The **Dataverse Network** ™ Project

The Dataverse Network Project is housed at the *Institute for Quantitative Social Science* (IQSS) at Harvard University.

Visit us as *http://TheData.org* and read all about it.

What is it?

The Project is an *open-source software development community*, housed at the IQSS. Via web application software, data *citation standards,* and statistical methods, the Dataverse Network project increases scholarly recognition and distributed control for authors, journals, archives, teachers, and others who produce or organize data; facilitates data access and analysis for researchers and students; and ensures long-term preservation whether or not the data are in the public domain.

The structure of the Dataverse Network (DVN) software application is shown in the following figure:
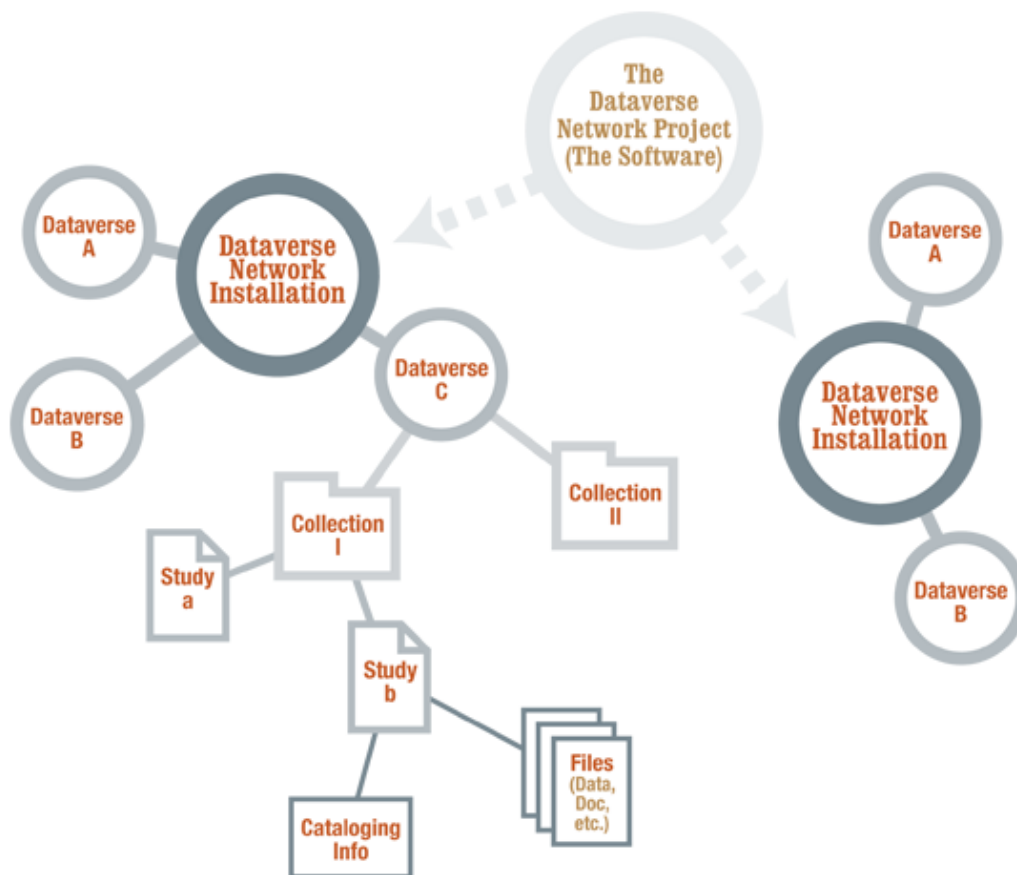
The following concepts apply to the DVN application:

☞ The *DVN* is a digital community library of research data, which enables sharing of data and the studies associated with that data. This digital library contains many collections and volumes of data. In addition, the DVN software offers value-added services to library users.

☞ A *dataverse* is a virtual collection, hosted by the DVN. That is, a dataverse is one view of the universe of data available in a DVN. A dataverse holds one or more collections of data, and can contain any number of individual volumes of data. Data can be uploaded to a dataverse, or linked to from another dataverse. The host DVN provides extensive services to each dataverse, including permanent data archiving, preservation, cataloging, dissemination, citation, searching, format conversion, subsetting, and online statistical analysis.

