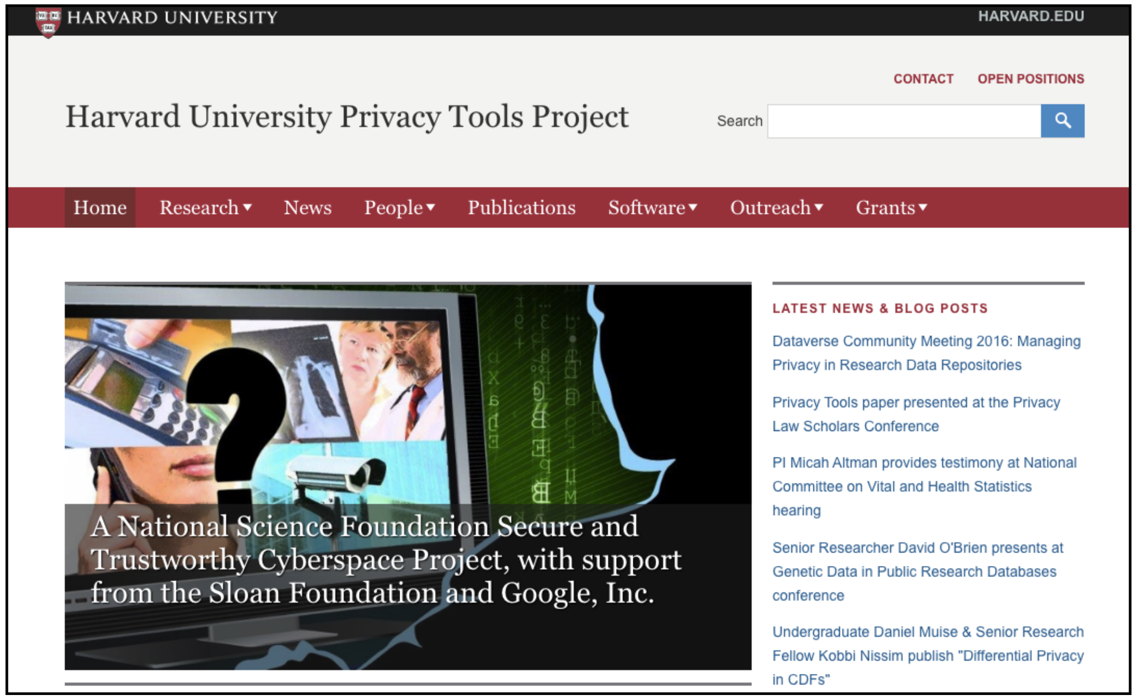
DataTags

Sharing Sensitive Data with Confidence

Mercè Crosas, Ph.D.
Chief Data Science and Technology Officer
Institute for Quantitative Social Science
Harvard University
Twitter: @mercecrosas Web: mercecrosas.com

RDA 8th Plenary, Denver, September 16, 2016 http://privacytools.seas.harvard.edu



HARVARD UNIVERSITY   HARVARD.EDU

CONTACT   OPEN POSITIONS

Harvard University Privacy Tools Project

Search

Home   Research ▾   News   People ▾   Publications   Software ▾   Outreach ▾   Grants ▾

A National Science Foundation Secure and Trustworthy Cyberspace Project, with support from the Sloan Foundation and Google, Inc.

LATEST NEWS & BLOG POSTS

Dataverse Community Meeting 2016: Managing Privacy in Research Data Repositories

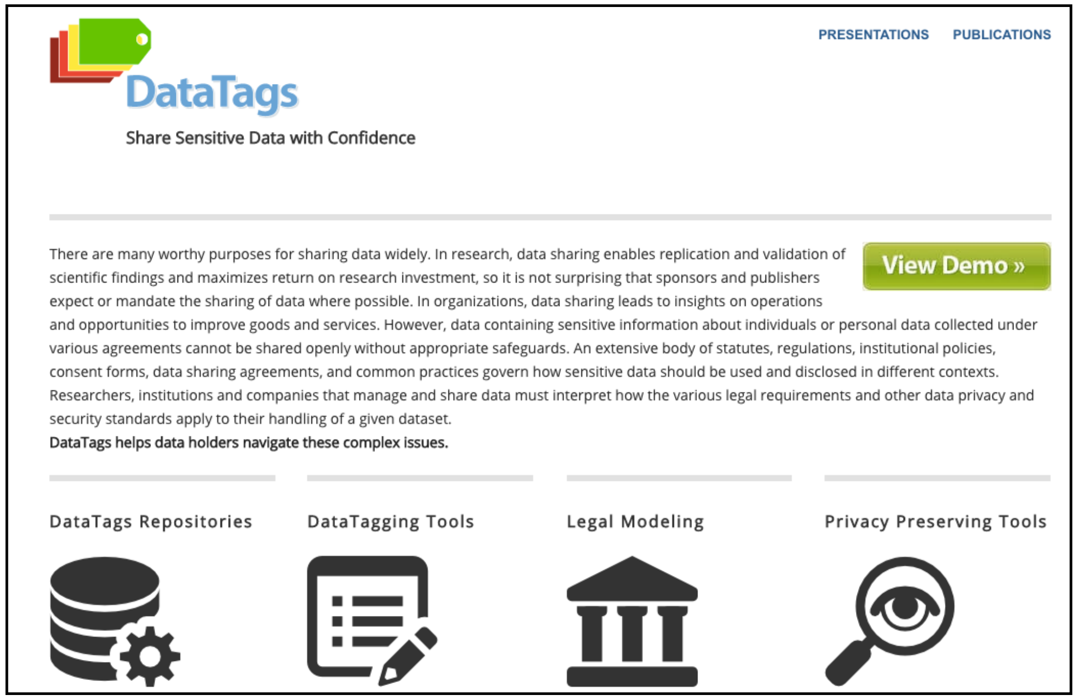
Privacy Tools paper presented at the Privacy Law Scholars Conference

PI Micah Altman provides testimony at National Committee on Vital and Health Statistics hearing

Senior Researcher David O'Brien presents at Genetic Data in Public Research Databases conference

Undergraduate Daniel Muise & Senior Research Fellow Kobbi Nissim publish "Differential Privacy in CDFs"

http://datatags.org



DataTags

Share Sensitive Data with Confidence

PRESENTATIONS   PUBLICATIONS

There are many worthy purposes for sharing data widely. In research, data sharing enables replication and validation of scientific findings and maximizes return on research investment, so it is not surprising that sponsors and publishers expect or mandate the sharing of data where possible. In organizations, data sharing leads to insights on operations and opportunities to improve goods and services. However, data containing sensitive information about individuals or personal data collected under various agreements cannot be shared openly without appropriate safeguards. An extensive body of statutes, regulations, institutional policies, consent forms, data sharing agreements, and common practices govern how sensitive data should be used and disclosed in different contexts. Researchers, institutions and companies that manage and share data must interpret how the various legal requirements and other data privacy and security standards apply to their handling of a given dataset.

