



DataTags in Dataverse

Mercè Crosas

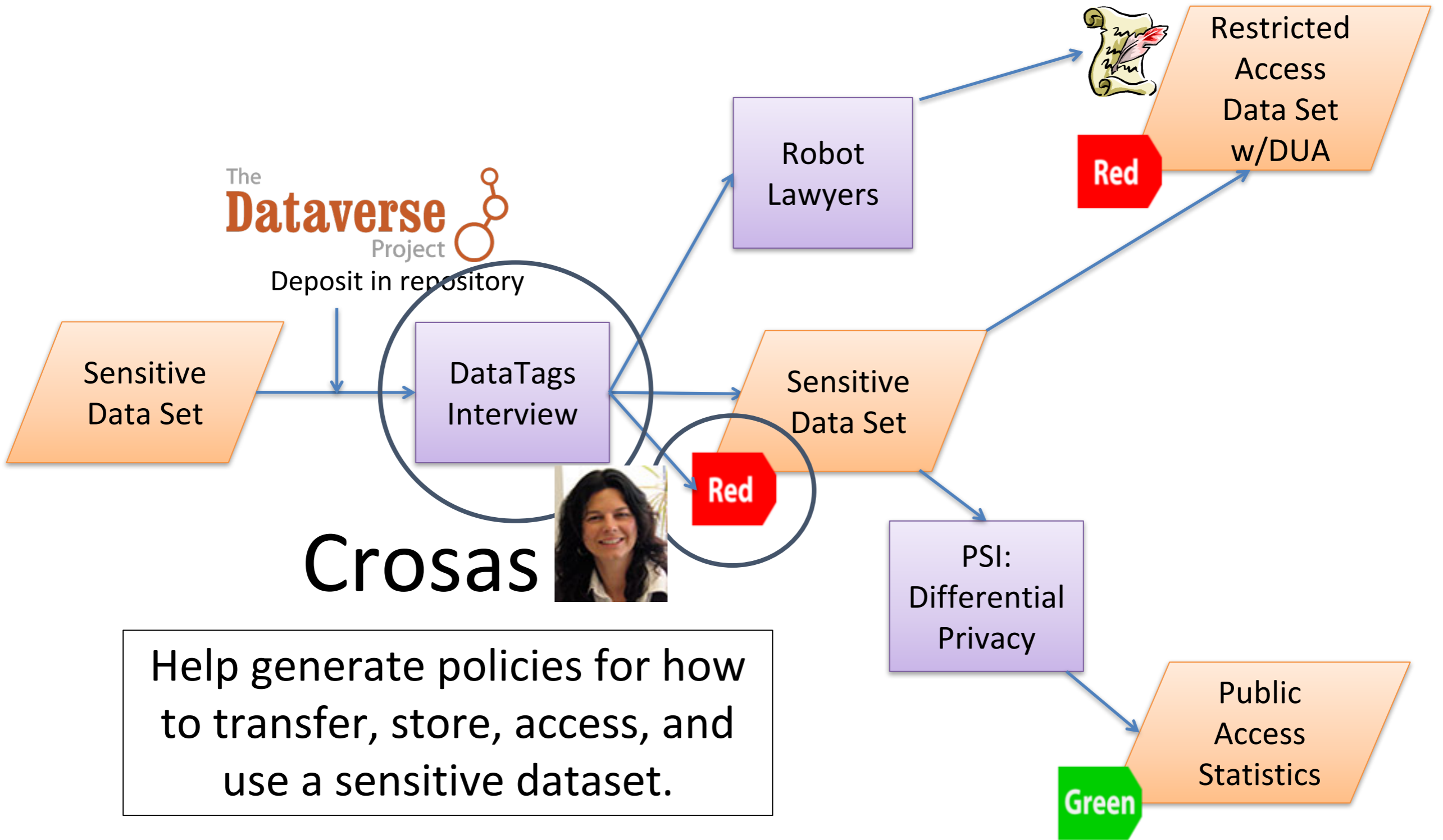
Institute for Quantitative Social Science

Harvard University

Harvard Privacy Tools Workshop, December 11, 2017

The **Dataverse** Project

Deposit in repository



Crosas



Help generate policies for how to transfer, store, access, and use a sensitive dataset.