☞ A *study* is an individual entity, or volume, within the dataverse collection, which has one set of metadata that describes that volume. Metadata is *data about data*, and in the DVN the study metadata is called the *Cataloging Information* for the study. Each study can have any number of associated files, including documentation files and data sets. One set of metadata and all of the associated files comprise the study.

☞ If you include data files in a study that are in specific formats, that data is *subsettable* in the DVN. That is, these data are recognized and converted into preservation format, or ingested, into the DVN. If you include Stata or SPSS data files, the DVN provides services for subsetting your data, recoding it, and analyzing the data in many ways. Subsetting for other data formats is in development.

What Does the DVN Do?

The Project developed and maintains the DVN application in response to digital storage issues:

☞ Accessible, public distribution of contents

☞ Recognition for author, publisher, and distributor

☞ Persistence of data

☞ Verification of data

☞ Validation of data

☞ Authorization controls

☞ Legal protection

Accessibility

The DVN provides an interface that is easy to use. Your DVN can harvest collections and studies from other repositories that are compliant with the Open Archives Initiative (OAI), and provide access to all contents through a single interface. DVN software includes powerful browse and search features that enable users to look across all data sets, across those within a specific DVN, or only within a specific dataverse. Or users can search for a dataverse that offers a specially organized view on some relevant area. You can download studies and data with one click, or analyze data online without downloading anything.
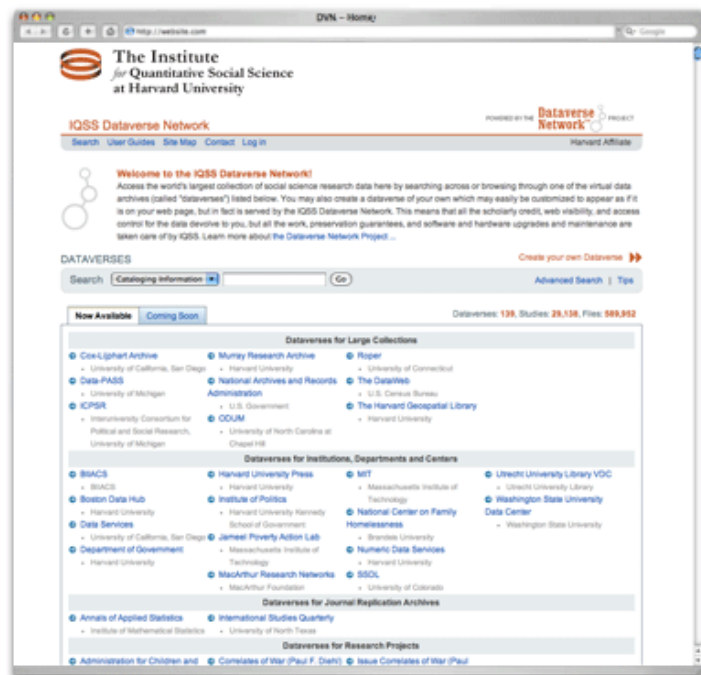
A DVN installation, like any library, can be open to the public or closed to specific membership. Anyone can use the DVN hosted by IQSS at *http://dvn.iq.harvard.edu/dvn/*, shown in the above figure. Institutions that host other DVNs determine who can use their services.

Recognition, Persistence, and Verification

When you contribute a study to a dataverse, the DVN assigns a citation to your study automatically. This citation gives recognition to study and data creators in terms of web visibility (within their own dataverse) and scholarly citation credit (through the combination of our data citation standard and links to the authors print publications).

This citation comprises several components: authors, year, title, and unique identifiers. Citations also can have optional features in a standard format, such as "Murray Research Archive [distributor]." With this citation, all participants in a study are credited and recognized.

A unique, global, persistent identifier is part of the citation for every study. When you contribute subsettable data to a study, the DVN also generates a *Universal Numerical Fingerprint* (UNF) identifier for the study. Each subsettable data set also receives a UNF, and the data is preserved in a format-independent state by the DVN.

The DVN hosted by the IQSS assigns a unique global identifier that complies with the international *Handle* System, in the form *hdl*. This identifier is designed to persist even if Uniform Resource Locators (URLs) no longer are used. A handle is resolved into a URL, but if the URL changes in the future the handle remains the same.

