- 07768-0003-Setup.sps (SPSS Syntax - Unknown file size - 0 downloads) with a "Download" button and "Setup" label.

Advanced Features

- **Tabular Data Sets**

- UNF
- Download a subset of variables
- Recode or subset cases
- Get Summary and Descriptive Statistics
- Run Advanced Statistical Analysis
- Data Visualization (for Time Series)

- **Network Data Sets**

- Write a Manual Query to create a subset of the data
- Use one of the pre-defined automatic queries (Largest Graph, Neighborhood)
- Run a Network Measure (Page Rank, Degree, Unique Degree, In Largest Component)

- **Harvesting from other Dataverse Networks** (or OAI Providers)

- **API for Automated Access** (Search, Metadata, Data)

POLITICAL PARTICIPATION AND EQUALITY IN SEVEN NATIONS, 1966-1971
DATA FILE: 07768-0001-DATA.TXT

Download Subset Recode and Case-Subsetting Descriptive Statistics **Advanced Statistical Analysis**

Selected Variables
V2
V4
V5
V6
V7

Select variables from table below (select all)

Choose a Statistical Model

- ✓ Choose a Statistical Model
- Categorical Data Analysis
 - Cross-Tabulation
- Ecological Inference Models
 - Hierarchical Multinomial-Dirichlet Ecological Inference Model for R x C Tables
- Event Count Models
 - Negative Binomial Reg for Event Count Dep Vars
 - Poisson Reg for Event Count Dep Vars
- Models for Continuous Bounded Dependent Variables
 - Exponential Reg for Duration Dep Vars
 - Gamma Reg for Cont, Positive Dep Vars
 - Log-Normal Reg for Duration Dep Vars
 - Weibull Reg for Duration Dep Vars
- Models for Continuous Dependent Variables
 - Least Squares Reg for Cont Dep Vars
 - Linear regression for Left-Censored Dep Variable
- Models for Dichotomous Dependent Variables
 - Logistic Reg for Binary Dep Vars
 - Probit Reg for Binary Dep Vars
 - Rare Events Logistic Reg for Binary Dep Vars
- Models for Ordinal Dependent Variables
 - Ordinal Logistic Reg for Ordered Cat Dep Vars
 - Ordinal Probit Reg for Ordered Cat Dep Vars

Variable Information Table

<input type="checkbox"/>	Type			
<input type="checkbox"/>	Continuous			
<input checked="" type="checkbox"/>	Continuous			
<input type="checkbox"/>	Continuous	V3	COMMID NAME OF COMM	
<input checked="" type="checkbox"/>	Continuous	V4	SEX SEX OF RESP	
<input checked="" type="checkbox"/>	Continuous	V5	AGE RESP AGE LAS	
<input type="checkbox"/>	Continuous	V6	AGEGRP RESP AGE IN	
<input checked="" type="checkbox"/>	Continuous	V7	MARITAL MARITAL STAT	
<input type="checkbox"/>	Continuous	V8	NHSHOLD NO. PERSONS	
<input type="checkbox"/>	Continuous	V9	NUNDER20 NO.IN HSHOLD	

Show
10 Variables

1 2 3 4 >>> >|

Summary

Development Effort

- **Open Source Software** (Apache 2 license)
 - Java EE6 application in Glassfish 3.1
 - Front End: JSF 2 (also Javascript and jQuery)
 - Back End: EJB 3.1, CDI, JPA 2
 - PostgreSQL database
 - Lucene Search Engine Library
 - Rserve for statistical analysis
- **Development Team**
 - 3* dedicated developers
 - 1 QA engineer
 - 1 (shared) UI designer

Today's Agenda

Research Data Sharing:

Dataverse at the Library

Dataverse Pilot

How is Dataverse being Used?

Wendy Gogel – Manager of Content and Projects, HUIT - LTS

Dataverse Development and Features

Gustavo Durand – Manager of the Dataverse Network, IQSS

Dataverse Architecture and Infrastructure

Bill Horka – Backend Systems Software Developer, IQSS

Software

- **Required**

- Linux or Mac OS X
 - RHEL or CentOS recommended
- Java
 - Oracle JRE
- Glassfish
 - Java Application Server
- PostgreSQL
 - DB Server

