

Dataverse with DataTags: Sharing Data you can't share

Mercè Crosas, Ph.D. [@mercecrosas](#)

Director of Data Science

Institute for Quantitative Social Science, Harvard University

Michael Bar-Sinai [@michbarsinai](#)

Architect, Senior Software Engineer,

Institute for Quantitative Social Science, Harvard University

<http://datascience.iq.harvard.edu>

Introduction to Dataverse

Dataverse Software

- A framework for publishing, citing and preserving research data:
<http://thedata.org>
- Open-source, available at GitHub
- Started in 2006 at IQSS
- Can support all data types across multiple disciplines
- APIs to integrate with journal systems and other repositories

Dataverse Repository

- **Harvard** hosts a Dataverse instance **free and open** to all research data:
<http://thedata.harvard.edu>
- More than 53,000 datasets, with 735,000 files
- Dataverses can be created for researchers, journals, organizations, educators, ...
- It federates with > 10 Dataverse installations around the world .

Find and publish data at: <http://thedata.harvard.edu>



Share, Cite, Reuse, Archive Research Data
Scientific data for reproducible research

Get credit for and keep control of your data, while preservation is guaranteed

Harvard Dataverse

POWERED BY THE **Dataverse Network** PROJECT v. 3.6.2

Search [Create Account](#) [Log In](#)

[Advanced Search](#) [Tips](#)

We're redesigning Dataverse and want your feedback! Please check out our [Beta Site](#)

The Harvard Dataverse Network is open to all scientific data from all disciplines worldwide. It includes the world's largest collection of social science research data. [Learn more about the Dataverse Network.](#)

Dataverses

[Create Dataverse](#)

706 Dataverses

A **Dataverse** is a container for research data studies, customized and managed by its owner.

RECENTLY RELEASED DATAVERSES

- | | |
|---|--------------|
| Eben N. Broadbent | Jun 2, 2014 |
| USoc: Quantitative Methods over the Undergraduate Life Course | May 30, 2014 |

Studies

53,896 Studies, **739,606** Files, **1,015,093** Downloads

A **study** is a container for a research data set. It includes cataloging information, data files and complementary files.

RECENTLY RELEASED STUDIES

- | | |
|--|-------------|
| Replication data for: Neoliberal Reform and Protest in Latin American Democracies: A Replication and Correction by Solt, Frederick; Kim, Dongkyu; Lee, Kyu Young; Willardson, Spencer; Kim, Seokdong | Jun 3, 2014 |
|--|-------------|

Dataverse Features

June 2014

Dataverse allows you to:

- Get a formal citation for your data
- Link your data set to the original publication(s)
- Publish multiple versions of your datasets
- Set terms of use for your data
- Restrict data files, while metadata and documentation can be kept public (but we encourage **open data**, when possible)
- Brand your dataverse banner with your logo, image or colors
- Track downloads for your data, and enable a guestbook
- List data sets from other dataverses in your dataverse

Dataverse 4.0 (Fall 2014)

Harvard Dataverse
The Harvard Dataverse is open to all scientific data from all disciplines worldwide. It includes the world's largest collection of social science research data.

Search this Dataverse... Find Advanced Search Add Data

1 to 10 of 12 results

Publication Status
 Unpublished (8)
 Published (3)
 Draft (1)

Affiliation
 Harvard University (6)
 IQSS (2)
 European Union (1)
 McGill University (1)
 NASA (1)

Publication Date
 2014 (4)

Author Name
 Smith, John (2)

Author Affiliation
 IQSS (2)

Keyword
 election (2)

Subject
 Law (2)

Contributor Type
 Data Collector (2)

Production Date
 4 (2)

Deposit Date
 2014 (2)

Draft
 Results from the 2004 Election in Mississippi
 Smith, John, 2014, "Results from the 2004 Election in Mississippi", http://dx.doi.org/10.5072/FK2/12, Harvard Dataverse
 Data for the results of the elections in 2004 that took place in the state of Mississippi. This includes all federal, state, and local elections.
 Host Dataverse: Department of Government Dataverse

Harvard Business Dept Dataverse
 Harvard University
 The Harvard University Business Department.