DataTags helps data holders navigate these complex issues.

View Demo »

DataTags Repositories   DataTagging Tools   Legal Modeling   Privacy Preserving Tools

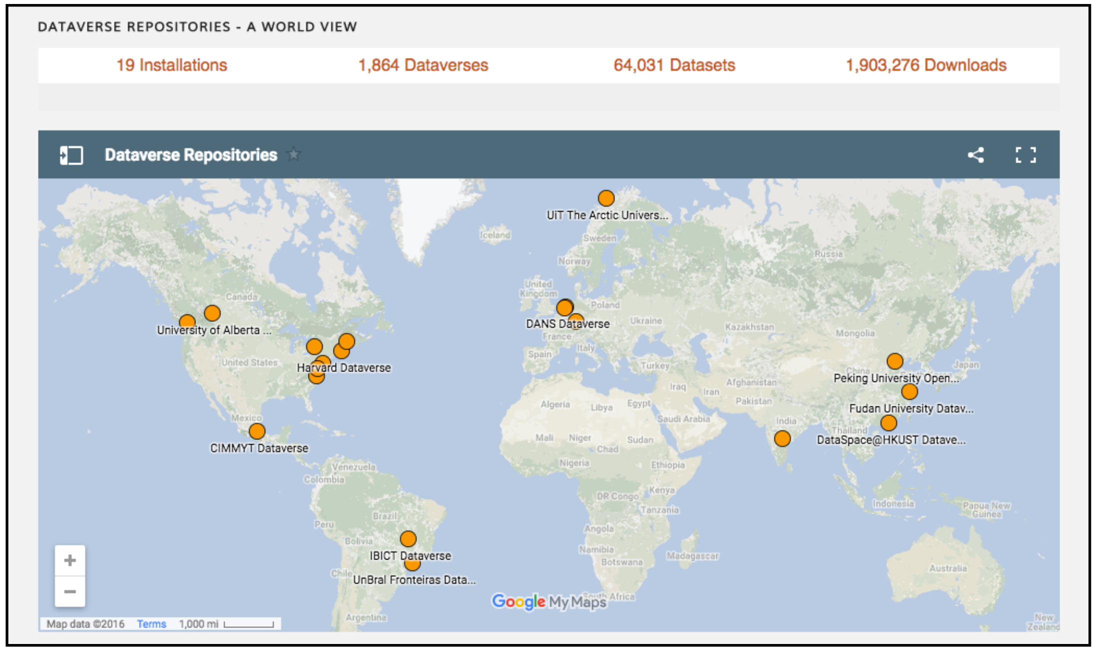
DataTags:

- is part of the Harvard University Privacy Tools Project
- now is being integrated with Dataverse

http://dataverse.org

The Dataverse Project

Open source research data repository software



DATAVERSE REPOSITORIES - A WORLD VIEW

19 Installations	1,864 Dataverses	64,031 Datasets	1,903,276 Downloads

Dataverse Repositories

UiT The Arctic Univers...
University of Alberta ...
DANS Dataverse
Harvard Dataverse
Peking University Open...
Fudan University Datav...
DataSpace@HKUST Datave...
CIMMYT Dataverse
IBICT Dataverse
UnBral Fronteiras Data...
Google My Maps
Map data ©2016 Terms 1,000 mi

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

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.

Requirements for a DataTags Repository

Requirements for a DataTags Repository

1. Supports more than one datatag

Requirements for a DataTags Repository

1. Supports more than one datatag
2. Each file in the repository must have one and only one datatag

Requirements for a DataTags Repository

1. Supports more than one datatag
2. Each file in the repository must have one and only one datatag
    - additional requirements cannot weaken the file security

Requirements for a DataTags Repository

1. Supports more than one datatag
2. Each file in the repository must have one and only one datatag
   - additional requirements cannot weaken the file security
   - and cannot required the same or more security than a more restrictive datatag

Requirements for a DataTags Repository

1. Supports more than one datatag
2. Each file in the repository must have one and only one datatag
   - additional requirements cannot weaken the file security
   - and cannot required the same or more security than a more restrictive datatag
3. A recipient of a file from the repository must

Requirements for a DataTags Repository

1. Supports more than one datatag
2. Each file in the repository must have one and only one datatag
    - additional requirements cannot weaken the file security
    - and cannot required the same or more security than a more restrictive datatag
3. A recipient of a file from the repository must
    - satisfy file's access requirements,

Requirements for a DataTags Repository

1. Supports more than one datatag
2. Each file in the repository must have one and only one datatag
   - additional requirements cannot weaken the file security
   - and cannot required the same or more security than a more restrictive datatag
3. A recipient of a file from the repository must
   - satisfy file's access requirements,
   - produce sufficient credentials as requested,

Requirements for a DataTags Repository

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

Requirements for a DataTags Repository

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

Repositories today do not have an easy, standard way to support sensitive data

The Dataverse Project Terms of Use

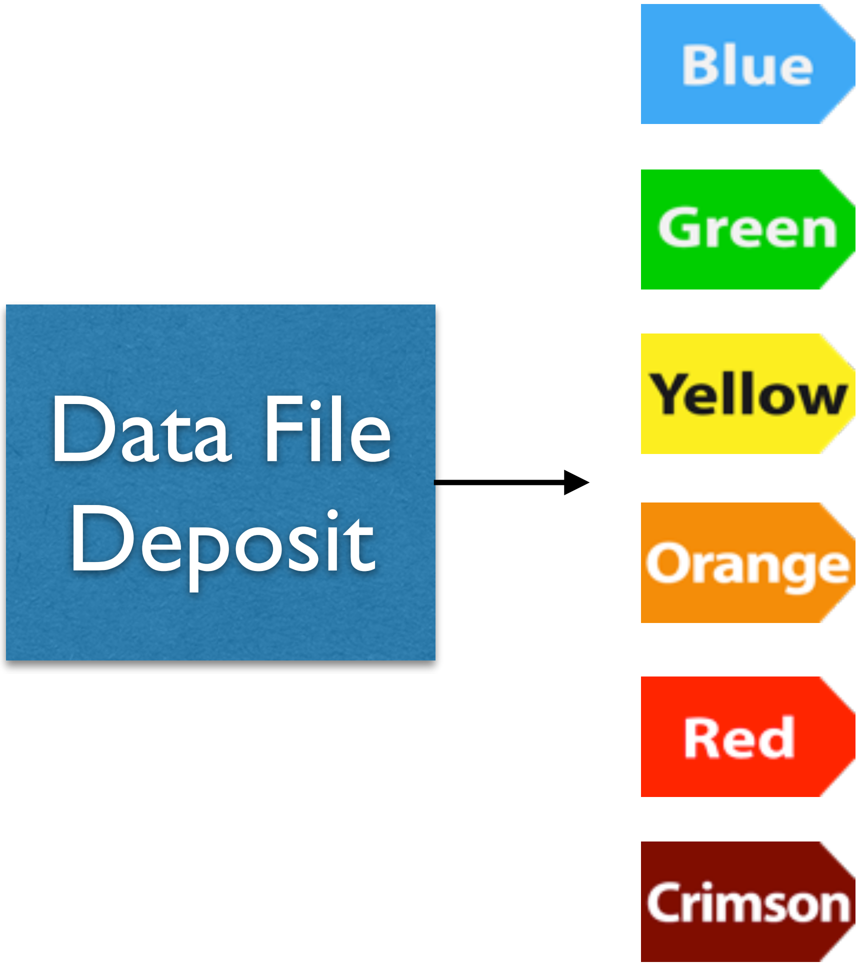
"User Uploads must be void of all identifiable information, such that re-identification of any subjects from the amalgamation of the information available from all of the materials (across datasets and dataverses) uploaded under any one author and/or user should not be possible."