This Talk

- Dataverse
- DataTags
- DataTags integrated with a Dataverse repository
- Workflow step by step: deposit and access
- Next Steps and Lessons learned

Dataverse

The Dataverse Project



- An open-source software platform to build data repositories
- Used to share, find, cite, and archive data
- Incentivizes data sharing by giving a data citation for each dataset with credit to data authors
- Main features:
 - Customization by branding or embedding a dataverse in your website
 - Standard and extensible metadata
 - Versioning (and soon provenance!)
 - Access control: guestbooks, file restrictions, terms of use and licensing
 - Extraction of variable (column) information from tabular data files
 - Integration with data explorations
 - APIs for depositing, searching, and accessing data and metadata
- Compliant with the FAIR principles: Findable, Accessible, Interoperable, Reusable data

Dataverse software used world-wide

- Installed in 27 sites around the world
- Each site might support an institution or the entire research community
- Sites can be federated by sharing metadata
- Has a growing and active open-source community



Harvard Dataverse repository

- Since 2007, hosts 2,400 dataverses
- 75,000 datasets, local and harvested
- 25,000 datasets with 219,000 files, deposited to Harvard Dataverse
- 40,000 files in tabular standardized format, with variable information
- 2.8 Million downloads
- 20,000 registered users
- 36,000 unique visitors per month, average for 2017
- 300 new datasets per month, average for 2017
- 50% social science research data, 13% biomedicine & health

DataTags



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 levels of security and access requirements

A DataTags Repository must satisfy the following conditions:

1. Supports more than one **datatag**
2. Each file in the repository must have one and only one **datatag**
 - a. additional requirements cannot weaken the file security
 - b. and cannot required the same or more security than a more restrictive datatag
3. A recipient of a file from the repository must:
 - a. satisfy file's access requirements,
 - b. produce sufficient credentials as requested,
 - c. and agree to any terms of use required to acquire the file.
4. Provides technological guarantees for requirements 1, 2 and 3.

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 and their respective policies

Sweeney L, Crosas M, Bar-Sinai M. [Sharing Sensitive Data with Confidence: The Datatags System](#).
Technology Science. 2015.

DataTags vs Harvard Security Levels

Blue

Level 1:

No sensitive data; open data

Green

Level 1:

Low risk de-identified data

Yellow

Level 2:

Confidential information by University standards; no material harm

Orange

Level 3:

Confidential information that could cause material harm (non-level 4 FERPA)

Red

Level 4:

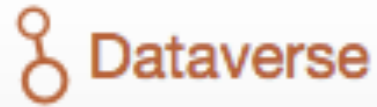
High-risk confidential information (SSN)

Crimson

Level 5* (Level 4.5, on the network)
Information that would cause severe harm

Integrating DataTags in Dataverse

Repositories prior to DataTags



Harvard Dataverse

“User Uploads must be void of all identifiable information, such that re-identification of any subjects from the amalgamation of the information available from all of the materials (across datasets and dataverses) uploaded under any one author and/or user should not be possible.”



“Submitter represents and warrants that the Content does not contain any information (i) which identifies, or which can be used in conjunction with other publicly available information to personally identify, any individual;”

