

# Dataverse, Journals, and Sensitive Data

Gustavo Durand

Dataverse Technical Lead / Architect

Data-PASS Pre-APSA Workshop - August 30, 2017

Dataverse

# Dataverse

- Overview, Features, and Technology
- Development Process
  - Transparency, Strategic Goals, Roadmap
- Collaborations
- Community

- **An open-source platform to publish, cite, and archive research data**
- Built to support multiple types of data, users, and workflows
- Developed at Harvard's Institute for Quantitative Social Science (IQSS) since 2006
- Development funded by IQSS and with grants, in collaboration with institutions around the world
- 15 on the core team - developers, designers, UI/UX, metadata specialists, curation manager

- Persistent IDs / URLs
  - DataCite
  - Handle
- Automatically Generated Citations with attribution
- Compliant with FAIR and data citation principles
- Domain-specific Metadata
- Versioning
- File Storage
  - Local
  - Swift (OpenStack)
  - S3 (Amazon)

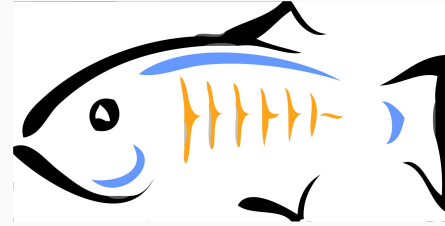
- Multiple Sign In options
  - Native
  - Shibboleth
  - OAuth (ORCID)
- Dataverses within Dataverses
- Branding
- Widgets

- Permissions
- Access Controls and Terms of Use
- Publishing Workflows
- Private URLs
- Upload / Download Workflows
  - Browser
  - Dropbox
  - Rsync (for big data “packages”)