Making Dataverse a DataTags Repository to share sensitive data

Making Dataverse a DataTags Repository to share sensitive data

Data File Deposit

Making Dataverse a DataTags Repository to share sensitive data



Data File Deposit

Blue

Green

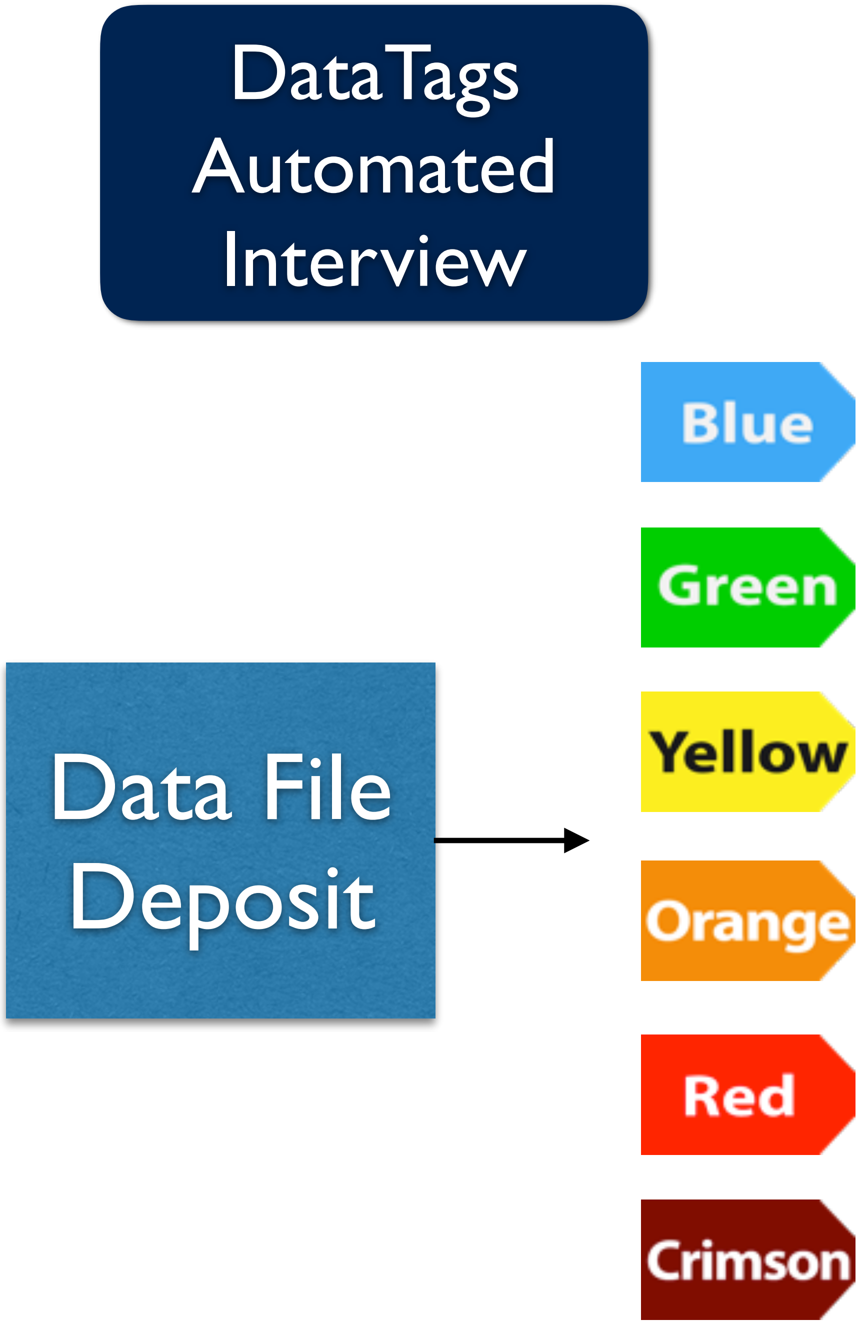
Yellow

Orange

Red

Crimson

Making Dataverse a DataTags Repository to share sensitive data



DataTags Automated Interview

Data File Deposit

Blue

Green

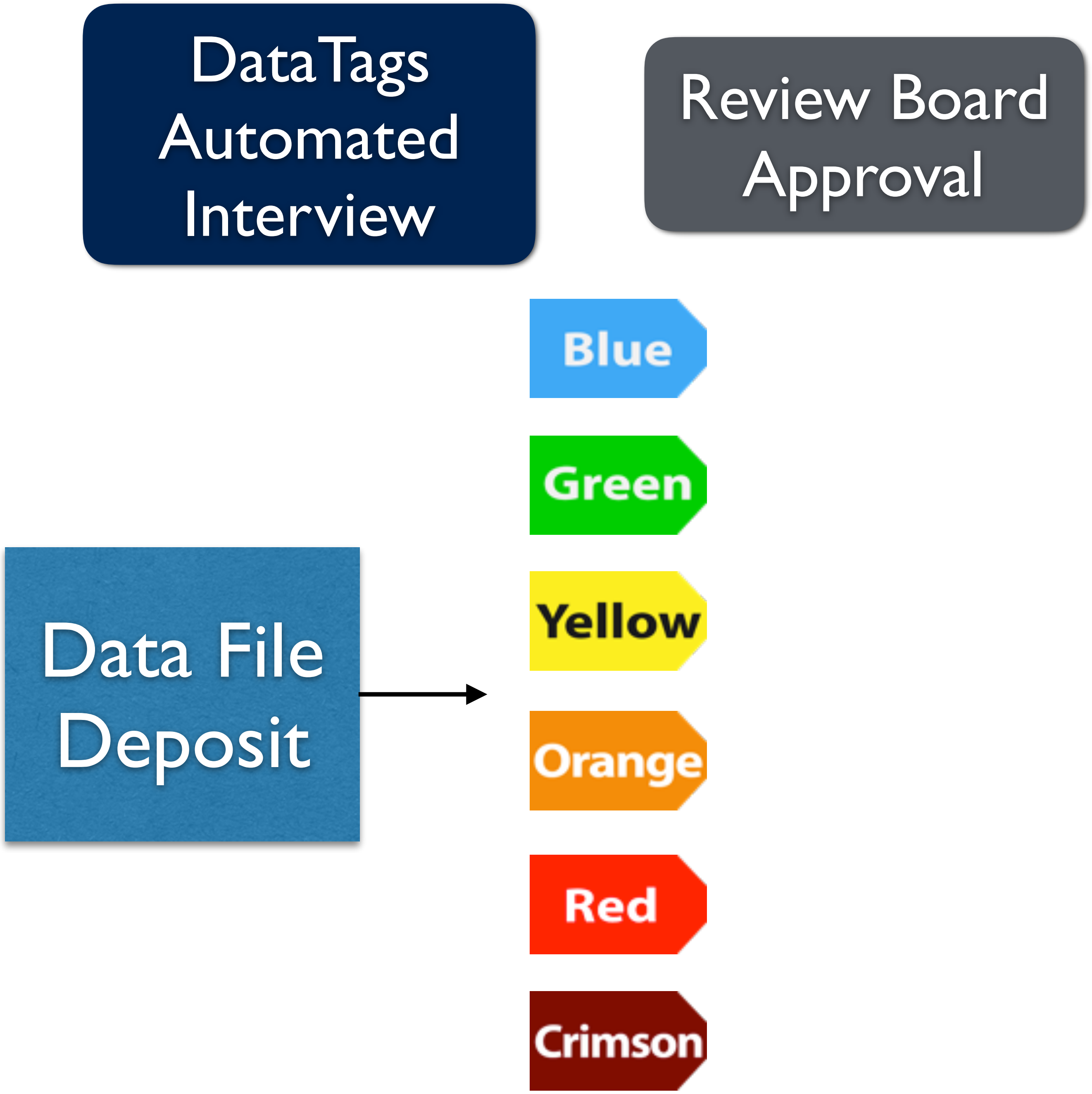
Yellow

Orange

Red

Crimson

