

# Persistent IDs throughout Dataverse

Gustavo Durand  
Technical Lead, Dataverse Project  
IQSS, Harvard University

PIDapalooza  
November 10, 2016

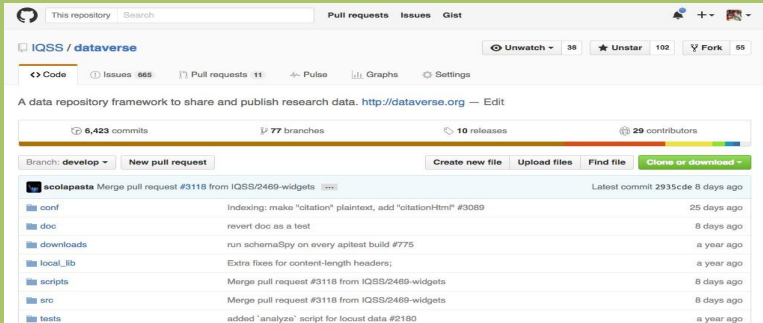
Software framework for publishing,  
citing and preserving research data  
(open source on [github](#) for others to install)

**Provides incentives for researchers to share:**

- Recognition & credit via data citations
- Control over data & branding
- Fulfill Data Management Plan requirements



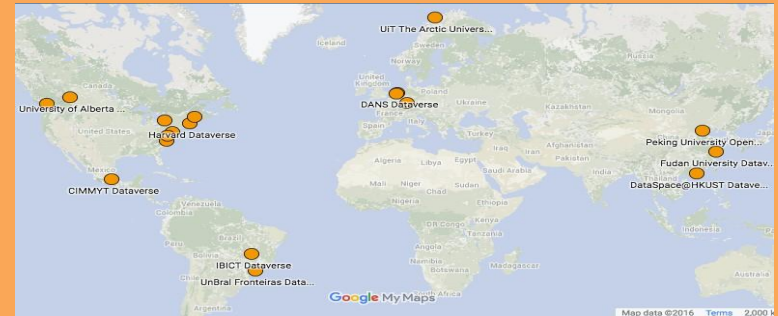
## Open Source Software Since 2006



The screenshot shows the GitHub interface for the `IQSS/dataverse` repository. At the top, it displays navigation links for Pull requests, Issues, and Git. Below this, the repository name `IQSS/dataverse` is shown along with statistics: 38 Unwatch, 102 Unstar, and 55 Fork. The repository description reads: "A data repository framework to share and publish research data. <http://dataverse.org> — Edit". It also shows 6,423 commits, 77 branches, 10 releases, and 29 contributors. A table of recent commits is visible, with the latest commit by `scolopasta` on 8 days ago, titled "Merge pull request #3118 from IQSS/2469-widgets".

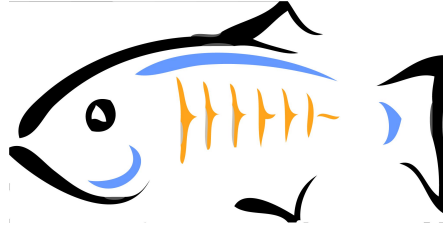
20+ data repositories worldwide

Example: Harvard Dataverse open to researchers, journals and research institutions worldwide to deposit data.



# Dataverse Technology

**Glassfish Server 4.1**



**Java SE8**

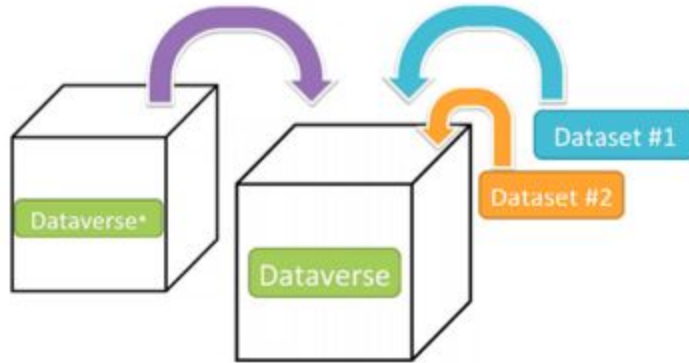
**Java EE7**

- Presentation: JSF (PrimeFaces), RESTful API
- Business: EJB, Transactions, Asynchronous, Timers
- Storage: JPA (Entities), Bean Validation