- APIs
  - SWORD
  - Native
- Harvesting (OAI-PMH)
  - Client
  - Server



## Glassfish Server 4.1



## Java SE8

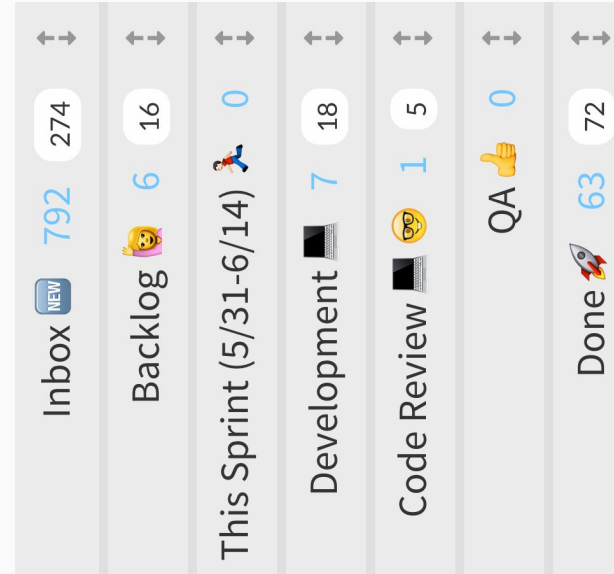
## Java EE7

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

**Storage:** Postgres, Solr, File System / Swift / S3

# Dataverse Development Process

- Inbox
- Backlog
- This Sprint
- Development
- Code Review
- QA
- Done

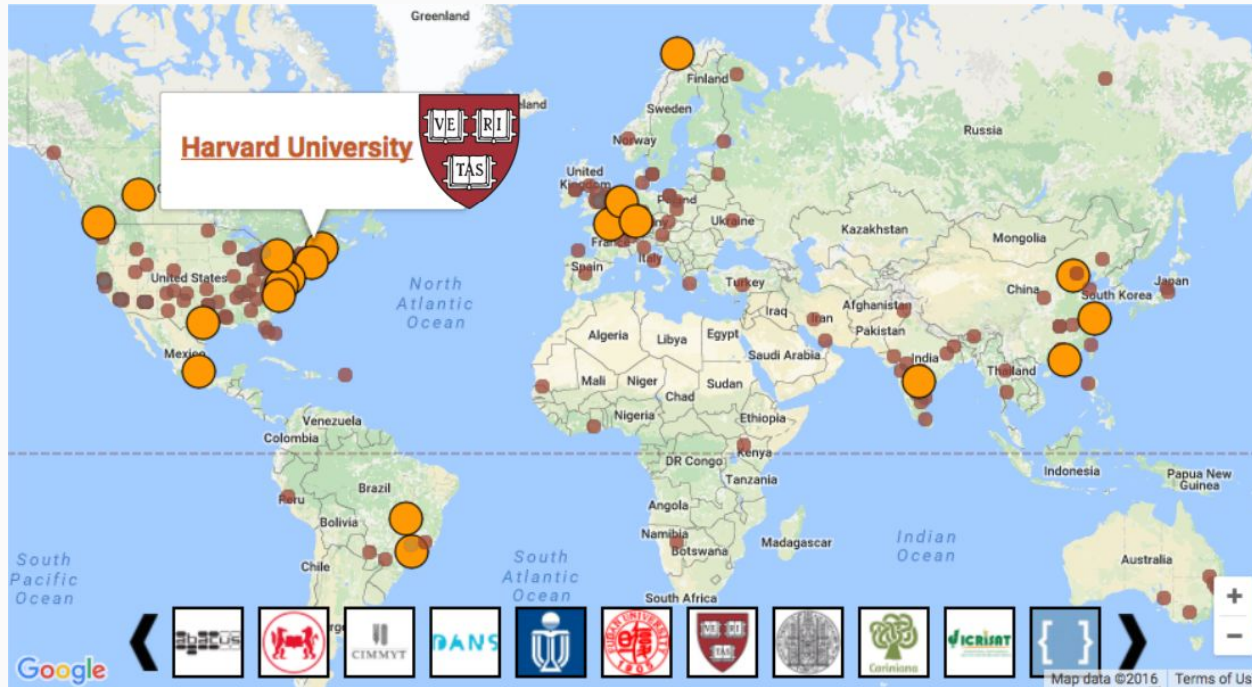


<http://dataverse.org/goals-roadmap-and-releases>

<https://waffle.io/IQSS/dataverse>

- SBGrid Data
  - Large Data and Support
- Massachusetts Open Cloud
  - Big Data Storage and Compute Access (OpenStack)
- DANS/CIMMYT
  - Handles Support
- ResearchSpace
  - API Java Client Library
- (soon) Provenance
  - W3C PROV

- 26 installations around the world



- 40+ code contributors outside of the Core Team
- Hundreds of members of the Dataverse Community - developers, researchers, librarians, data scientists
  - Dataverse Google Group
  - Dataverse Community Calls
  - Dataverse Community Meeting

# Community



# Journals

# Journals

- Overview
- Permissions
- Demo
  - Review Workflow
  - Private URLs




<https://dataverse.org/journals>

- We recommend four ways that journals can use Dataverse repositories to ensure that authors make data available and get credit for their research, with links to and from associated published articles
  - Set up a journal dataverse
  - Set up a journal dataverse with data curation & verification
  - Integrate your journal's manuscript submission system with Dataverse
  - Recommend Dataverse to authors

## Robust Permission System:

- System Roles
- Custom Roles (can be defined per installation)
- Groups
  - Explicit
  - IP
  - Shibboleth
- Inheritance
  - dataverse -> dataset
  - dataset -> file

### Roles

 All the roles set up in your dataverse, that you can assign to users and groups.