Making Dataverse a DataTags Repository to share sensitive data



DataTags Automated Interview

Review Board Approval

Data File Deposit

Blue

Green

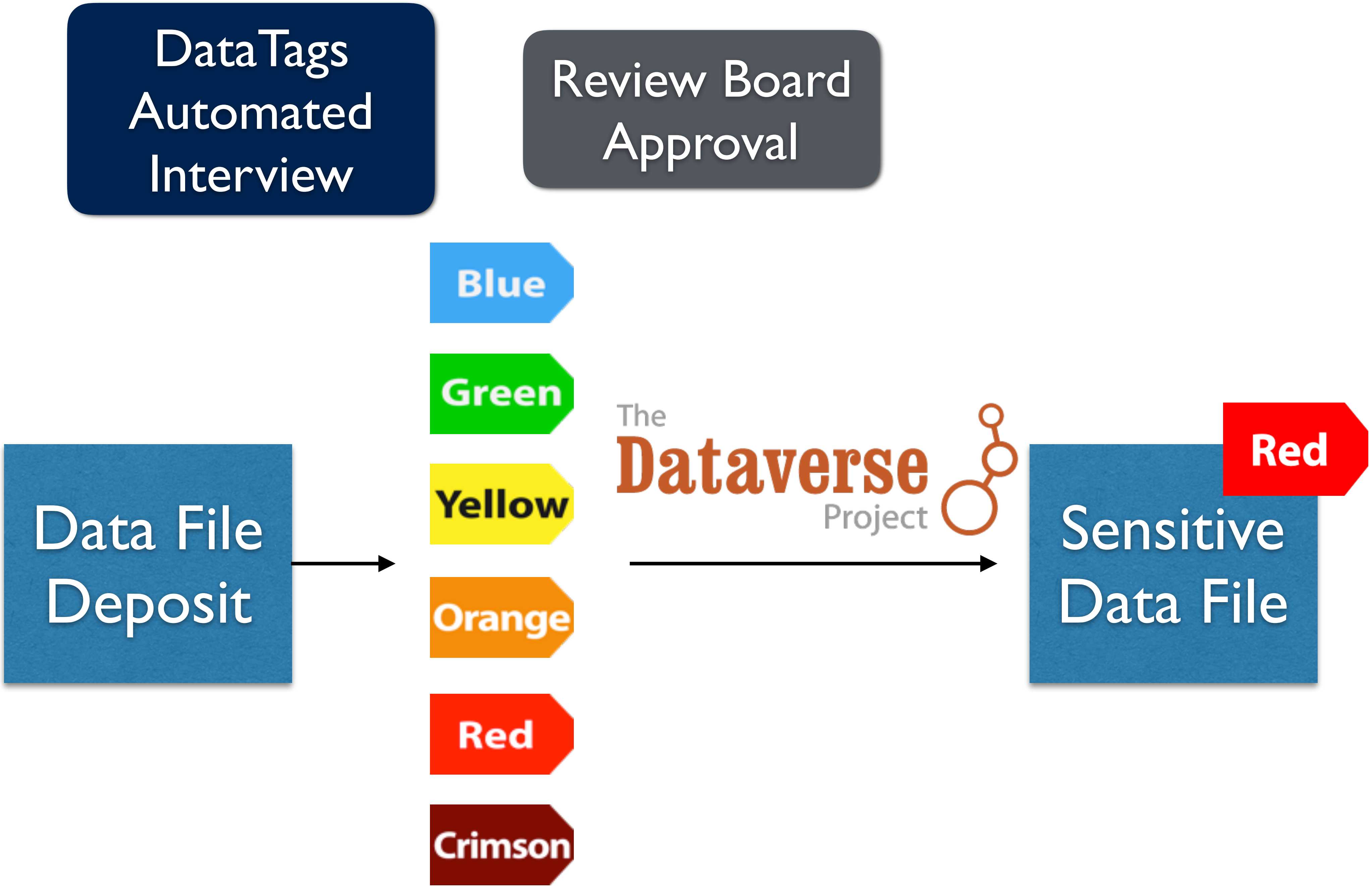
Yellow

Orange

Red

Crimson

Making Dataverse a DataTags Repository to share sensitive data



DataTags Automated Interview

Review Board Approval

Data File Deposit

Blue

Green

Yellow

Orange

Red

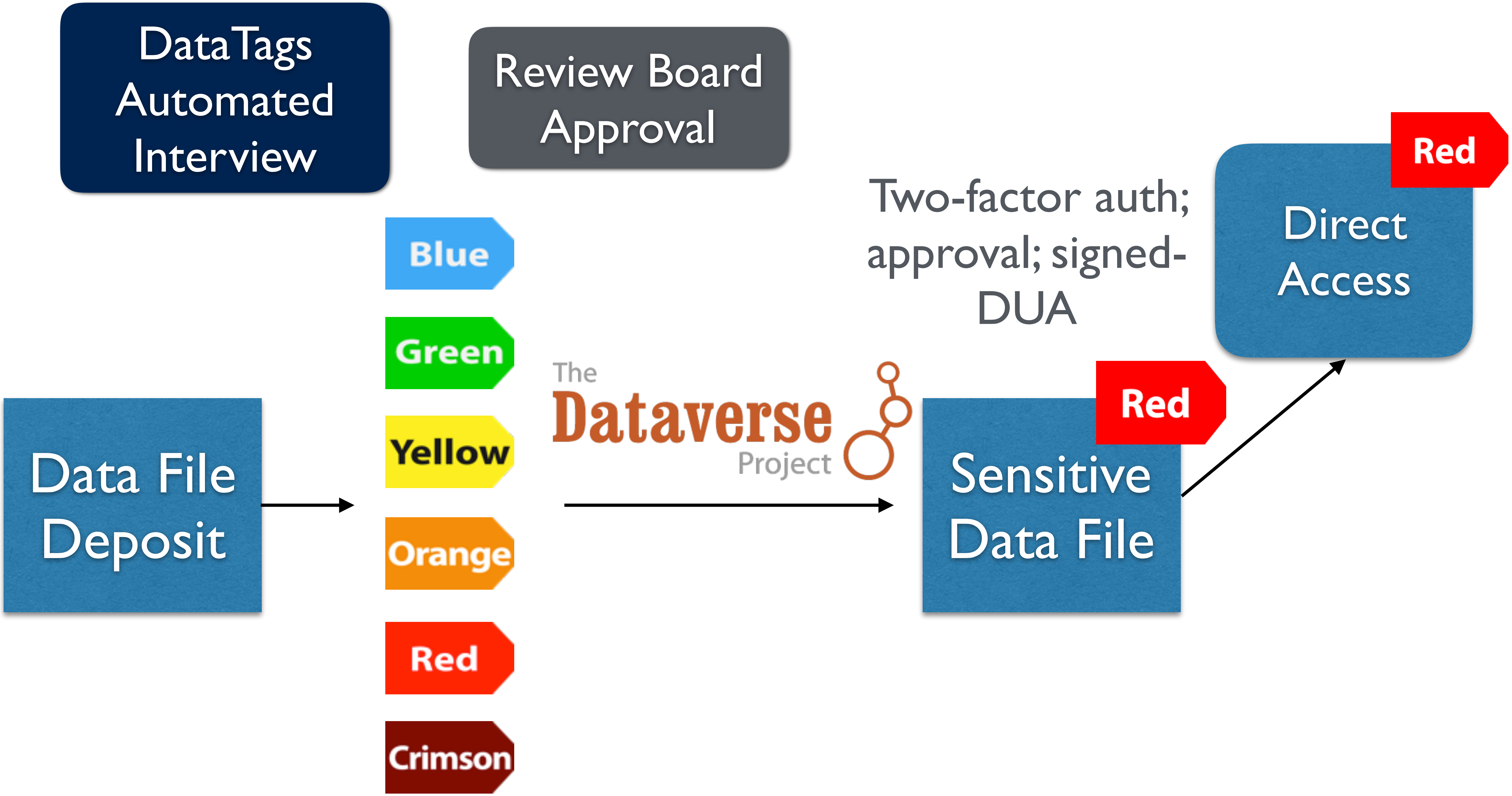
Crimson

The Dataverse Project

Red

Sensitive Data File