**Storage:** Postgres, Solr, File System

# What is a Dataverse or Dataset?

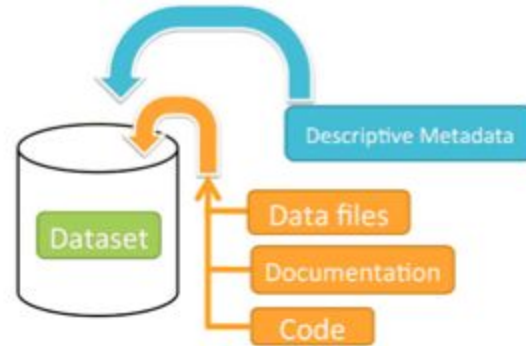
Schematic Diagram of a **Dataverse** in Dataverse 4.0



Container for your **Datasets** and/or **Dataverses\***

\* Dataverses can now contain other Dataverses (this replaces Collections & Subnetworks)

Schematic Diagram of a **Dataset** in Dataverse 4.0



Container for your data, documentation, and code.

Image created by: Eleni Castro

# Data Citation in Dataverse complies with the **Data Citation Principles**

Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014

Altman, Crosas, The Evolution of Data Citation: From Principles to Implementation, IASSIST Quarterly; 2013

# Data Citation Generated by Dataverse

Authors

Published Year

Export Formats  
for users

## Correlates of Father Participation in Family Work, 1979-1981

Rosalind C. Barnett; Grace K. Baruch, 2007, "Correlates of Father Participation in Family Work, 1979-1981", <http://hdl.handle.net/1902.1/00620>, Harvard Dataverse, V3

If you use these data, please add this citation to your scholarly resources. Learn about [Data Citation Standards](#).

Download Citation ▾

EndNote XML  
RIS Format

Persistent Identifier:  
Handle or DOI

Repository  
Name

Dataset Title

Version

## Increasing Food Security and Farming System Resilience in East Africa through Wide-Scale Adoption of Climate-Smart Agricultural Practices

Winowiecki, Leigh; Laderach, Peter; Mwongera, Caroline; Twyman, Jennifer; Mashisia, Kelvin; Okolo, Wendy; Eitzinger, Anton; Rodriguez, Beatriz, 2015, "Increasing Food Security and Farming System Resilience in East Africa through Wide-Scale Adoption of Climate-Smart Agricultural Practices", <http://dx.doi.org/10.7910/DVN/28703>, Harvard Dataverse, V7

Download Citation ▾

# Persistent Identifier Resolves to Dataset Landing Page



RESEARCH PROGRAM ON  
**Climate Change,  
Agriculture and  
Food Security**



[CCAFS - Climate Change, Agriculture and Food Security Dataverse \(CCAFS\)](#) <http://ccafs.cgiar.org/>

[Harvard Dataverse](#) > [CCAFS - Climate Change, Agriculture and Food Security Dataverse](#) >  
**Increasing Food Security and Farming System Resilience in East Africa through Wide-Scale Adoption of Climate-Smart Agricultural Practices**



## Increasing Food Security and Farming System Resilience in East Africa through Wide-Scale Adoption of Climate-Smart Agricultural Practices

Winowiecki, Leigh; Laderach, Peter; Mwongera, Caroline; Twyman, Jennifer; Mashisia, Kelvin; Okolo, Wendy; Eitzinger, Anton; Rodriguez, Beatriz, 2015, "Increasing Food Security and Farming System Resilience in East Africa through Wide-Scale Adoption of Climate-Smart Agricultural Practices", <http://dx.doi.org/10.7910/DVN/28703>, Harvard Dataverse, V7

 [Download Citation](#)

If you use these data, please add this citation to your scholarly resources. [Learn about Data Citation Standards.](#)

### Description

The overall project goal is to improve food security and farming system resilience of smallholder mixed crop-livestock farmers in East Africa while mitigating climate change through wide-scale adoption of climate-smart agriculture (CSA). The project integrates interdisciplinary approaches, including participatory research, integrating a meta-analysis of CSA practices, real-time land and soil health assessments, crop suitability modelling, socio-economic appraisals and multi-dimensional trade-off analyses, as well as on-farm participatory evaluations of CSA to identify, test, implement, and outscale locally appropriate CSA practices.

### Subject

Earth and Environmental Sciences; Medicine, Health and Life Sciences; Social Sciences; Other

### Keyword

Climate Smart Agriculture, Food Security, Land Health, Soil, Socio-Economic, Adaptation, Mitigation



# The Persistent Identifier applies to the entire Dataset, not to individual Files

Files

Metadata

Terms

Versions

Search this dataset...