- **Recommended**

- R
 - Statistical Computing Language
- HANDLE.NET
 - Unique ID service

- **Optional**

- LOCKSS
 - Data & Metadata Replication Service
- z39.50
 - Search Gateway

Hardware

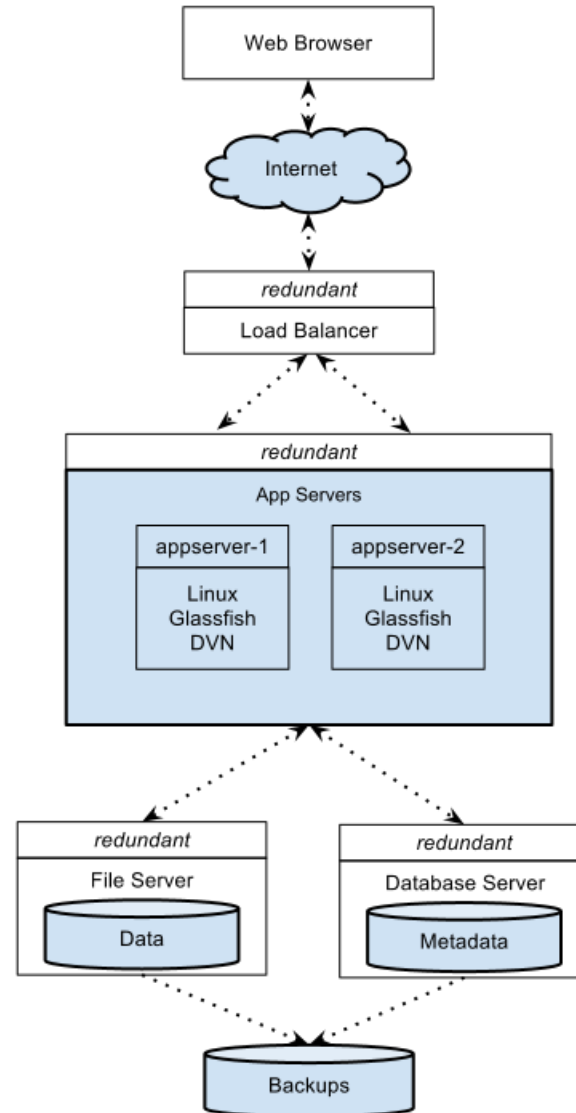
- **Minimal**

- Single Linux VM
 - Java, Glassfish, PostgreSQL
 - DVN Application
 - Local data and metadata storage