Making Dataverse a DataTags Repository to share sensitive data



DataTags Automated Interview

Review Board Approval

Data File Deposit

Blue

Green

Yellow

Orange

Red

Crimson

The Dataverse Project

Sensitive Data File

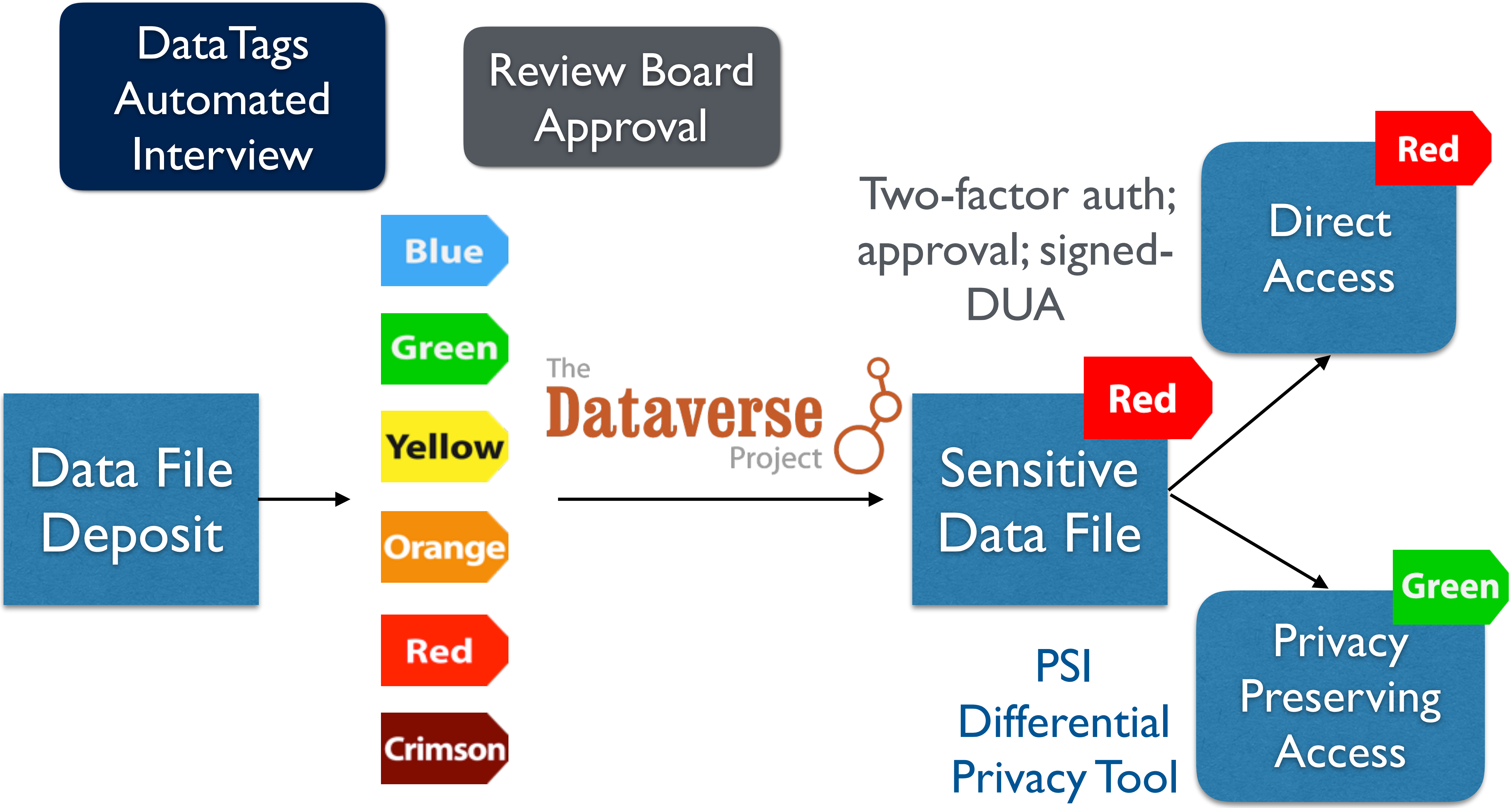
Red

Two-factor auth; approval; signed-DUA

Direct Access

Red

Making Dataverse a DataTags Repository to share sensitive data



DataTags Automated Interview

Review Board Approval

Data File Deposit

Blue

Green

Yellow

Orange

Red

Crimson

The Dataverse Project