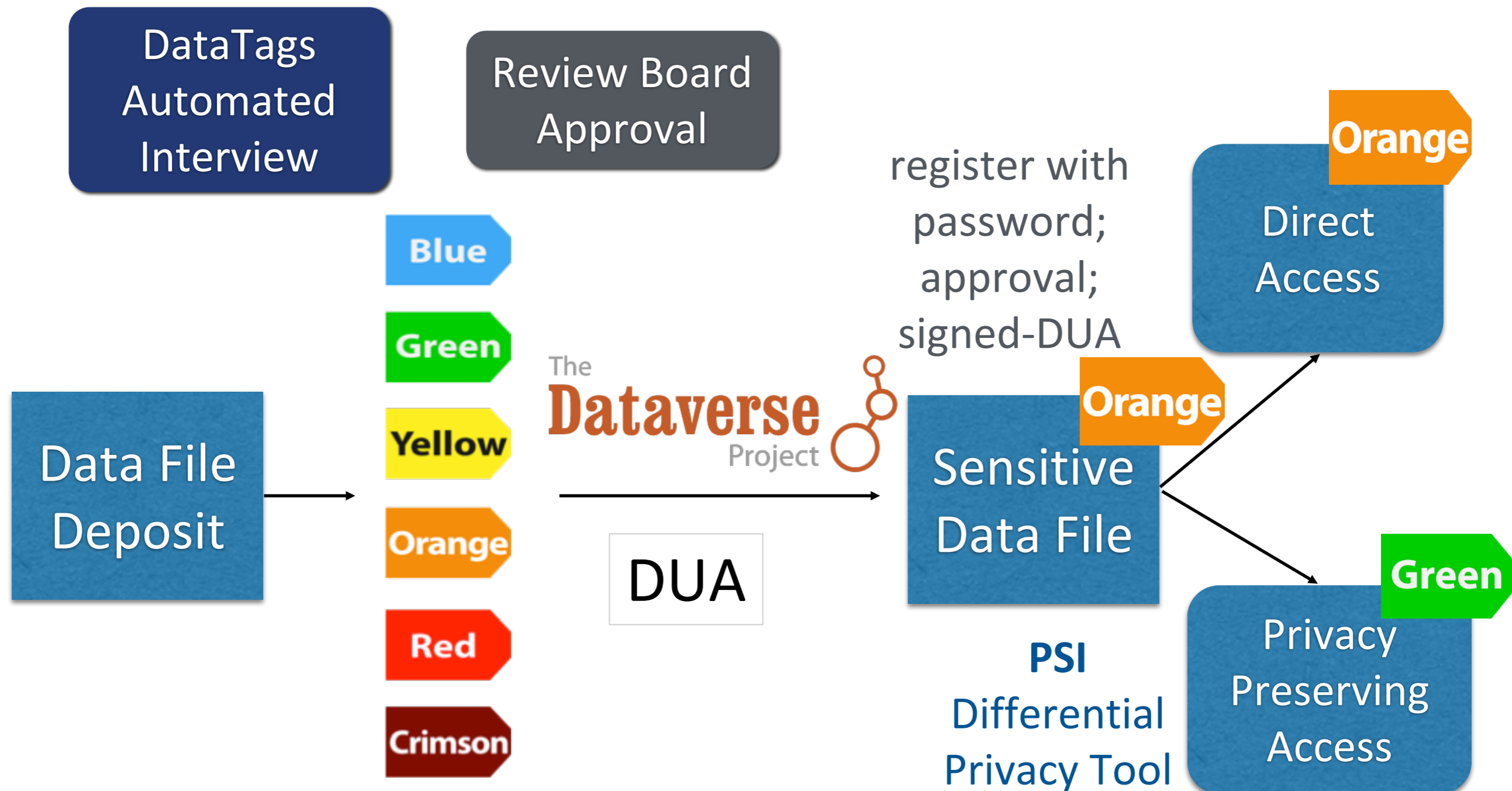


GenBank

“If you are submitting human sequences to GenBank, do not include any data that could reveal the personal identity of the source. It is our assumption that you have received any necessary informed consent authorizations that your organizations require prior to submitting your sequences.”

Repositories with DataTags

Datatags integrated with a Dataverse Repository



Step 1. Create a dataset, default no sensitive data (datatag **blue**)

Harvard Dataverse > Murray Research Archive Dataverse >

Add Dataset - Create an unpublished draft dataset in Murray Research Archive Dataverse. Add additional metadata, files, terms and permissions before publishing. ×

New Dataset

Draft **Unpublished** **Blue**

To-Do List: Before Publishing your Dataset

- Required Citation Metadata -- These are the only required fields to create a dataset citation.
- Additional Metadata -- All other metadata fields that are not required to create a dataset citation.
- Sensitive Data Level -- The DataTag applied to this dataset is **Blue**, which means it does not contain sensitive data or identifiable information.
- Files -- The data files associated with this dataset.
- Terms -- These are the terms of use, access, etc.
- Permissions -- Who can access my dataset and data files.

Host Dataverse Murray Research Archive Dataverse

Dataset Template ⓘ Changing the template will clear any fields you may have entered data into.