Department of Government Dataverse
 Harvard University
 Datasets from Harvard University's Department of Government.
 Preview Recently Released Datasets [+]

Unpublished
 International Cosmos Journal Dataverse
 NASA
 Datasets from articles published in the International Cosmos Journal

Unpublished
 Climate Change in Massachusetts Dataverse

- New UI
- New rich, faceted search
- Reformatting and metadata extraction for more data types (excel, CSV, RData, Stata, SPSS, FITS)
- Metadata standards for social sciences, astronomy, biomedical sciences.
- Integration with a new data exploration and analysis tool for tabular data: **TwoRavens**

Try Dataverse 4.0 Beta:
<http://dataverse-demo.iq.harvard.edu>



Dataverse 4.0 will include a new interactive data exploration and analysis tool, **TwoRavens**, which integrates with **Zelig** statistical framework

TwoRavens

Estimate

Force

Reset

turnout

Variables

Subset

Summary

educate

Education

Mean: 12.2

Median: 12

Mode: NaN

Stand.Dev: 3.39

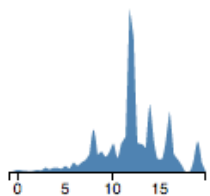
Minimum: 0

Maximum: 19

Valid: 15837

Invalid: 0

educate



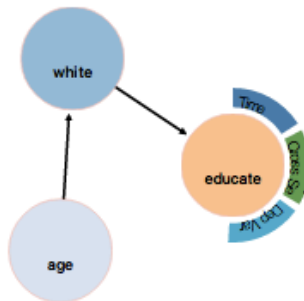
educate

log(d)

exp(d)

d^2

sqrt(d)



Results Table

Models

Set Covar.

Results

Title *

Replication Data for: Building a Bridge Betwe

Add 'Replication Data for' to Title

Author**Name ***

Castro, Eleni

Affiliation

IQSS

Contact E-mail *

ecastro@fas.harvard.edu

**Description ***

Research dataset for my publication on connecting journal articles and their underlying research data. Includes analysis of current data publication practices.

Citation Metadata:
Compliant with DataCite, Dublin Core, DDI study description.
Applies to all datasets.

Keyword

data publication

**Subject ***

- Mathematical Sciences
- Physics
- Social Sciences
- Other

Topic Classification

Term

Vocabulary



URL

Software

Name

Version



Series

Name

Information

Time Period Covered

Start

End



Date of Collection

Start

End



Country/Nation

Geographic Coverage

Geographic Unit

Geographic Bounding Box

West Longitude

East Longitude

North Latitude

South Latitude

Social Sciences and Humanities Metadata: Compliant with DDI

Type

- Image
- Mosaic
- EventList
- Spectrum
- Cube

Facility

Instrument

Spatial Resolution

Spectral Resolution

Time Resolution

Bandpass

Central Wavelength (m)

Wavelength Range

Minimum (m)

Maximum (m)

Dataset Date Range

Start

End

**Astronomy Metadata:
Compliant Virtual Observatory
(VO) schema; extract metadata
from FITS files**

Design Type

- Case Control
- Cross Sectional
- Not Specified
- Parallel Group Design
- Perturbation Design

Factor Type

- Age
- Biomarkers
- Developmental Stage
- Cell Surface Markers
- Cell Type/Cell Line

Measurement Type

- DNA Methylation Profiling (Bisulfite-Seq)
- DNA Methylation Profiling (MeDIP-Seq)
- Histone Modification (ChIP-Seq)
- Protein-RNA Binding (RIP-Seq)
- Transcription Factor Binding (ChIP-Seq)