Red

Sensitive Data File

Two-factor auth; approval; signed-DUA

Red

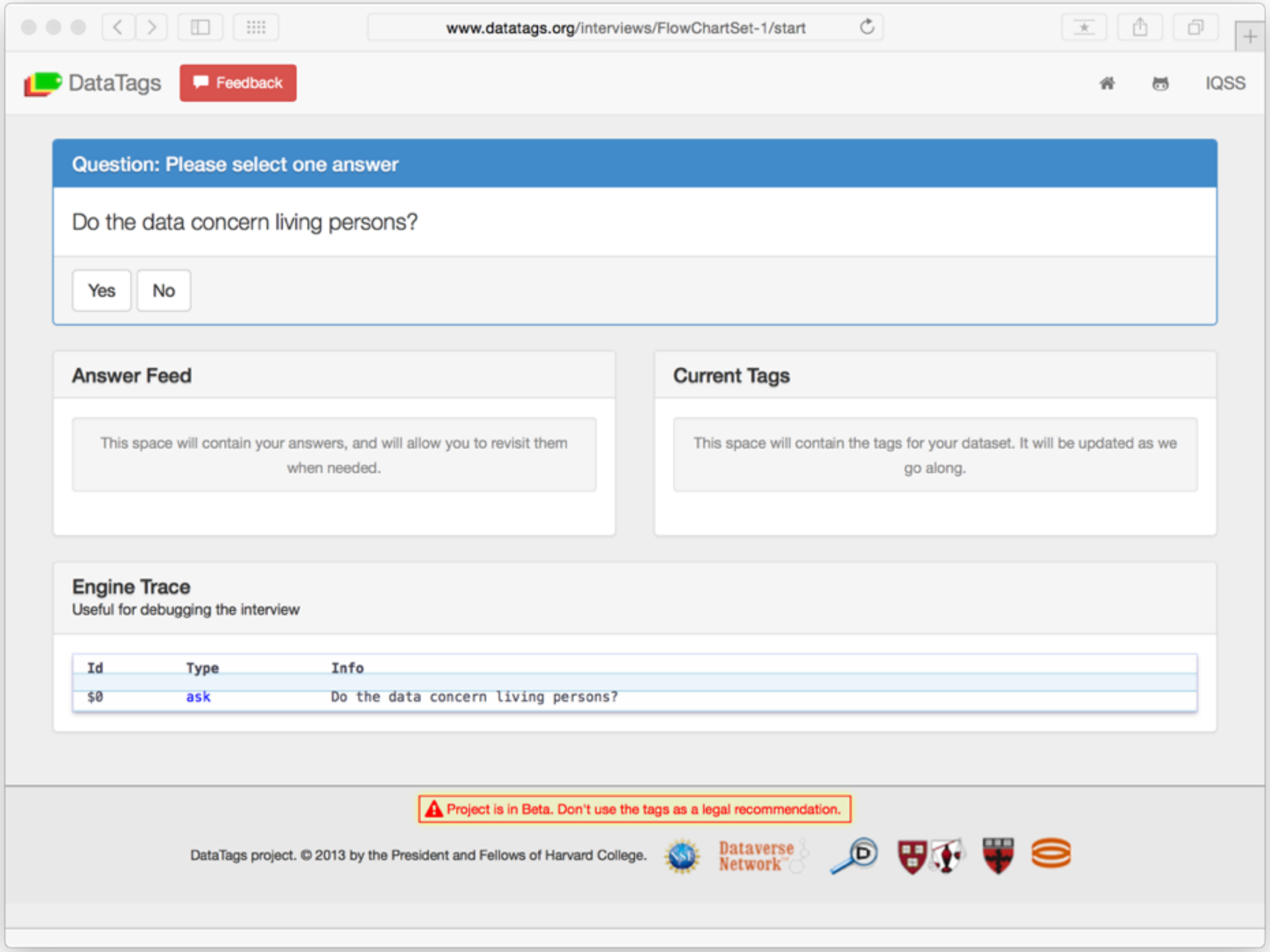
Direct Access

PSI Differential Privacy Tool

Green

Privacy Preserving Access

The DataTags automated interview …



Question: Please select one answer

Do the data concern living persons?

Yes No

Answer Feed

This space will contain your answers, and will allow you to revisit them when needed.

Current Tags

This space will contain the tags for your dataset. It will be updated as we go along.

