

# Sensitive Data Support in Dataverse

Gustavo Durand  
Dataverse Technical Lead / Architect  
IASSIST 2018 - May 31, 2018

# 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

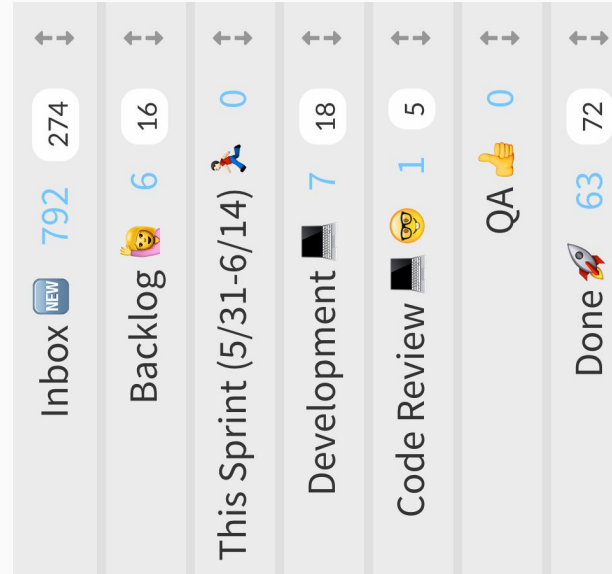
## Java EE7

- 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



<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
- Provenance
  - W3C PROV

- 33 installations around the world



- 50+ 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



Sensitive Data

# Sensitive Data

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



- Encrypted Transit
- Encrypted Storage ([#4113](#), [#4379](#))
- Require verification of e-mail address ([#3300](#))
- Complex Passwords
  - :PVMinLength, :PVMaxLength
  - :PVCharacterRules, :PVNumberOfCharacteristics
  - :PVDictionaries
  - :PVGoodStrength
- Mitigate against password guessing
- Bulk Removal of Roles / Permissions

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 in Dataverse

The screenshot shows the 'Security + Access' dialog box in the Dataverse interface. The background is a dimmed view of the 'Untitled Dataset' page, which includes tabs for 'Draft', 'Unpublished', and 'Blue' (the active tab). The left sidebar lists various dataset components: 'Citation Metadata (Re)', 'Files - All the data, doc', 'Metadata - All other m', 'Terms - CCO - "Public', 'Permissions - Who can', and 'Publish - Make this dat'. The main content area of the dialog is titled 'Security + Access' and contains the following elements:

- DataTags Security Level \***: A dropdown menu currently set to 'Blue - Unrestricted, no sensitive data or identifiable information'. A help icon (?) is to the right.
- Learn about sensitive and identifying information, please refer to the DataTags section of our [User Guide](#).**
- Need help to determine what restrictions may be appropriate for your data?**
- Take DataTags Questionnaire**: A button with a document icon and a help icon (?) next to it.
- DataTag**: A section with a dropdown menu set to 'Blue' and the text 'Non-confidential, unrestricted, accessible by anyone' with a help icon (?) to the right.
- File Access**: A section with a download icon, a dropdown menu set to 'Public', and the text 'Unrestricted file anyone off the street can download' with a help icon (?) to the right.
- Buttons**: 'Save Changes' and 'Cancel' buttons at the bottom.

At the bottom of the dialog, there is a faint line of text: 'Changing the template will delete any metadata you may have entered data into.' Below this, a dropdown menu is visible with the value 'None' and a help icon (?) to its right.

# DataTags in Dataverse

The screenshot shows the 'Security + Access' modal for an 'Untitled Dataset'. The modal is overlaid on a background showing the dataset's 'Publish Dataset To-Do List' with items like 'Citation Metadata', 'Files', 'Metadata', 'Terms', 'Permissions', and 'Publish'. The modal has a title bar with a close button. Inside, the 'DataTags Security Level' is set to 'Orange - Restricted, sensitive data or identifiable information'. A link to the 'User Guide' is provided. A message asks for help in determining restrictions, with a button to 'Take DataTags Questionnaire'. Below, the 'DataTag' is set to 'Orange' with the description 'Sensitive personal information, restricted, accessible by DUA'. The 'File Access' is set to 'Restricted' with the description 'Registered user, log in required...'. At the bottom are 'Save Changes' and 'Cancel' buttons.