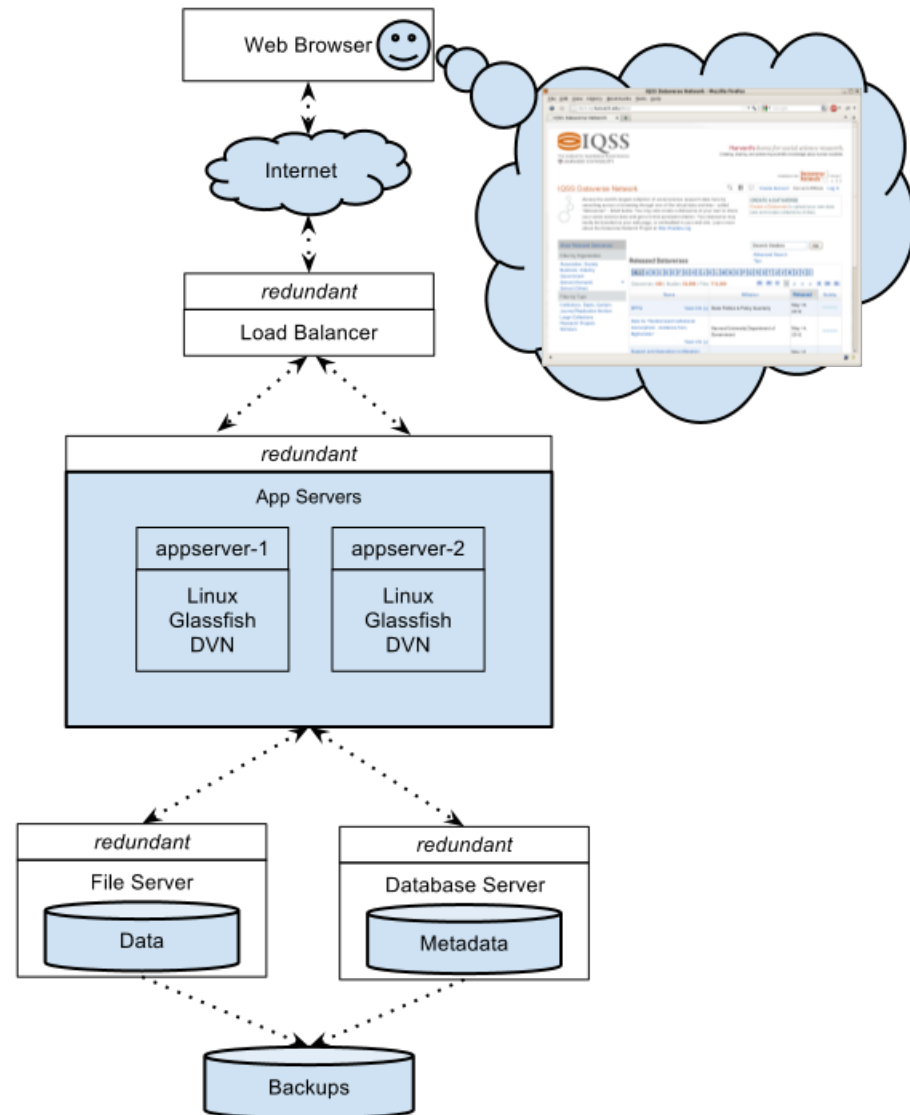
- **High Availability**

- DVN App servers
 - Load-balanced
 - Java, Glassfish, DVN App
- PostgreSQL DB servers
 - Fault-tolerant
- SAN/NAS storage servers
 - Fault-tolerant