Find

7 Files

Download

## 0000 Increasing FS and farming system resilience in East Africa through wide-scale adoption of CSA.pdf

Adobe PDF - 438.8 KB - Apr 23, 2015 - 52 Downloads

MD5: 47a940551eed82c5f5e2e6ed9c698aab;

This document gives a brief description of the project, including the type data collected.

Please download and read this before downloading any other.

00 ReadMe

Download

## 0200 CSA-RA manual V2.pdf

Adobe PDF - 4.6 MB - Feb 2, 2016 - 0 Downloads

MD5: 9f022e49f891ec6b4a80c7dad9107f1d;

CSA-RA Manual updated including the CSA Prioritization Workshops

02 Manuals

Download

## 0300 CIAT SAGCOT CSA-RA Report.pdf

Adobe PDF - 1.8 MB - Apr 23, 2015 - 70 Downloads

MD5: 7c8a695de400ea0c40ff3ed6fac34aac;

A Climate Smart Agriculture Rapid Appraisal (CSA-RA) was carried out by CIAT in collaboration with Sokoine University of Agriculture (SUA) for the Southern Agricultural Growth Corridor of Tanzania (SAGCOT) in September 2014. The CSA-RA aimed to assess within and between district variations in farming systems, agricultural management practices, challenges for current agricultural practices, and climate vulnerability, in order to inform targeting of climate smart agriculture (CSA). The CSA-RA used key-informant interviews, participatory workshops, transect walks, farmer interviews, as well as a gender-disaggregated methods to gather information on important

Download


# The same Persistent Identifier applies to All Versions of the Dataset

Only major versions (not minor) appear in the generated data citation

Version	Description	Author	Date
<input type="checkbox"/> 7.0	Files (Added: 2; Removed: 1; Changed File Metadata: 5); <a href="#">View Details</a>	Paola Camargo	February 2, 2016
<input type="checkbox"/> 6.1	Citation Metadata: Contact (1 Changed); Additional Citation Metadata: 0; Terms of Use/Access Changed <a href="#">View Details</a>	Cathy Garlick	April 23, 2015
<input type="checkbox"/> 6.0	Files (Added: 2; Removed: 1); <a href="#">View Details</a>	Juliana Muriel Osorio	March 3, 2015
<input type="checkbox"/> 5.0	Files (Changed File Metadata: 3); <a href="#">View Details</a>	Juliana Muriel Osorio	February 18, 2015
<input type="checkbox"/> 4.0	Files (Added: 1); <a href="#">View Details</a>	Juliana Muriel Osorio	January 30, 2015
<input type="checkbox"/> 3.0	Citation Metadata: Author (2 Added); <a href="#">View Details</a>	Juliana Muriel Osorio	January 23, 2015
<input type="checkbox"/> 2.0	Files (Added: 1); <a href="#">View Details</a>	Juliana Muriel Osorio	January 22, 2015
<input type="checkbox"/> 1.0	This is the first published version.	Hector F. Tobon R.	January 16, 2015

# Citation for Quantitative (tabular) Data


Authors, Published Year, Dataset Title, **Persistent Identifier**,  
Repository Name, Version, **Universal Numerical Fingerprint**  
**(UNF)**, [File name], [var 1], [var 2], [var...]



Checksum  
independent  
of file format



Specify File in  
Dataset



Specify a subset of  
variables in Tabular  
Data File

Following: Altman, King, A Proposed Standard for the Scholarly Citation of  
Quantitative Data, D-Lib, 2007

# Dataverse – DataCite Workflow

## EZID API

1. Dataset Created in Dataverse
2. Mint DOI with status “reserved” in EZID, send citation metadata
3. Dataset published in Dataverse
4. Change status to “public” in EZID
5. New version of Dataset
6. Send updated citation metadata

## DataCite API

1. Dataset Created in Dataverse
2. Reserve local DOI in Dataverse
3. Dataset published in Dataverse
4. Mint DOI in DataCite, send citation metadata
5. New version of Dataset
6. Send updated citation metadata

# Additional Metadata in Dataverse

## Citation Metadata

- Authors
- Title
- Description
- Dates
- Contact
- Subject
- ...