Engine Trace

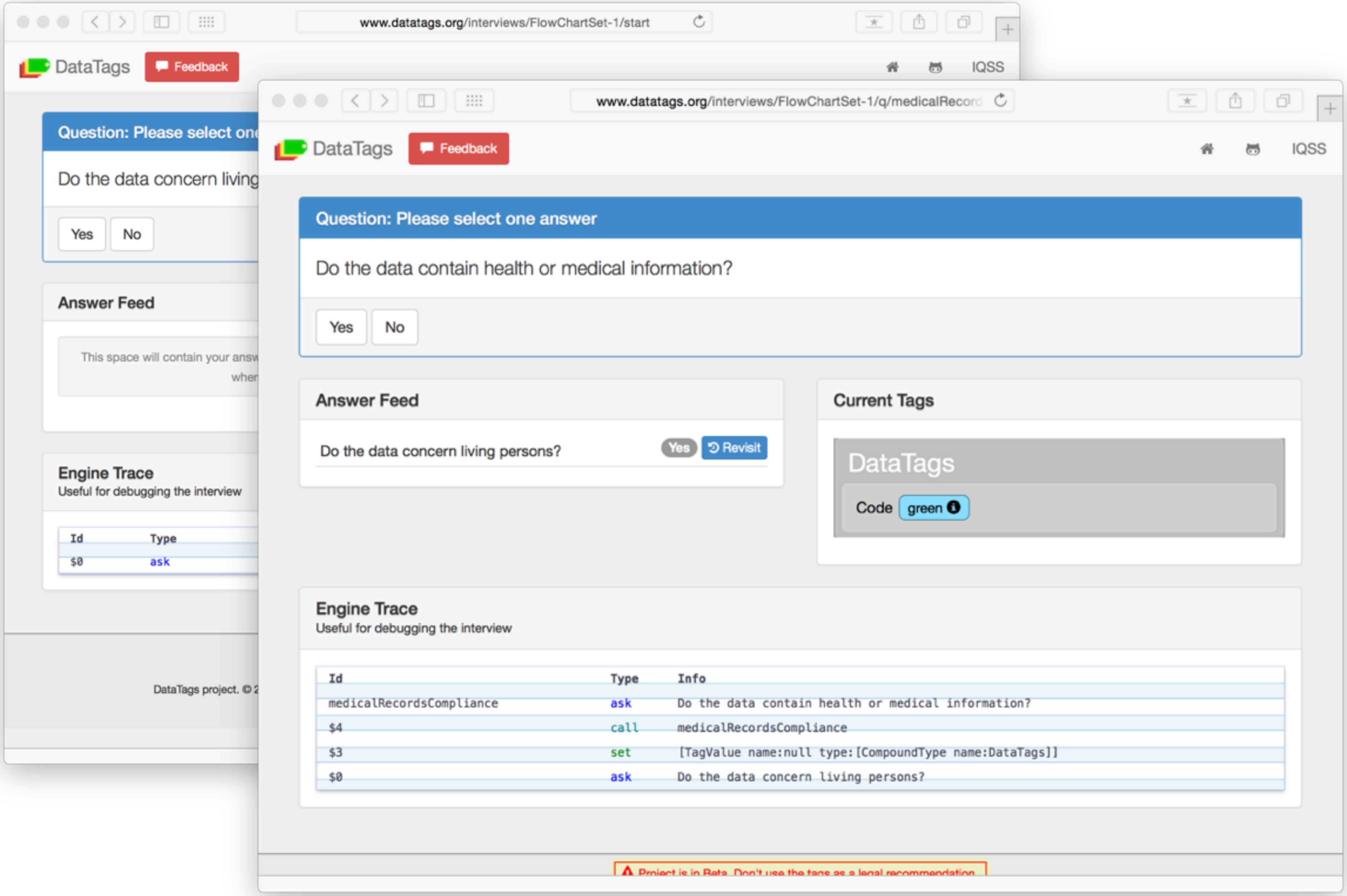
Useful for debugging the interview

```
Id        Type          Info
$0        ask           Do the data concern living persons?
```

Project is in Beta. Don't use the tags as a legal recommendation.

DataTags project. © 2013 by the President and Fellows of Harvard College.

The DataTags automated interview …



www.datatags.org/interviews/FlowChartSet-1/start

DataTags Feedback IQSS

Question: Please select one

Do the data concern living

Yes No

Answer Feed

This space will contain your answ
wher

Engine Trace
Useful for debugging the interview

Id	Type
$0	ask

DataTags project. © 2

www.datatags.org/interviews/FlowChartSet-1/q/medicalRecord

DataTags Feedback IQSS

Question: Please select one answer

Do the data contain health or medical information?