Bio Metadata:
Compliant with ISA-Tab schema,
plus biomedical ontologies

Organism

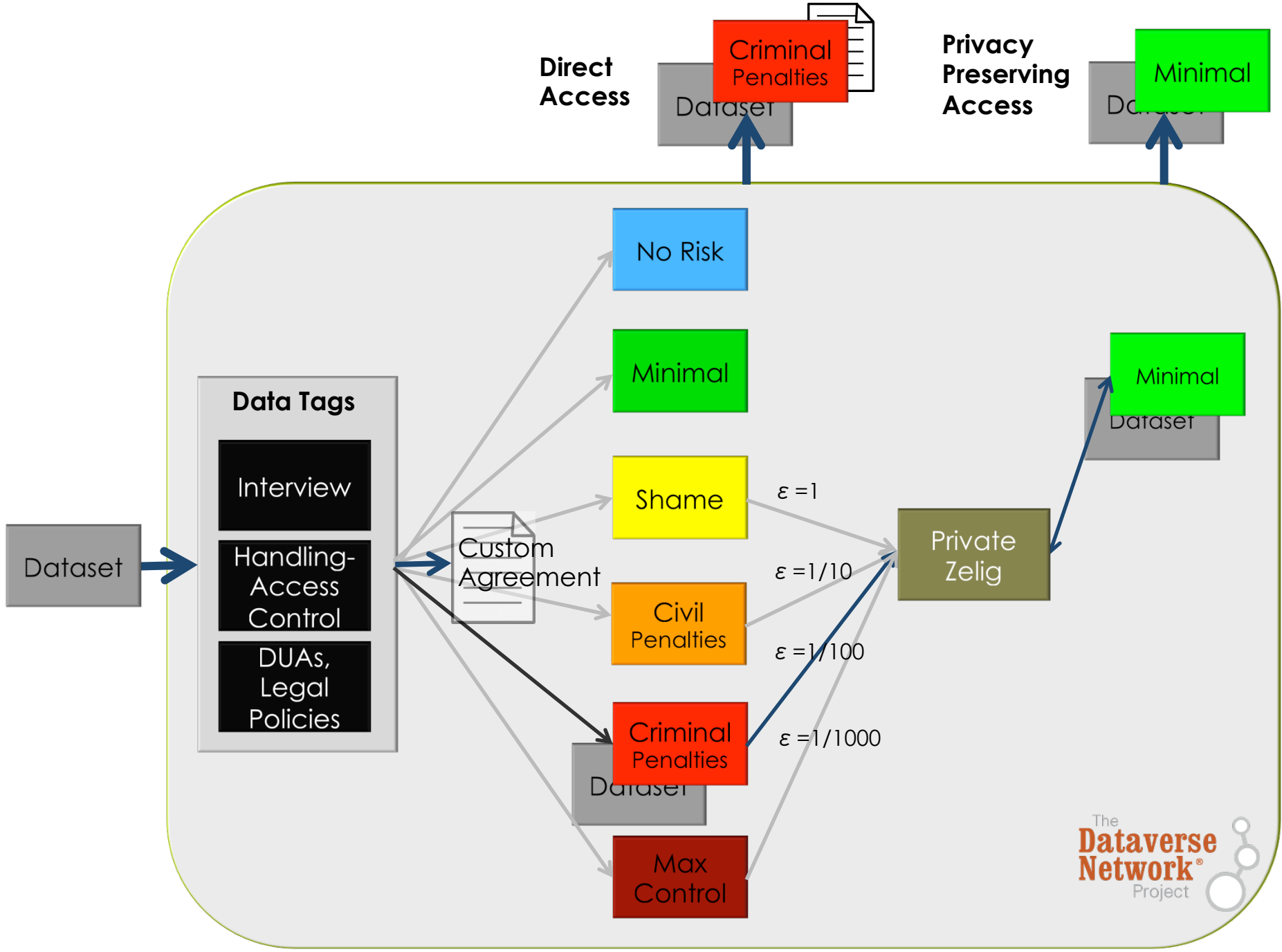
- Danio rerio
- Homo sapiens
- Mus musculus
- Rattus norvegicus

Cell Type



Sharing Data You Can't Share

- ▣ Dataverse is part of a 4 years NSF funded project on **Privacy Tools for Sharing Sensitive Data**
<http://privacytools.seas.harvard.edu/> (*with Harvard SEAS, Berkman Center, Data Privacy Lab, and IQSS*).
- ▣ This project includes:
 - ▣ **DataTags:** A framework that provides data handling prescriptions to comply with numerous privacy regulations and data user agreements
 - ▣ **Private Zelig:** A differential privacy version of the Zelig statistical framework