## Domain Metadata

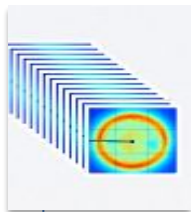
- Life Sciences: based on ISA-Tab (and OBI and NCBI taxonomy)
- Other domains (social science, astronomy)

## File Metadata

- File header metadata
- File description, type
- Variable metadata

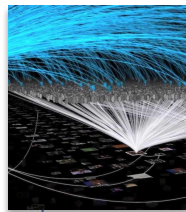
Coming Soon

# What's Coming Next

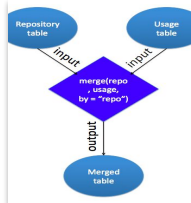


SBGrid Data Repository,  
Biomedical Dataverse (Sliz  
HMS, Crosas IQSS)

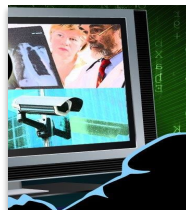
THE LEONA M. AND HARRY B.  
**HELMSLEY**  
CHARITABLE TRUST



Social Science Big Data (King,  
Crosas at IQSS)



Data Provenance (Seltzer  
SEAS, Crosas, King IQSS)



Privacy Tools to share sensitive  
data (SEAS, Berkman, Privacy  
Lab, IQSS, MIT)



# File Page

- Individual Landing Page per File
- Citation specific to the file
- Metadata contains standard file info; additionally metadata specific to that file's format
- Versions tab displays changes in metadata, in addition to reference to previous file reference (file replacement)

The screenshot shows the Dataverse interface for a file named 'turnout4.tab'. At the top, the Dataverse logo and navigation links (Search, About, Guides, Support, Sign Up, Log In) are visible. The breadcrumb trail reads: Root Dataverse > Tom Brady Dataverse > Free Samples > turnout4.tab. A progress bar shows '1 Download'. On the right, there are icons for email, share, and a 'Download' button. The file title 'turnout4.tab' is displayed above a citation box containing: 'Brady, Tom, 2016, "Free Samples", doi:10.5072/FK2/FVXYPF, Root Dataverse, V1, UNF:6:/5DHSW+eMuph53f1Ou3w==; turnout4.tab [fileName], UNF:6:r1QaJIAAIFbQGLxXI3fLg== [fileUNF]'. A 'Cite File' button and a link to 'Learn about Data Citation Standards' are also present. Below the citation, the file is described as 'Tabular Data - 32.6 KB - Last Updated: Oct 25, 2016' with '5 Variables, 2000 Observations'. A row of buttons includes 'Data', 'Playbook', 'Geospatial', and 'Time Series'. At the bottom right, an 'Export Metadata' button is visible. The 'File Metadata' section is expanded, showing a table with two rows: 'Original File MD5' with the value '21a807b044e14ae4cfc9fc5e9df22ea7' and 'UNF' with the value 'UNF:6:r1QaJIAAIFbQGLxXI3fLg=='. Navigation tabs for 'Metadata', 'Provenance', and 'Versions' are located above the metadata table.

**turnout4.tab**

Brady, Tom, 2016, "Free Samples", doi:10.5072/FK2/FVXYPF, Root Dataverse, V1, UNF:6:/5DHSW+eMuph53f1Ou3w==; turnout4.tab [fileName], UNF:6:r1QaJIAAIFbQGLxXI3fLg== [fileUNF]

Learn about Data Citation Standards.

Tabular Data - 32.6 KB - Last Updated: Oct 25, 2016  
5 Variables, 2000 Observations - UNF:6:r1QaJIAAIFbQGLxXI3fLg==

Data Playbook Geospatial Time Series

Export Metadata

File Metadata

Original File MD5	21a807b044e14ae4cfc9fc5e9df22ea7
UNF	UNF:6:r1QaJIAAIFbQGLxXI3fLg==



# ORCID Authentication

- Developed in collaboration with the SBGrid Data Repository (Harvard Medical School)
- Allows user to login via their ORCID account via OAuth
- Store ORCID iD with account
- Future ideas: allow users to pre populate dataset with ORCID iD; allow users to search for authors / contributors via ORCID iD

ORCID  
Harvard/IQSS/Dataverse  
has asked for the following access to your ORCID Record

Read limited information from your record

Allow this permission until I revoke it.  
*You may revoke permissions on your account settings page. Unchecking this box will grant permission this time only.*

This application will not be able to see your ORCID password, or other private info in your ORCID Record. [Privacy Policy](#).