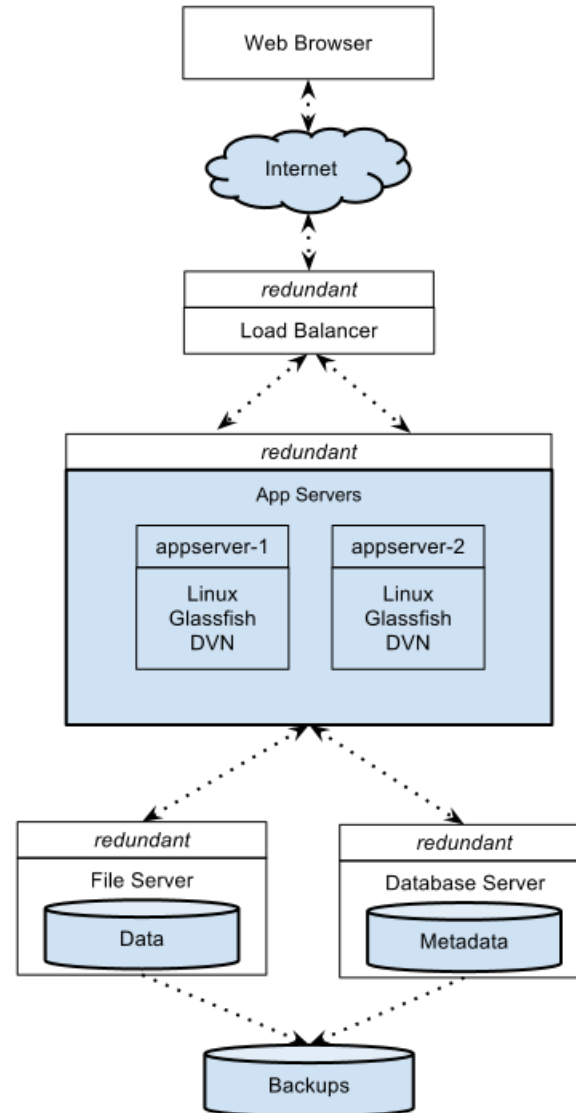
DVN Service Topology



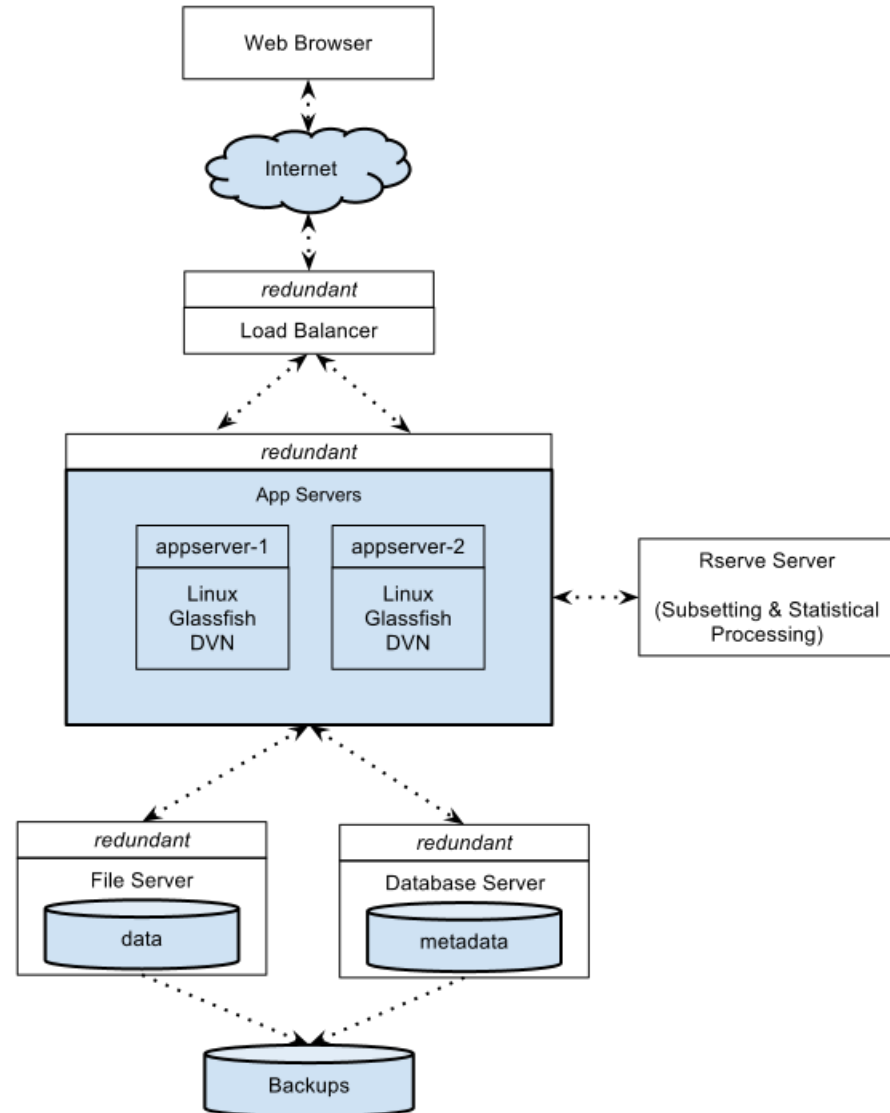
DVN Service Topology: User Interaction