The UNF component is a string of letters and numbers computed from the data set that uniquely identifies its content. It guarantees that users can verify that data retrieved is identical to data used earlier. If the data change, the UNF changes. Four features make the UNF especially useful:

☞ The UNF algorithm's cryptographic technology ensures that the alphanumeric identifier changes when any portion of the data set changes.

☞ The UNF is determined by the content of the data, not the format in which it is stored.

☞ Knowing only the UNF, researchers can be confident that they reference a specific data set that cannot be changed, even if they do not have permission to see the data.

☞ The UNF's noninvertible, cryptographic properties guarantee that acquiring the UNF of a data set conveys no information about the content of the data. Authors can take advantage of this property to distribute the full citation of a data set—including the UNF—even if the data is proprietary or highly confidential, all without the risk of disclosure.

For example, the following citation includes a Handle.net ID and a UNF, is available in the ICPSR dataverse at the IQSS DVN, and contains data from a seminal study conducted in 1959-60:

"Almond, Gabriel; Verba, Sidney, 1984-06-20, "Civic Culture Study, 1959-1960", hdl:1902.2/7201 UNF:3:rN4NdVDG3yDThIqzlhnCwg== Inter-university Consortium for Political and Social Research [Distributor]."
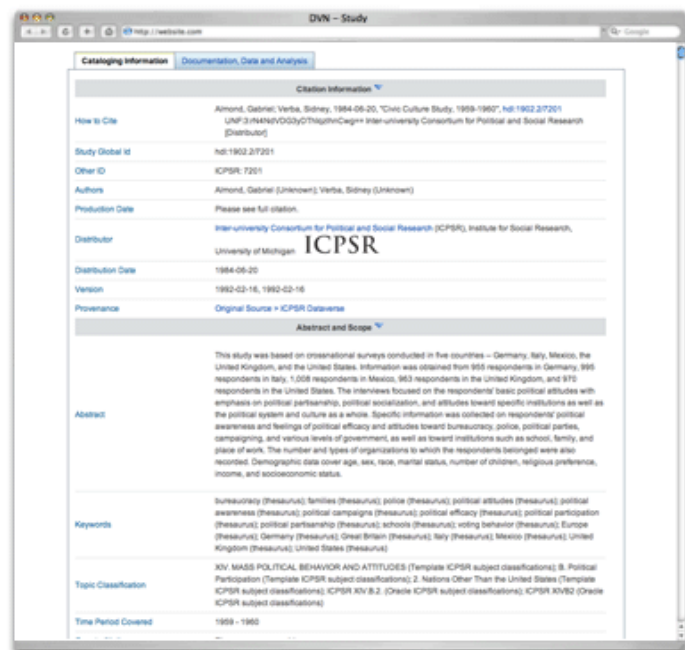
Validation

By viewing the UNF citation for specific data, users can validate that the data they choose to use is correct, without being able to see the data.

Authorization and Legal Protection

The DVN enables you to control distribution and contribution of both studies and data. You can enable or disable contributions to your dataverse. You also can enable and disable downloads of your study files.

In addition, to protect your dataverse and your contributors, you can apply Terms of Use at both the dataverse and study levels. The DVN also can apply Terms to users and contributors.

You Control Your Dataverse

A dataverse offers you a secure facility to upload data and associated documentation. You can organize your studies in collections that you find useful. And, you can enable others to share their data, too, and provide a venue for making your collected works publicly accessible and properly recognized. You set the Terms of Use for each study, or for your entire dataverse. The result is that you control your data, and no one else.

Create Dataverse

Creating a dataverse is easy. You create an account using basic information (name, email address) and provide a name for your dataverse. That is all you need to do.