Sign into ORCID or [Register now](#)

Personal Account  Institutional Account

Sign in with your ORCID account

Email or iD \*

ORCID Password

[Forgotten password?](#)

[Deny](#) [Authorize](#)

Dataverse  
Root Dataverse > Account

[My Data](#) [Notifications](#) [Account Information](#) [API Token](#)

Username tbrady12

Given Name Tom

Family Name Brady

Email tbrady12@mailinator.com ✓ Verified

ORCID ID <http://orcid.org/0000-0002-1825-0097>

Affiliation New England Patriots

Position Quarterback

# Handles

- Developed in collaboration with DANS (Data Archiving and Networked Services, Netherlands)
- Previously available in DVN 3.x
- Dataverse 4.0 released with ability to migrate datasets with handles, but not register new ones
- New Interface will allow for easier inclusion of additional Persistent Identifier services



The screenshot shows a Dataverse dataset page for Gary King. The header includes the Dataverse logo and navigation links (Search, About, Guides, Support, Sign Up, Log In). The dataset title is "GARY KING" and the creator is "Gary King Dataverse (Harvard University)" with the URL "http://gking.harvard.edu/". The dataset is titled "Replication data for: Why Are American Presidential Election Campaign Polls So Variable When Votes are So Predictable?". It has 453 downloads and a "Metrics" bar. The citation information is: "Gelman, Andrew; King, Gary, 2007, 'Replication data for: Why Are American Presidential Election Campaign Polls So Variable When Votes are So Predictable?', hdl:1902.1/SBBXEUSSCW, Harvard Dataverse, V4". A "Cite Data" button and a link to "Learn about Data Citation Standards" are also visible. The description section contains the following text:

**Description**

As most political scientists know, the outcome of the U.S. Presidential election can be predicted within a few percentage points (in the popular vote), based on information available months before the election. Thus, the general election campaign for president seems irrelevant to the outcome (except in very close elections), despite all the media coverage of campaign strategy. However, it is also well known that the pre-election opinion polls can vary wildly over the campaign, and this variation is generally attributed to events in the campaign. How can campaign events affect people's opinions on whom they plan to vote for, and yet not affect the outcome of the election? For that matter, why do voters consistently increase their support for a candidate during his nominating convention, even though the conventions are almost entirely predictable events whose effects can be rationally forecast? In this exploratory study, we consider several intuitively appealing, but ultimately wrong, resolutions to this puzzle, and discuss our current understanding of what causes opinion polls to fluctuate and yet reach a predictable outcome. Our evidence is based on graphical presentation and analysis of over 67,000 individual-level responses from forty-nine commercial polls during the 1988 campaign and many other aggregate poll results from the 1952–1992 campaigns. We show that responses to pollsters during the campaign are not generally informed or even, in a sense we describe, "rational." In contrast, voters decide which candidate to eventually support based on their

Things We're Thinking About

# Future Data Citation Extensions

- Provenance Metadata to be used in citation services
- Extended Domain Metadata (e.g., Life Sciences) to be used in citation services
- Support for Privacy, Sensitive Datasets:
  - A DataTag (blue, green, yellow, orange, red, crimson) assigned to each dataset that defines its sensitive level, with security and access requirements
- Support for Large (Streaming) Datasets:
  - Many files per Dataset. E.g., Primary Structure Dataset with thousands of images
  - Large Streaming Dataset. E.g., Geospatial Tweets

# Citations for Big Data: Large, Streaming, or Sensitive Datasets

Authors, Published Year, Title, **Persistent Identifier**, Repository Name, Version, [Subset: Query or Variable], [DataTag]

- Be able to cite entire Big Data dataset (with one Persistent Identifier), as well as specify granularity when needed
- Should the query be a RESTful url?
- Should the subset be defined by variable/attributes metadata?
- Should the DataTag be part of the citation for sensitive data?

# Persistent IDs for Files

Authors, Published Year, Dataset Title, **Persistent Identifier**,  
Repository Name, Version, **Universal Numerical Fingerprint**  
**(UNF)**, **[File name]**, **[var 1]**, **[var 2]**, **[var...]**

- Ability to provide a Persistent Identifier directly to the File Landing Page
- What should the format of the identifier be? Related to the Dataset persistent Identifier or independent?
- Should there be a Persistent ID for each file? (should this be configurable per installation, dataverse?)
- Should we use Template Identifiers?

# Persistent IDs for Versions

Authors, Published Year, Dataset Title, **Persistent Identifier**,  
Repository Name, Version, **Universal Numerical Fingerprint**  
(UNF)

- Ability to provide a Persistent Identifier directly to the cited Dataset Version
- What should the format of the identifier be? Related to the Dataset persistent Identifier or independent?
- Should there be a Persistent ID for each version (minor versus major)? (should this be configurable per installation, dataverse?)
- Should we use Template Identifiers?

# Thank You!

Gustavo Durand

**gdurand@iq.harvard.edu**

Project website: **dataverse.org**

- Community Info
- Guides
- Metrics
- Roadmap / Source Code