None

*Asterisks indicate required fields

Citation Metadata

Title *

Author *

Name *	<input type="text" value="Crosas, Mercè"/>	Affiliation	<input type="text" value="Dataverse.org"/>	<input type="button" value="+"/>
Identifier Scheme	<input type="text" value="Select..."/>	Identifier	<input type="text"/>	

Step 2. Wait, take the DataTags Questionnaire because you might have sensitive data!

Sensitive Data Level -- The DataTag applied to this dataset is **Blue**, which means it does not contain sensitive data or identifiable information. Edit

Files -- The data files associated with this dataset. Edit

Terms -- These are the terms of use, access, etc. Edit

Permissions -- Who can access my dataset and data files. Edit

Summary Files Metadata Explore Terms Provenance Versions

Crosas, Mercè, 2017, "Sample Data from the Harvard Dataverse, DRAFT VERSION" ?

[Learn about Data Citation Standards](#)

Sensitive Data Level ×

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

Sensitive Data Level *

DataTag * This dataset is **Blue**, or non-confidential, CC0, and files are unrestricted, or accessible by anyone.

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

I confirm the information above is accurate, correct and true.

Description

This study gathered information from seven Harvard University graduates in Divinity, Education, Law, and Public Health, and additional participants from the Harvard Divinity School. Data was collected by means of a mailed questionnaire, with 1,255 men and women responding.

Variables assessed from the questionnaire included: children and child-care arrangements, partner's work, and evaluations of one's own education and career as compared to other male and female colleagues. Nadelson and Notman's study of medical school alumnae (see Log# 00629) included many similar questions.

The Murray Archive holds additional analogue materials for this study. If you would like to access this material, please apply to use

Step 3: The DataTags Questionnaire ...

The image displays three overlapping browser windows from the DataTags website, illustrating the questionnaire interface. Each window shows a question, an answer feed, and an engine trace table.

Top Window: URL: `www.datatags.org/interviews/FlowChartSet-1/start`. Question: "Do the data concern living people?". Answer Feed: "This space will contain your answer when...". Engine Trace table:

Id	Type
58	ask

Middle Window: URL: `www.datatags.org/interviews/FlowChartSet-1/q/medicalRecordsCompliance`. Question: "Do the data contain health information?". Answer Feed: "Do the data concern living people?". Engine Trace table:

Id
medicalRecordsCompliance
54
53
58

Bottom Window: URL: `www.datatags.org/interviews/questionnaireId/q/MR7`. Question: "Do the data contain information from a covered entity or business associate of a covered entity?". Terms: **Business associate** (A business associate is any person or organization, including a subcontractor, that acts on behalf of, or provides services to, a covered entity involving the use or disclosure of protected health information. This includes, but is not limited to, legal, actuarial, accounting, consulting, claim processing, data analysis, administration, utilization review, quality assurance, billing, benefit management, practice management, and re-pricing activities.) and **Covered entity** (A covered entity is a health plan, health care clearinghouse, or health care provider that transmits any health information in electronic form.). List: Health plans include health insurance companies, health maintenance organizations [HMOs], company health plans, and government programs that pay for health care, such as Medicare, Medicaid, and the military and veterans health care programs; Health care providers include doctors, clinics, psychologists, dentists, chiropractors, nursing homes, and pharmacies; Health care clearinghouses include entities that process nonstandard health information they receive from another entity into a standard, i.e. standard electronic format or data content, or vice versa. Answer Feed: "Do the data contain information related to substance abuse diagnosis, referral, or treatment?". Current Tags: DataTags, Code green.

... recommends datatag **Orange**, and generates a machine-readable datatag for your dataset (and files)

www.datatags.org/interviews/questionnaireId/accept

DataTags Feedback

Dataset Can be Accepted

Your dataset is tagged as **Orange**

May include sensitive, identifiable personal information, shared with verified and/or approved recipients under agreement.