Contribute Studies and Data

You can create studies in your dataverse and upload any number of files associated with those studies. Other dataverse owners might permit you to contribute studies in their dataverses, or they might create collection within their dataverses and select studies from within your dataverse to include in those collections.
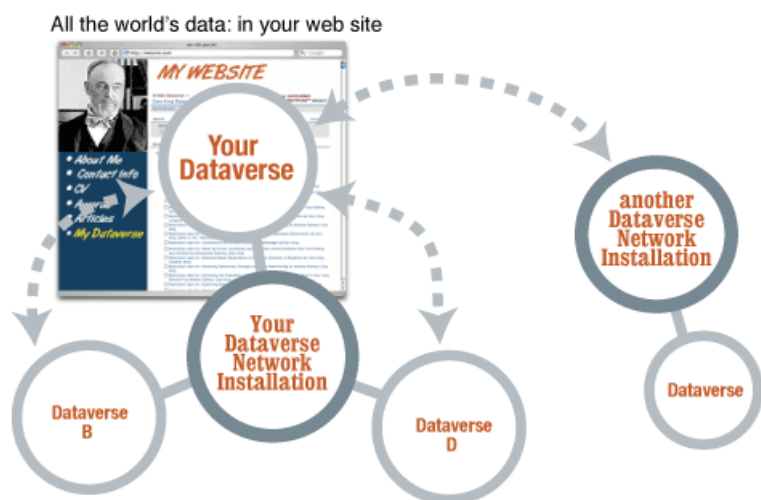
When you create a study, you enter associated metadata that comprises the Cataloging Information about the study. Metadata can include such information as the survey for which the data were collected or are otherwise connected with, how the data were obtained, and keyword summaries of its content. It also might include more detailed information, such as variable names and labels, value labels, missing value codes, data structures, and other types of extensive machine-readable documentation. You can define standards for minimal required metadata fields in your dataverse, but in general the more metadata you provide, the more services the DVN can offer to other researchers. Additional metadata improves searchability of your study (search is based on metadata) so other users can more easily find and understand your study. You can make your metadata public but restrict access to the data files.

After uploading, the DVN assigns to each study a formal citation to these data, to be used in your study results and to be cited in other publications.

Collect Studies

You decide which data to include in your dataverse: only data sets that you produced, any data associated with your community, or links to collections of data available in other dataverses.

For example, if your web site is for a project, research center, or other entity, the main page of your dataverse could offer an expandable list of projects, subjects within projects, issues within subjects, surveys within issues, and data sets corresponding to specific surveys.

All the world's data: in your web site

You can create collections of studies within your own dataverse, or of studies from other dataverses. There are three ways to define a collection:

☞ Select individual studies for a collection.

☞ Define a query and assign the results to a collection.

☞ Creating a link to existing collections from another dataverse.

Brand Dataverse

You can brand your dataverse as your own, with the same look and feel as your web site. Copy your style information or HTML from your personal web site, and paste that code to your dataverse header and footer. Then link to your dataverse from your main web page, and from your main web page to your dataverse. The result is two web sites that appear as one, but your web site is hosted by your provider, and your dataverse is hosted by the DVN. You leverage the services and support of the DVN to incorporate your data into your web site, without having to install or back up anything.


Gary King's website

Gary King's dataverse

Set Terms, Restrictions, and Privileges

There are three types of controls and protections that you can assign to different functions in your dataverse:

☞ Terms of Use – Require a user to accept Terms when uploading or download in your dataverse. These Terms support legal protections on the use of your data.

☞ Restrictions – Restrict users from accessing either studies or data sets, based on user account. You can restrict all users or enable selected users to access your material.

☞ Privileges – Allow users to contribute their data to your dataverse, or to upload or download in your dataverse, if you choose. You also can create user accounts and assign these privileges to those users.

Your DVN or Ours?

Institutions and large organization can install the DVN software to host shared-access digital library services, and to control access to their library's contents. In addition, a DVN can harvest studies from other DVNs. Community members use these digital storage and archive services to view and create dataverses and studies. A DVN provides all the services of a digital repository: ease of use, recognition, persistence, preservation (verification), validation, access control, and legal protections.

A Dataverse Network (DVN)
may harvest studies from other DVNs
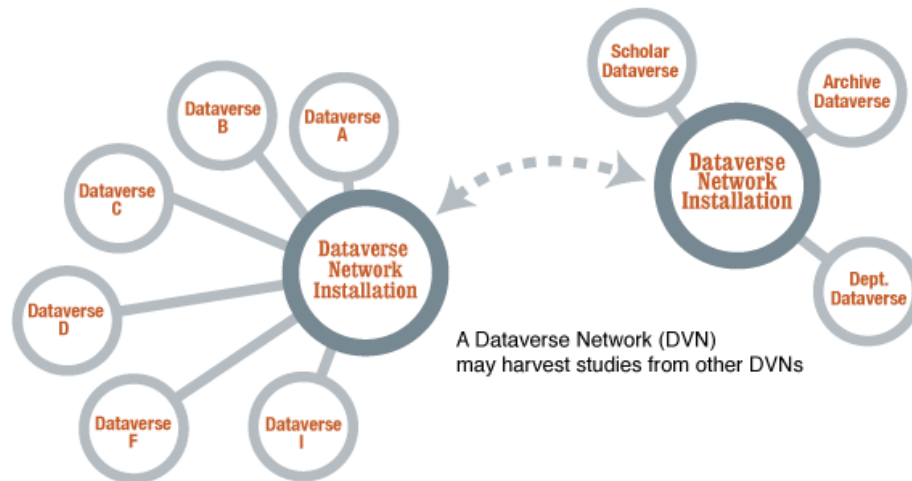
Access Controls