**Untitled Dataset**

Draft Unpublished **Blue**

Publish Dataset To-Do List

- Citation Metadata (Re)
- Files - All the data, doc
- Metadata - All other m
- Terms - CCO - "Public
- Permissions - Who can
- Publish - Make this dat

Host Dataverse

Dataset Template

None

### Security + Access

**DataTags Security Level \*** Learn about sensitive and identifying information, please refer to the DataTags section of our [User Guide](#)

Orange - Restricted, sensitive data or identifiable information ?

Need help to determine what restrictions may be appropriate for your data?

[Take DataTags Questionnaire](#) ?

**DataTag**

Orange Sensitive personal information, restricted, accessible by DUA ?

**File Access**

Restricted Registered user, log in required... ??? ?

Save Changes Cancel

# DataTags in Dataverse

**Add Dataset** - Create an unpublished draft dataset in Gary King Dataverse. For more information on how to add a dataset, see our [User Guide](#).

## Untitled Dataset

**Draft** **Unpublished**

Publish Dataset To-Do List

- Citation Metadata (Required)
- Files - All the data, documents, and other files
- Metadata - All other metadata
- Terms - CCO - "Public Domain"
- Permissions - Who can access this dataset
- Publish - Make this dataset available

Host Dataverse

Dataset Template

\* Asterisks indicate required

Citation Metadata

Files

### DataTags Terms of Use

Learn about sensitive and identifying information, please refer to the DataTags section of our [User Guide](#).

**DataTag \*** **Orange** Sensitive personal information, restricted files accessible by DUA

By selecting the Orange DataTags Security Level, you warrant the following.

You warrant that there is no high-risk or confidential information that requires strict controls or information that the disclosure of which beyond specific recipients would likely cause serious harm (social, psychological, reputational, financial, legal) to individuals, groups or other entities.

You further warrant that there is no information commonly used to establish identity that is protected by state, federal or foreign privacy law and regulations, including but not limited to social security numbers, bank account numbers, biometric data, credit card numbers, forms containing SSN such as I9, government issued identifiers, human subject research data classified as level 4 by Harvard IRB, individually identifiable genetic, health or medical information, national security information, and trade secrets.












In addition, you warrant that:

1. The User Submission does not infringe upon the copyrights or other intellectual property rights, including, but not limited to patent, trademark, trade secret, copyright, right of publicity or other right of any third party;
2. The User Submission does not violate any relevant or applicable state, federal, or foreign laws and regulations;
3. The User Submission does not contain software viruses or any other computer codes, files, or programs that are designed or intended to disrupt, damage, limit or interfere with the proper function of any software, hardware, or telecommunications equipment or to damage or obtain unauthorized access to any system, data files, or other information of Harvard Dataverse or any third party;
4. You have obtained the all relevant, obligatory, and applicable approvals for posting the User Submission with the content included and in the format uploaded, including but not limited to approvals from the

☐ I accept this warrant

**Accept + Continue** **Cancel**

# DataTags in Dataverse

DataTags Security Level ?			
 Blue	 Green	 Yellow	 Orange
2 files	3 files	6 files	2 files
Access Requirement ?			
 Public	 Public	 Restricted	 Restricted
Access Method ?			
Public, anyone can download	Authenticated, registered users can download	Authenticated, registered users must request access and accept <i>click-through</i> Data Use Agreement, if approved can download	Authenticated, registered users must request access and accept <i>sign</i> Data Use Agreement, if approved can download
		Request Access	Disabled  ?
License Options ?			
<div>CC0 - "Public Domain Dedication" ▼</div> <div> PUBLIC DOMAIN</div> <p>Creative Commons waiver/license (CC0 is default)</p> <p><a href="#">CC0 public domain dedication</a></p>		<div>+ Upload Data Use Agreement File</div> <div></div> <div> Preview Data Use Agreement</div> <p>No waiver/license, Data Use Agreement outlines use</p> <p><a href="#">Acceptable Use Policy</a></p>	





<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**.

# PSI - Budgeteer

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 ( $\delta$ ):  $1 \times 10^{-6}$

Search variable names

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: 18

Upper Bound: 50

Delete variable

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

Show Epsilon

Confidence Level

( $\alpha$ ) 0.05 ?

Reserve budget for future users

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

PSI ( $\Psi$ ): a Private data Sharing Interface

<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**