DataTags

Legal

MedicalRecords

HIPAA **safeHarborDeidentified**

EducationRecords

PPRA **protectedDeidentified** **consent**

ContractOrPolicy **no**

GovernmentRecords

DPPA **highlyRestricted**

Code **orange**

Assertions

Step 4: Select **Orange** datatag in Dataverse and agree to terms of use for that datatag

The screenshot shows a web interface for a dataset in Dataverse. A modal dialog titled "Sensitive Data Level" is open in the center. The dialog contains the following elements:

- Sensitive Data Level:** A dropdown menu is set to "Orange - Confidential information in this dataset".
- DataTag:** A text block stating "This dataset is **Orange**, or moderately confidential, and is restricted unless users are approved and have signed a DUA." Below this is a link: "Need help to determine what restrictions may be appropriate for your data?" and a button: "Take DataTags Questionnaire".
- Terms of acceptance:** A text area with the placeholder text "Terms of acceptance for DataTags...".
- Confirmation:** A checked checkbox with the text "I confirm the information above is accurate, correct and true."
- Buttons:** "Accept" and "Cancel" buttons at the bottom.

The background shows a dataset page with tabs for "Summary", "Files", and "Metadata". The dataset title is "Crosas, Mercè, 2017, 'Sample Data', Dataverse, DRAFT VERSION". There are buttons for "Share" and "Contact".

Step 5: Your dataset is tagged **Orange**; users may request access for raw data if they sign DUA

Harvard Dataverse > Murray Research Archive Dataverse >

Alumnae Study: Graduate and Professional Schools at Harvard University, 1979

Version 2.0 **Orange**

Summary Files Metadata Explore Terms Provenance Versions

Susan McGee Bailey; Barbara Burrell, 2007, "Alumnae Study: Graduate and Professional Schools at Harvard University, 1979", hdl:1902.1/00536, Harvard Dataverse, V2, UNF:3:j4wTC/d94/OzhhW0lg3LSg==

[Cite Dataset](#) [Share](#) [Contact](#)

[Learn about Data Citation Standards](#)

Metrics 5,750 Views 209 Downloads 33 Citations **Files** 57 Total Files 5 Documentation 22 Tabular 30 Data [Download All Files \(57\)](#)

Description

This study gathered information on the career development, family responsibilities, and professional standing of graduates from seven Harvard University graduate and professional schools: the Graduate School of Arts and Sciences, Dental Medicine, Design, Divinity, Education, Law, and Public Health. All degree recipients from the class of 1972 at each school were surveyed, with additional participants from the Dental School classes of 1968-1978, and the Divinity School classes of 1974 and 1976. Data were collected by means of a mailed questionnaire in the spring of 1979. Of the 3,000 eligible degree-holders, a total of 1,620 or 63% responded, including 1,255 men and 365 women.

Variables assessed from the questionnaire included educational background, employment history, career goals and job satisfaction, children and child-care arrangements, partner's work, and evaluations of one's own education and career as compared to other male and female colleagues. Nadelson and Notman's study of medical school alumnae (see Log# 00629) included many similar questions.

The Murray Archive holds additional analogue materials for this study. If you would like to access this material, please apply to use the data

Related Publication

Nevo, Aviv, and Rosen, Adam M., (2012) " Identification With Imperfect Instruments." Review of Economics and Statistics 94:3, 659-671.

Notes

Manuscript Number NCOMMS-17-15407B

Subject

Social Sciences

Keyword

electoral integrity project, elections, electoral integrity, electoral fraud, voting behavior

Preview

Table A.1: Survey Questions Used in Class (Pages 1-2)

Question	Response	Yes	No	Other
Q1	Did you work for the government?	100	100	100
Q2	Did you work for a private company?	100	100	100
Q3	Did you work for a non-profit organization?	100	100	100
Q4	Did you work for a university?	100	100	100
Q5	Did you work for a research institution?	100	100	100
Q6	Did you work for a consulting firm?	100	100	100
Q7	Did you work for a financial institution?	100	100	100
Q8	Did you work for a media organization?	100	100	100
Q9	Did you work for a government contractor?	100	100	100
Q10	Did you work for a defense contractor?	100	100	100

Step 6: Other files in the **Orange** dataset might have less restrictive datatags
















Harvard Dataverse > Murray Research Archive Dataverse >

Alumnae Study: Graduate and Professional Schools at Harvard University, 1979 Version 2.0 Orange

Summary **Files** Metadata Explore Terms Provenance Versions

Search this dataset...