Each DVN has infrastructure to identify members, such as by Internet Protocol (IP) address, or through pin servers or passwords. Hosts can subscribe to data services and offer them to selected members, and make metadata available more widely so others can see what is available and request access.

Search Services

A DVN provides the following types of Search services:

☞ Search within all Network contents, in a dataverse only, or in specific collections.

☞ Search for terms in study metadata (title, author, abstract, time period, etc).

☞ Or, search for data set variable metadata (data labels and data names).

Data Subset and Analysis Services

When a study includes *subsettable* data sets, or data sets in SPSS and Stata formats, many services are available to users to study these data:

☞ Subset the data by observation (such as women 18–24 who voted for Bill Clinton in Massachusetts in 1992) and by variable (such as vote preference, unemployment status, and partisan identification).

☞ Run descriptive statistics and graphics for the data set or a subset.

☞ Perform sophisticated statistical analyses using R with Zelig (Imai, King, and Lau 2006, *http://gking.harvard.edu/zelig*), online, for the data set or a subset.

☞ Translate data to a format readable by your favorite statistical software, database program, or spreadsheet package.

☞ Download the data set, the subset, and the analysis.

The DVN incorporates Zelig statistical software, which provides a common framework for statistical analysis and software development, and a unified syntax for using statistical methods implemented in R (see R Project for Statistical Computing, http://www.r-project.org). This means that any method in Zelig also can be run by anyone through a dataverse, without knowing R. These models all can be run online, without having to download data or install, learn, or run your own statistical software.

The Project is developing support to provide subsetting and analysis functions for other data formats.

Try the IQSS DVN

At the IQSS DVN, you can access existing data, create your own dataverse, or contribute your data in the Murray Research Archive. You can download data from any available dataverse and use it in your own applications, or use the DVN subsetting and analysis tools online.

Anyone can create a dataverse at IQSS DVN:
1. Go to *http://dvn.iq.harvard.edu/dvn/* and click the *Create your own Dataverse* link.
2. Create an account using your first and last name, email address, and username and password.
3. Name your dataverse, and provide a short alias to use in the URL for your dataverse.
   ☞ Click Save, and you see a message that lists the URL for your dataverse and provides links to the user guides for assistance in uploading data, customizing your dataverse, and administrating other functions.

Or, you can contribute a study to the Murray Research Archive dataverse:
1. Go to *http://dvn.iq.harvard.edu/dvn/dv/mra* and click the *Become a Contributor* link.
2. Create an account using your first and last name, email address, and username and password.
   ☞ Click Save, and you see a message that provides links to the user guides for assistance in uploading data and viewing your studies.

IQSS DVN Data Backups
The IQSS and Harvard University guarantee to maintain all contributions and keep each available at the handle identifier assigned to it. Harvard guarantees the *data backup terms* for the IQSS DVN.

IQSS DVN Contributors
Data currently in the IQSS DVN includes the following:

☞ Large collections – Odum, ICPSR, Roper, Murray Research Archive, NARA, The DataWeb

☞ Institutions, Departments, and Centers – Harvard Department of Government, MIT Jameel Poverty Action Lab, University of CO SSDL

☞ Journals – International Studies Quarterly, Annals of Applied Statistics

☞ Research Projects – US Administration for Children and Families, Correlates of War

☞ Scholars -- Universities of Essex (UK) and Tsukuba (Japan), MIT, Stanford, USC, CIT, UCSD, Columbia, Cornell, Dartmouth, Princeton, and Johns Hopkins