**Admin** - A person who has all permissions for dataverses, datasets, and files.

`AddDataverse` `AddDataset` `ViewUnpublishedDataverse` `ViewUnpublishedDataset` `DownloadFile` `EditDataverse` `EditDataset` `ManageDataversePermissions` `ManageDatasetPermissions` `PublishDataverse` `PublishDataset` `DeleteDataverse` `DeleteDatasetDraft`

**Contributor** - For datasets, a person who can edit License + Terms, and then submit them for review.

`ViewUnpublishedDataset` `DownloadFile` `EditDataset` `DeleteDatasetDraft`

**Curator** - For datasets, a person who can edit License + Terms, edit Permissions, and publish datasets.

`AddDataverse` `AddDataset` `ViewUnpublishedDataverse` `ViewUnpublishedDataset` `DownloadFile` `EditDataset` `ManageDatasetPermissions` `PublishDataset` `DeleteDatasetDraft`

**Dataset Creator** - A person who can add datasets within a dataverse.

`AddDataset`

**Dataverse + Dataset Creator** - A person who can add subdataverses and datasets within a dataverse.

`AddDataverse` `AddDataset`

**Dataverse Creator** - A person who can add subdataverses within a dataverse.

`AddDataverse`

**File Downloader** - A person who can download a file.

`DownloadFile`

**Member** - A person who can view both unpublished dataverses and datasets.

`ViewUnpublishedDataverse` `ViewUnpublishedDataset` `DownloadFile`

If you have a **Contributor** role in a Dataverse you can **submit your dataset for review** when you have finished uploading your files and filling in all of the relevant metadata fields.

1. To Submit for Review, go to your dataset and click on the “Submit for Review” button, which is located next to the “Edit” button on the upper-right.
2. Once Submitted for Review: the Admin or Curator for this dataset will be notified to review this dataset before they decide to either “Publish” the dataset or “Return to Author”.
  - a. If the dataset is published the contributor will be notified that it is now published.
  - b. If the dataset is returned to the author, the contributor of this dataset will be notified that they need to make modifications before it can be submitted for review again.

Creating a **Private URL** for your dataset allows you to share your dataset (for viewing and downloading of files) before it is published to a wide group of individuals who may not have a user account on Dataverse. Anyone you send the Private URL to will not have to log into Dataverse to view the dataset.

1. Go to your unpublished dataset
2. Select the “Edit” button
3. Select “Private URL” in the dropdown menu
4. In the pop-up select “Create Private URL”
5. Copy the Private URL which has been created for this dataset and it can now be shared with anyone you wish to have access to view or download files in your unpublished dataset.

Sensitive Data

# Sensitive Data

- Dataverse 5
  - Infrastructure
  - DataTags
  - PSI (Differential Privacy)

- Encrypted Transit (already supported)
- Encrypted Storage ([#4113](#))
- Require verification of e-mail address ([#3300](#))
- Complex Passwords ([#3150](#))
  - :PVMinLength, :PVMaxLength
  - :PVCharacterRules, :PVNumberOfCharacteristics
  - :PVDictionaries
  - :PVGoodStrength
- Mitigate against password guessing ([#3153](#))
- Bulk Removal of Roles / Permissions ([#4055](#))

A **datatag** is a set of security features and access requirements for file handling.

A **datatags repository** is one that stores and shares data files in accordance with a standardized and ordered level of security and access requirements.



# DataTags Levels

Tag Type	Description	Security Features	Access Credentials
Blue	Public	Clear storage, Clear transmit	Open
Green	Controlled public	Clear storage, Clear transmit	Email- or OAuth Verified Registration
Yellow	Accountable	Clear storage, Encrypted transmit	Password, Registered, Approval, Click-through DUA
Orange	More accountable	Encrypted storage, Encrypted transmit	Password, Registered, Approval, Signed DUA
Red	Fully accountable	Encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA
Crimson	Maximally restricted	Multi-encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA

# DataTags

Harvard Dataverse > Gary King Dataverse > New Dataset

Host Dataverse: Gary King Dataverse

Metadata

Dataset Template: Changing the template to: None

\* Asterisks indicate required fields

Citation Metadata

Terms

Please provide additional information about your data, including applicable restrictions required by various legal regimes.