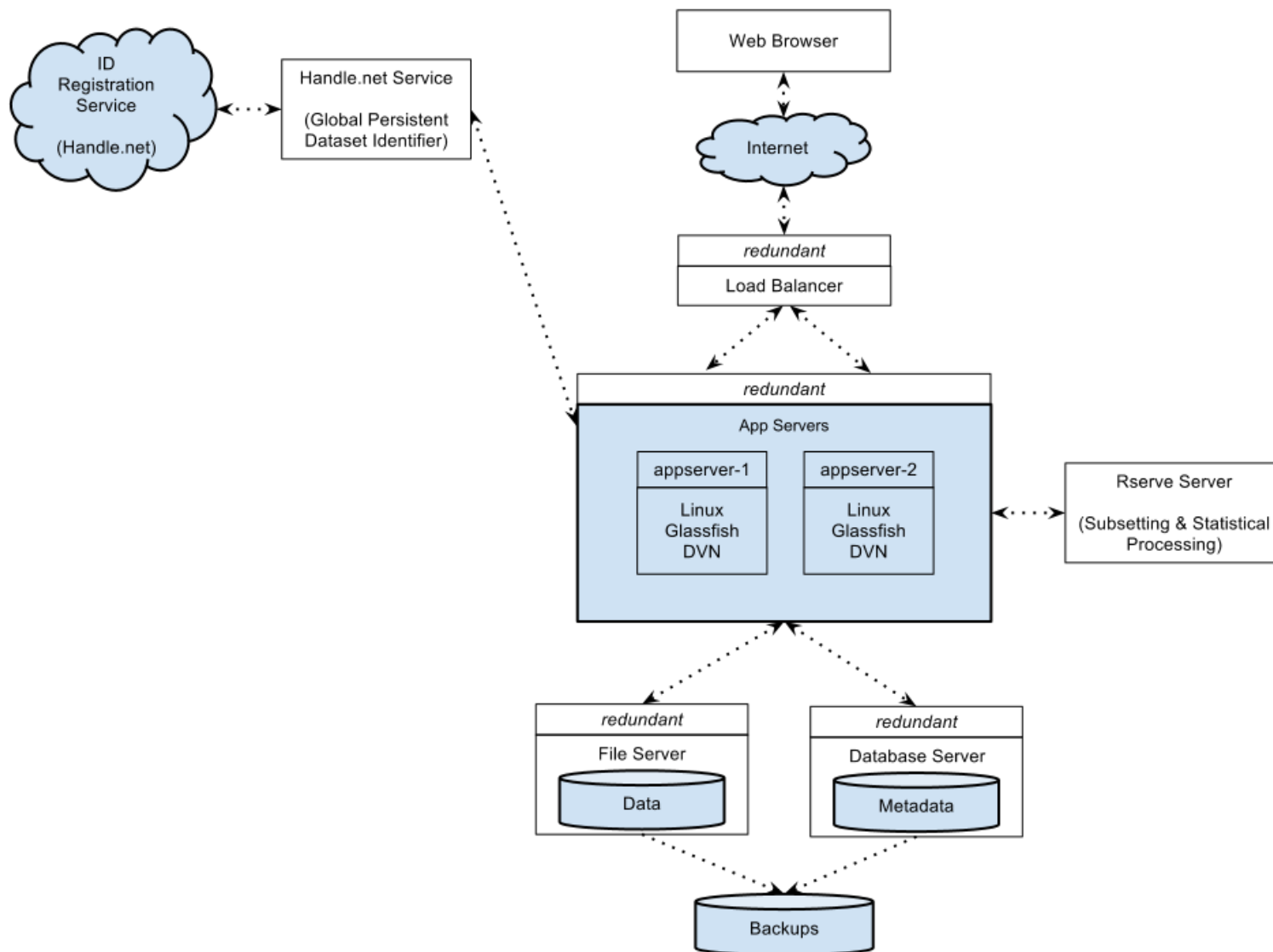
DVN Service Topology



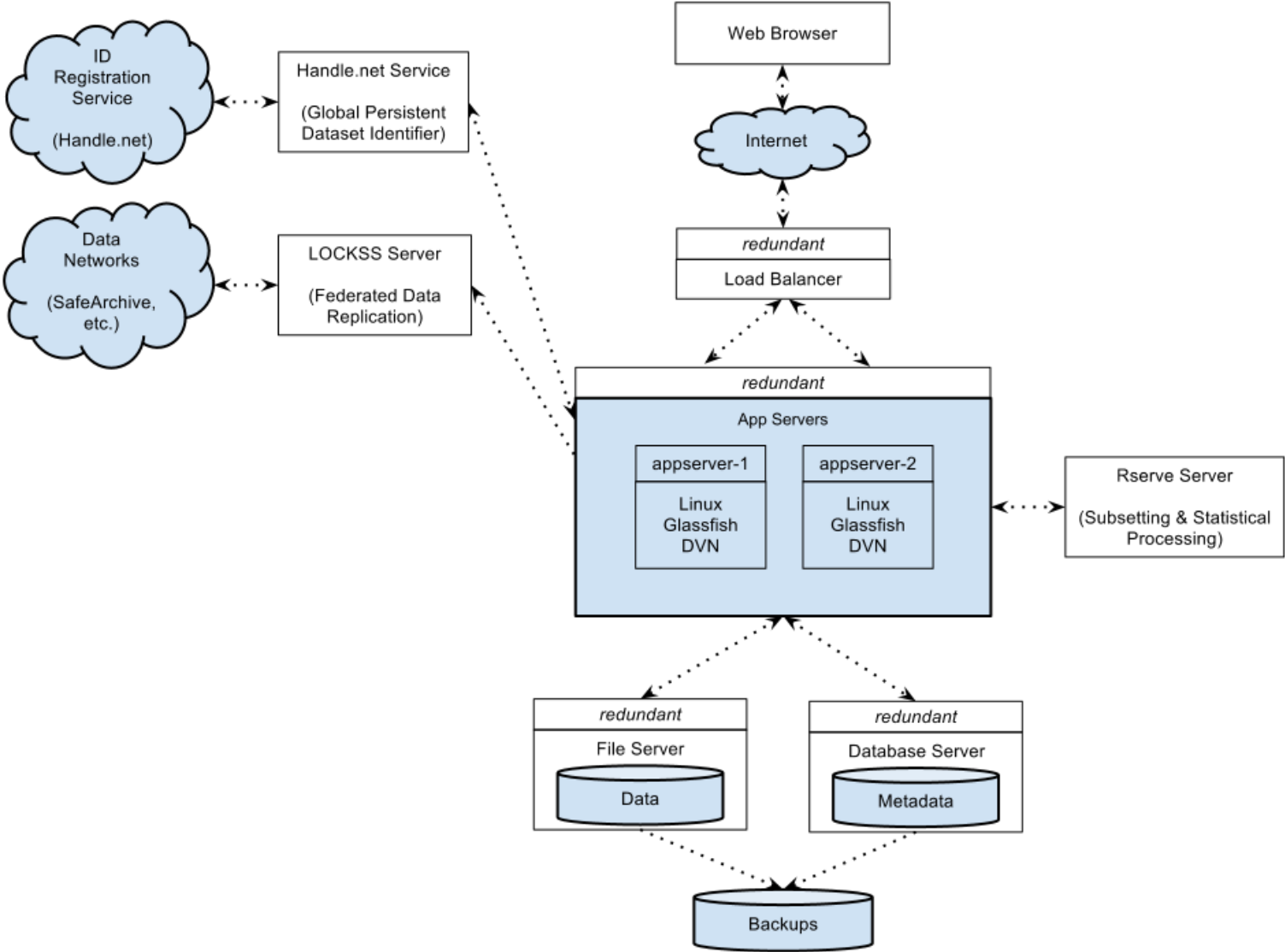
DVN Service Topology: Data Processing



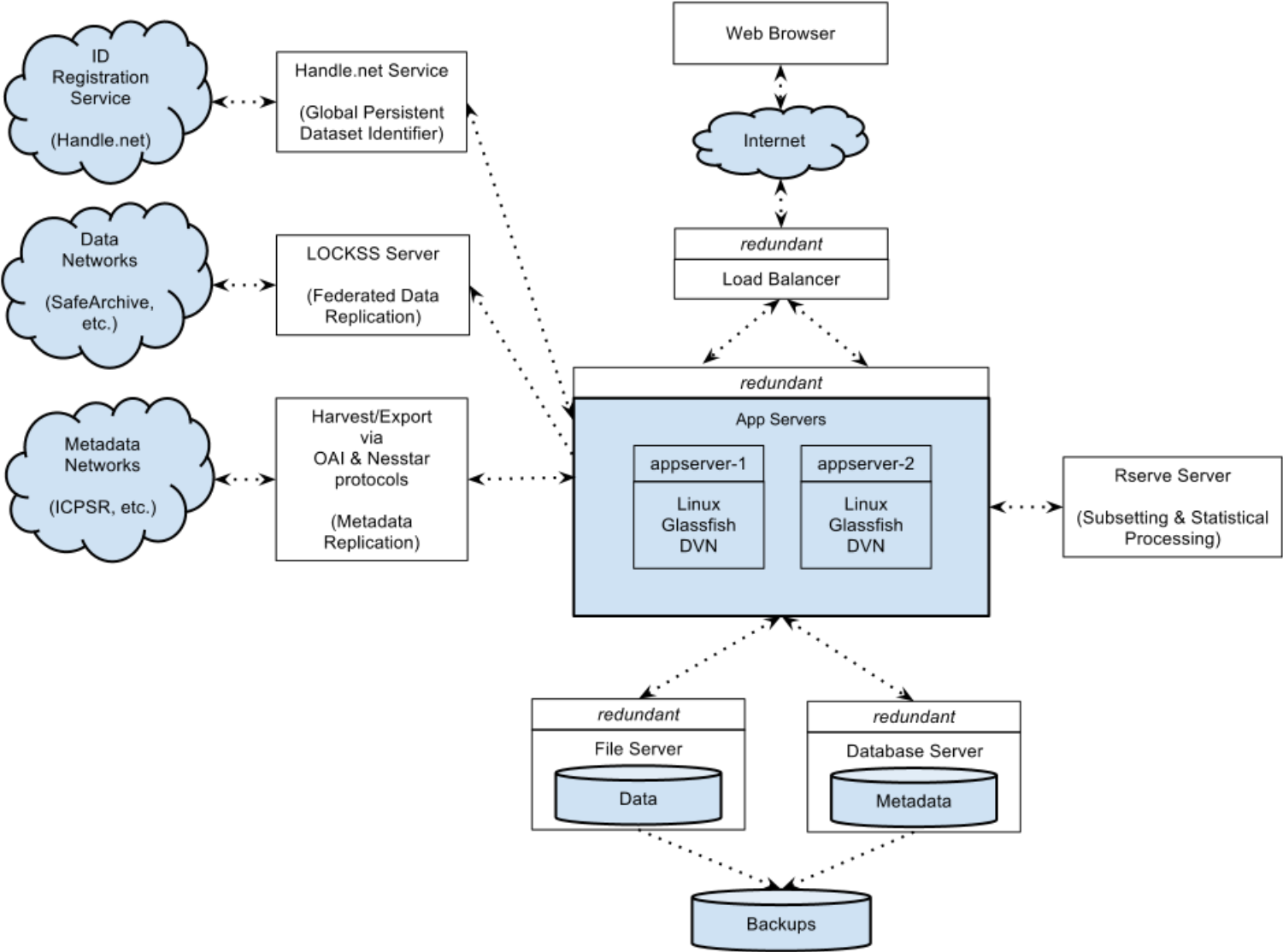