Community

The Project is a member of a large community of open-access partners. Currently, we harvest data from and export to other repositories: ICPSR, ODUM, Roper, National Archives, US Census, and more. International partnerships are underway with NESSTAR and the UN Universities. Dataverse owners include Chinese, Mexican, Japanese, and French participants. Our community of developers is active and growing. See our *Partners* page on TheData.org

Architecture

DVN software is written under the *Java Platform, Standard Edition* (Java SE), using the latest Java technologies, including Enterprise Java Beans (EJB) and Java Server Faces. It runs on top of the *GlassFish Application Server*. We use the *PostgreSQL database management system* software, but you can use other databases easily, such as Oracle or MySQL. The data analysis component uses *R* and Zelig for statistical computing.

The following figure depicts the DVN architecture.
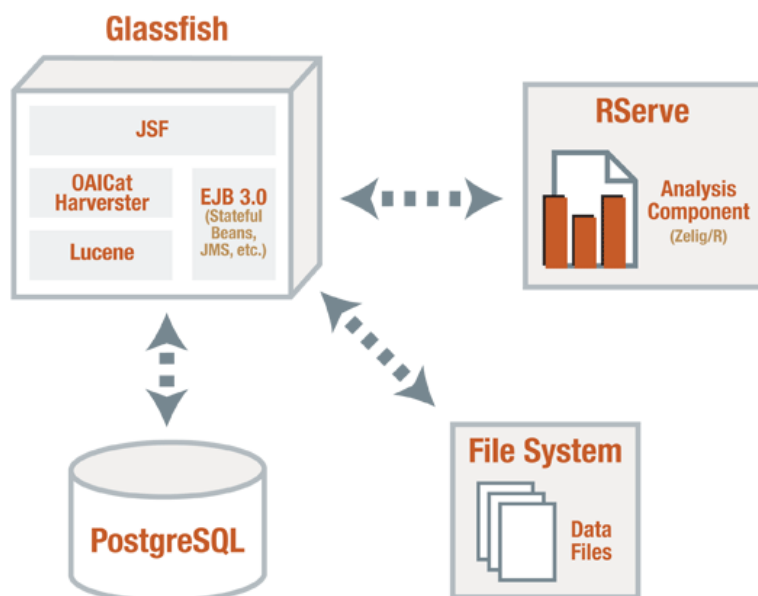
DVN software is open source, and is licensed under Affero, a version of General Public License. This license guarantees that the software is available now and that all future versions of it, including any versions created from it by others, remain available. Under this license, you have full access to all versions of the source code and a permanent license to use it and build on it as you see fit. All development is open to contributions through use of the *SourceForge.net development web site.*

Current and Future Features

☞ OAI support – Harvest metadata from one DVN to another. IQSS harvests from Odum, ICPSR, others, and exports to any OAI-compliant repository.

☞ Z39.50 support – Available at IQSS DVN, can be implemented for other DVNs.

☞ Handle.net support – Available at IQSS DVN, studies are registered with Handle.net and receive hdl identifier. Can be implement for other DVNs.

☞ Batch upload support – DVN supports upload of study files in batch format.

☞ Shibboleth support (coming soon) – Remote authorization enhancements coming.

☞ Web services support for data entry (coming soon) – DVN integrating services with other systems.

☞ Geo-referenced studies support and geo-spatial tools (coming soon) – DVN services to include map viewer for geo-related studies.

References

Refer to these sources for more information.

Learn More

Gary King. "An Introduction to the Dataverse Network as an Infrastructure for Data Sharing," *Sociological Methods and Research*, 32, 2 (November, 2007): 173--199, (Abstract: HTML | Article: PDF).

Resource URLs

☞ The Dataverse Network Project – *http://thedata.org*

☞ Partners page – *http://thedata.org/partner*s

☞ IQSS DVN – *http://dvn.iq.harvard.edu/dvn/*

☞ Murray Research Archive dataverse – *http://dvn.iq.harvard.edu/dvn/dv/mra*

☞ IQSS DVN data backup terms – *http://thedata.org/get_started/terms/backup*

☞ Gary King's dataverse and website – *http://dvn.iq.harvard.edu/dvn/dv/king* and *http://gking.harvard.edu/*

☞ Zelig, an R-based analysis tool – *http://gking.harvard.edu/zelig*