DataTags Level

Files

For more information about supported file formats, see the [DataTags Level](#) section of the User Guides.

+ Select Files to Add

Drag and drop files here

### Edit Tags

Does your data have any legal restrictions? Learn more in the [Topics Tags](#) section or the [DataTags Level](#) section of the User Guides.

**DataTags Level** Yellow

Potentially harmful personal information. [?](#)

Yellow - Potentially harmful person...

Take a survey about your data to determine recommended restrictions.

[Take DataTags Survey](#)

**Topic Tags**

Select a tag to describe the topic(s) of data this is (data, documentation, code).

Select...

**Custom Topic Tag**

Creating a new tag will add it as a tag option for all files in this dataset.

[Apply](#)

**Tabular Data Tags**

Select a tag to describe the type(s) of data this is (survey, time series, geospatial, etc).

Select...

[Save Changes](#) [Cancel](#)

PolicyModels is a system for creating models of policies, and can be used to perform interactive interviews which yield a concrete treatment that is both human readable and machine actionable.

**Question: Please select one answer**

Did the data have any restrictions on sharing, e.g. stated in an agreement or policy statement?

Yes No

**Answer Feed**

Is there any reason why we cannot store the data indefinitely? Limiting the time a dataset could be held interferes with good science practices such as replication, and should thus be avoided whenever possible. **No** [Revisit](#)

Do the data concern living persons? **No** [Revisit](#)

**Current Tags**

**DataTags**

Code **blue** ⓘ

Assertions

Identity **noPersonData** ⓘ

Handling

DUA

TimeLimit **none** ⓘ

<https://privacytools.seas.harvard.edu/datatags>

<https://datatags.org/>

What is Differential Privacy?

$$\Pr[T(M(X)) = 1] \leq e^\epsilon \Pr[T(M(X')) = 1] + \delta, \quad \forall T.$$

**Differential Privacy** is a formal, mathematical conception of privacy preservation.

It **guarantees** that any reported result does not reveal information about any one single individual, regardless of auxiliary information.

## Private data Sharing Interface



- **upload** private data to a secured Dataverse archive,
- decide / **budget** what statistics they would like to release about that data
- **release** privacy preserving versions of those statistics to the repository
- that can be **explored** through a curator interface without releasing the raw data
- including interactive **queries**.



The budgeteer allows users to select which statistics they would like to calculate and are given estimates of how accurately each statistic can be computed. They can also redistribute their privacy budget according to which statistics they think are most valuable in their dataset.

## Census\_PUMS5\_California\_Subsample

Privacy Loss Parameters [Edit Parameters](#) ?

Epsilon ( $\epsilon$ ): 0.1000  
Delta ( $\delta$ ):  $1 \times 10^{-6}$

- puma
- sex
- age**
- educ
- income
- latino
- black
- asian
- married

age

Variable Type: Numerical ?

Mean  
 Histogram  
 Quantile

The selected statistic(s) require the metadata fields below. Fill these in with reasonable estimates that a knowledgeable person could make without having looked at the raw data. **Do not use values directly from your raw data as this may leak private information.** [Click here for more information.](#)

Lower Bound:   
Upper Bound:

Delete variable

Variable Name	Statistic	Error	Hold	?
age	Mean	0.9586 ?	<input type="checkbox"/>	

Show Epsilon Confidence Level  
(a) 0.05 ?

Reserve budget for future users

[Submit Statistics and Generate Differentially Private Release](#) ?

<https://privacytools.seas.harvard.edu/differential-privacy>

<https://privacytools.seas.harvard.edu/psi>

# Thank you!

Please get in touch with us!

Google Group, Github, IRC, Twitter - [dataverse.org/contact](https://dataverse.org/contact)

[support@dataverse.org](mailto:support@dataverse.org)

**Dataverse Community Meeting 2018**

**June 13, 14, 15 at Harvard University**