Yes No

Answer Feed

Do the data concern living persons? Yes Revisit

Current Tags

DataTags

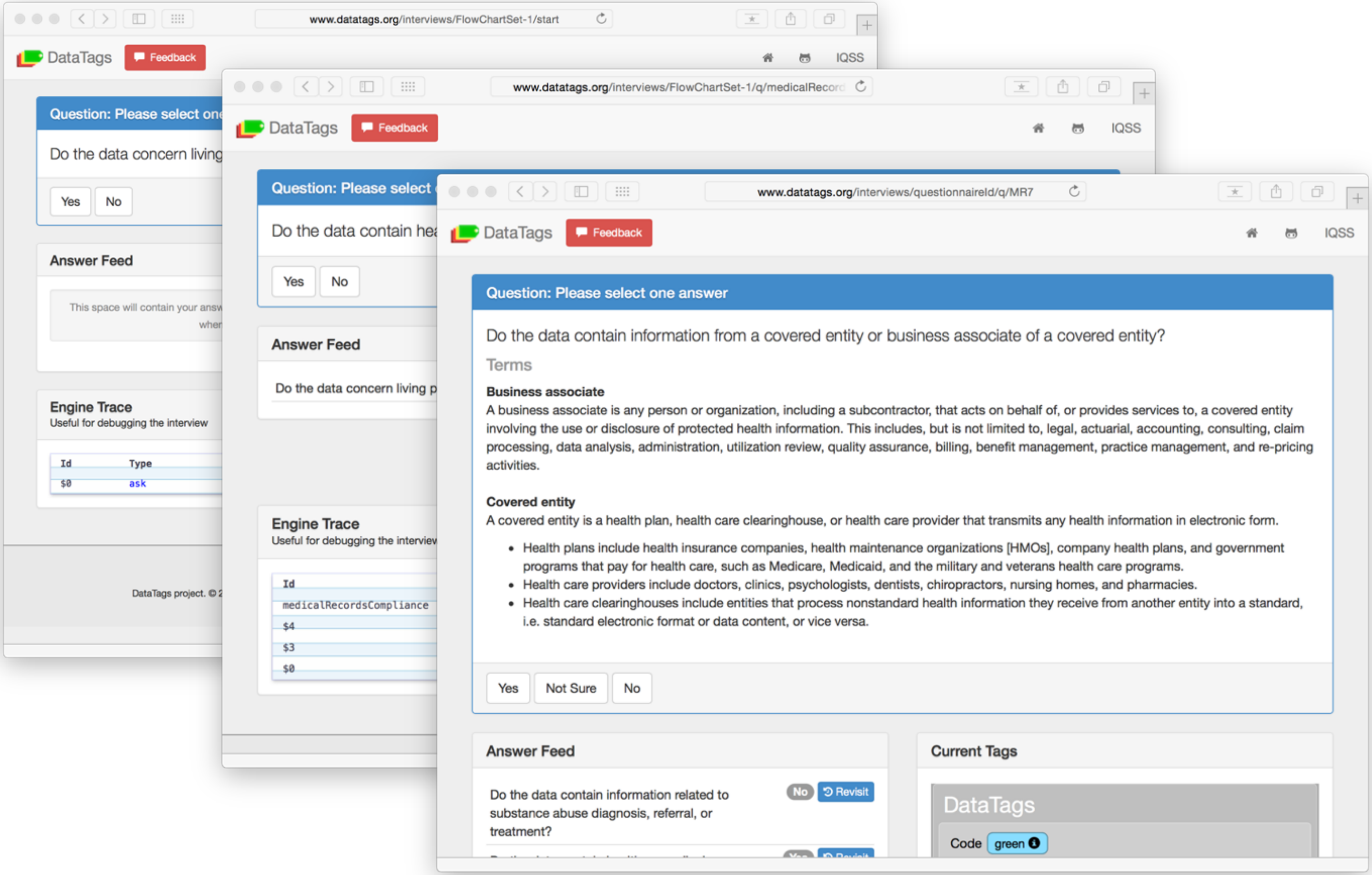
Code green

Engine Trace
Useful for debugging the interview

Id	Type	Info
medicalRecordsCompliance	ask	Do the data contain health or medical information?
$4	call	medicalRecordsCompliance
$3	set	[TagValue name:null type:[CompoundType name:DataTags]]
$0	ask	Do the data concern living persons?

⚠ Project is in Beta. Don't use the tags as a legal recommendation

The DataTags automated interview …



www.datatags.org/interviews/FlowChartSet-1/start

DataTags Feedback IQSS

Question: Please select one

Do the data concern living

Yes No

Answer Feed

This space will contain your answ
wher

Engine Trace
Useful for debugging the interview

Id	Type
$0	ask

DataTags project. © 2

www.datatags.org/interviews/FlowChartSet-1/q/medicalRecord

DataTags Feedback IQSS

Question: Please select

Do the data contain hea

Yes No

Answer Feed

Do the data concern living p

Engine Trace
Useful for debugging the interview

Id
medicalRecordsCompliance
$4
$3
$0

www.datatags.org/interviews/questionnaireId/q/MR7

DataTags Feedback IQSS

Question: Please select one answer

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.

Covered entity
A covered entity is a health plan, health care clearinghouse, or health care provider that transmits any health information in electronic form.

- 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.

Yes Not Sure No

Answer Feed

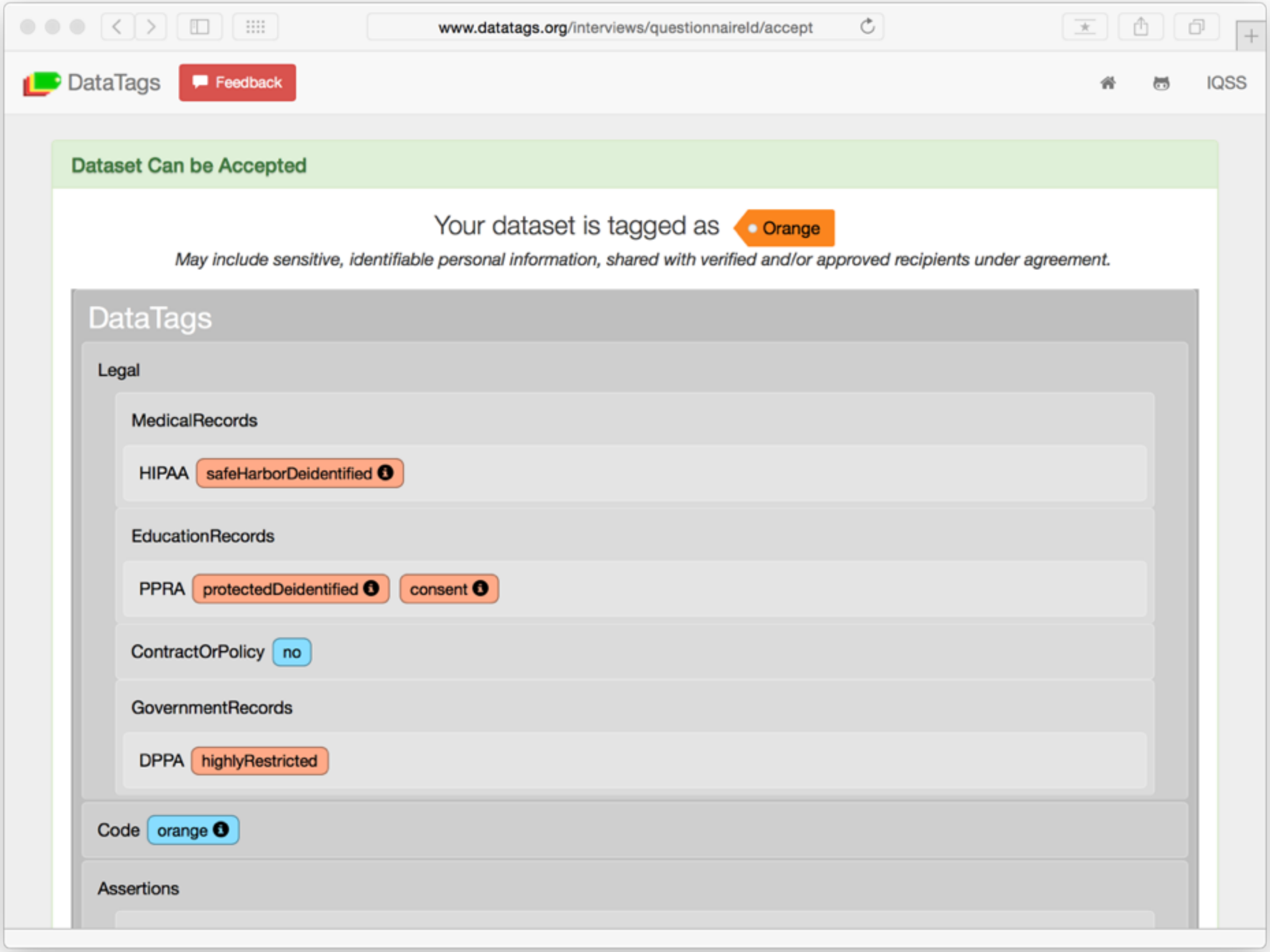
Do the data contain information related to substance abuse diagnosis, referral, or treatment? No Revisit

Current Tags

DataTags

Code green

… helps you generate a machine-readable datatag for your data



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

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 automated interview)

http://privacytools.seas.harvrd.edu

http://dataverse.org

Harvard University Privacy Tools Project funded by the National Science Foundation,
with support from the Alfred P. Sloan Foundation and Google Inc.