1 to 5 of 57 Files Type: All Tag: All Access: All Publication Date: All Metadata Source: All

<input type="checkbox"/>		/Example/Hierarchy/Structure/README.docx MS Word (docx) 12.5 KB Aug 26, 2017 11 Downloads MD5: c5cde698dea2f75d27157311cc54c8bc Documentation Green	 
<input type="checkbox"/>		/Example/Hierarchy/Structure/Database_ACM_PRIVACY_Authors_2006_2016.csv Comma Separated Values 1.9 MB Sep 27, 2017 17 Downloads MD5: 0f1816e5450ff1ea7bea820eabd7abc1 Data Orange Dataset of privacy researchers from 2006 to 2016 extracted from the ACM digital library	 
<input type="checkbox"/>		/Example/Hierarchy/Structure/Survey Round1 - Final Questionnaire - Code Book.pdf Adobe PDF 193.7 KB Aug 26, 2017 15 Downloads MD5: 6cb397c2303cf7dfe5d59df54f47e569 Documentation Green	 
<input type="checkbox"/>		/Example/Hierarchy/Structure/Database_ACM_PRIVACY_Authors_2006_2016.csv Comma Separated Values 1.9 MB Sep 27, 2017 17 Downloads MD5: 0f1816e5450ff1ea7bea820eabd7abc1 Data Orange Dataset of privacy researchers from 2006 to 2016 extracted from the ACM digital library	 
<input type="checkbox"/>		/Example/Hierarchy/Structure/Database_ACM_PRIVACY_Authors_2006_2016.csv Comma Separated Values 1.9 MB Sep 27, 2017 17 Downloads MD5: 0f1816e5450ff1ea7bea820eabd7abc1 Data Orange Dataset of privacy researchers from 2006 to 2016 extracted from the ACM digital library	 

Step 7: Users may explore an Orange data file, with results from PSI (differentially private stats)

Harvard Dataverse > Murray Research Archive Dataverse > Alumnae Study: Graduate and Professional Schools at Harvard University, 1979 >

00536Bailey-Burrell-Law-Data.tab

Version 2.0 Data Geospatial Orange

Summary Metadata **Explore** Terms Provenance Versions

Q 116 Results Download

ID	Name	Label	Categories	Valid Cases	Missing Cases	Minimum	Maximum
196545	PanelistIdQuestion	ARF - Panelist ID		493	0	133689	138228
196539	arf_Gender	ARF - Gender	2	493	0	1	2
196519	arf_Age_Rollup_fine	ARF - Age (7 groups)	7	493	0	2	7
196579	ARF_Education_Level_DV	ARF - Education [derived variable]	8	493	0	1	8
196527	arf_HH_Income	ARF - HH Income	8	493	0	1	8

First « 1 2 3 4 5 » Last Records Per Page 5

Variables

```

    graph TD
      year((year)) --> unicode((unicode))
      unicode2((unicode2)) --> unicode
  
```

Model Estimate

Predicted Values: Y|X

Expected Values: E(Y|X)

	Estimate	SE	t-value	Pr(< t)
(Intercept)	57.8	0.774	74.7	0.00
unicode2	0.0100	2.80e-7	3.57e+4	0.00

Formula: normal(unicode ~ unicode2)

Next Steps and Lessons Learned

- DataTags up to **orange** to be integrated with **Dataverse v5.0** (Q2 2018)
- **Harvard Dataverse** repository to support DataTags once v5 is released (Harvard Security Level 3)
- **Other Dataverse repositories** will be able to support DataTags, if they satisfy hosting/archiving security requirements
- **Other repository or file systems** can use DataTags (not only for Dataverse)
- Institutions might have additional requirements than those imposed by DataTags; **review with IT**
- Liability issues require careful review of terms of use and all information text in the application; **review with OGC, IRB, Berkman/Law school**
- Balance between user-friendly, fast deposit and well-informed steps legally vetted; **lots of UX research and testing**

Thanks!

References:

Sweeney L, Crosas M, Bar-Sinai M. Sharing Sensitive Data with Confidence: The Datatags System. Technology Science. 2015. Technology Science

Bar-Sinai M, Sweeney L, Crosas M. DataTags, Data Handling Policy Spaces, and the Tags Language. Proceedings of the International Workshop on Privacy Engineering. 2016. IEEE Symposium on Security and Privacy

<http://datatags.org> (Demo of DataTags questionnaire)

<http://privacytools.seas.harvard.edu>

<http://dataverse.org>