DVN Service Topology: Connections



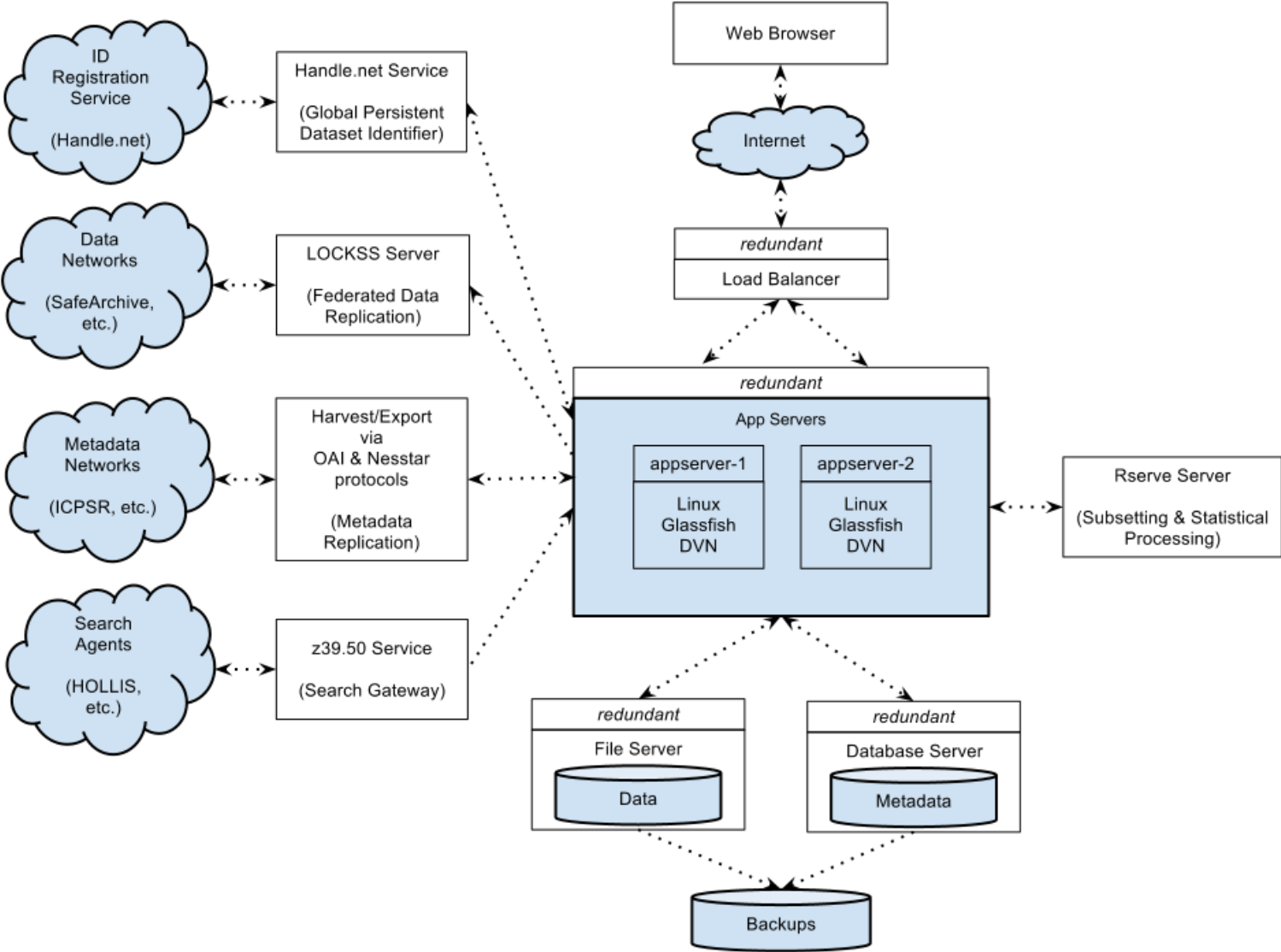
DVN Service Topology: Connections



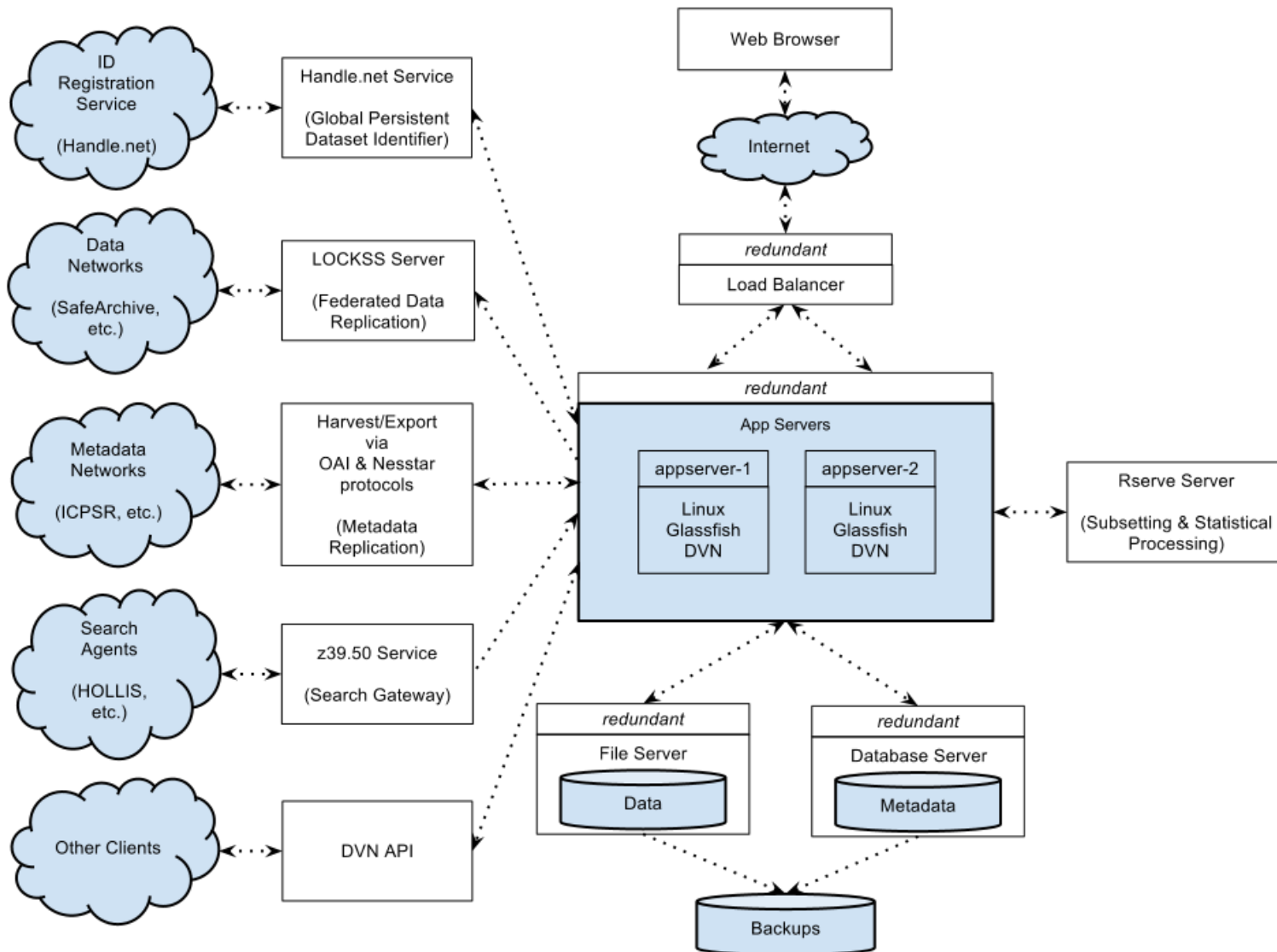
DVN Service Topology: Connections



DVN Service Topology: Connections



DVN Service Topology: Connections



Thank you.

The DVN Project Site

<http://thedata.org>

The Social Sciences DVN

<http://dvn.iq.harvard.edu>

The Astronomy DVN

<http://theastrodata.org>

Wendy Gogel | Gustavo Durand | Bill Horka

| May 31, 2012