Try our new Beta version: <http://datatags.org>

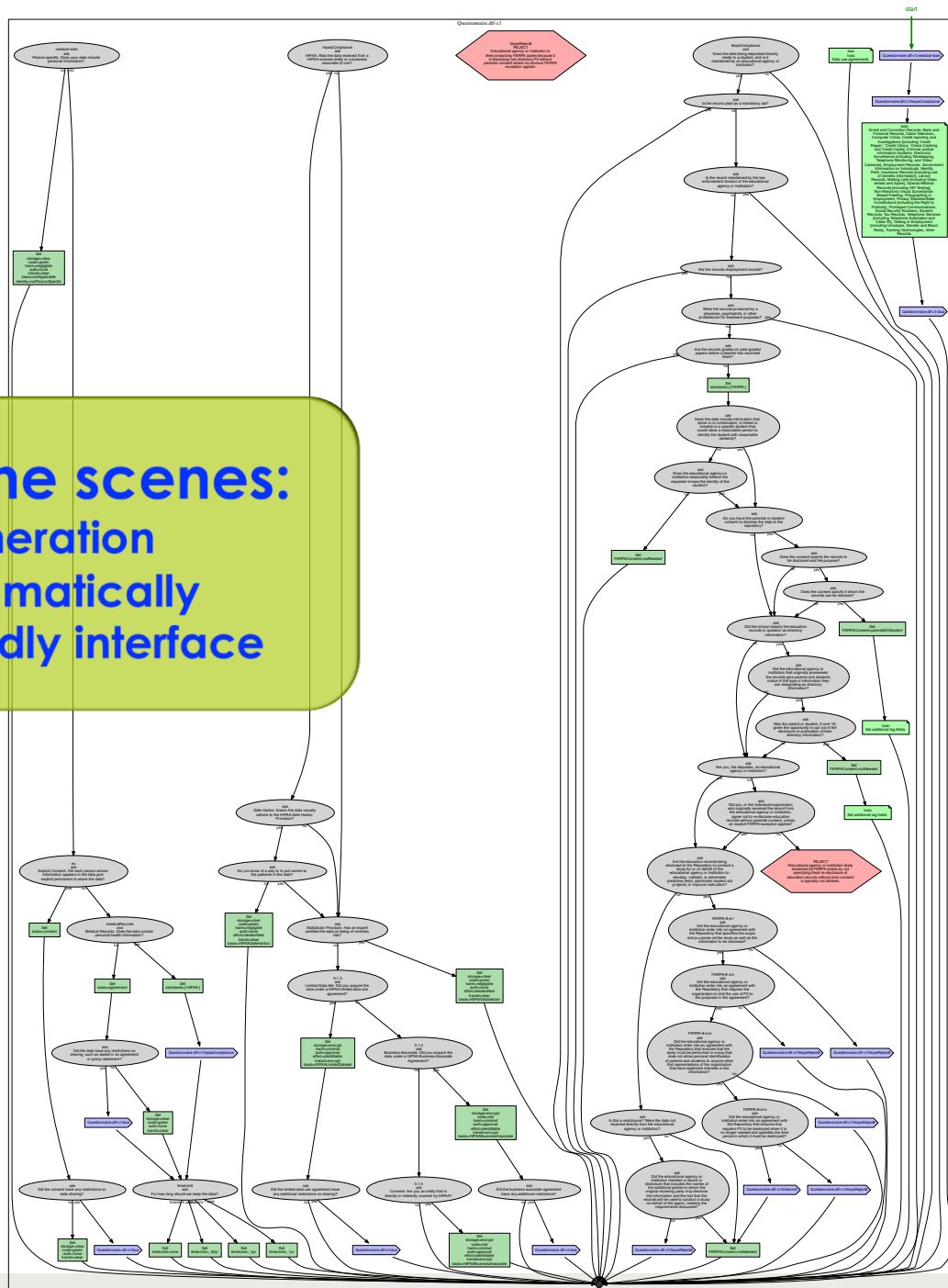
Harm Levels and Their Appropriate Tags

The tags below denote are the minimal handling requirements, based on the harm level inherent to the data. The tags resulting from the tagging interview may be more restrictive, due to data use agreements, contracts etc. Hover/touch tags for explanation

Level	DUA Agreement Method	Authentication	Transit	Storage
No Risk	None	None	Clear	Clear
Minimal	None	Email or OAuth	Clear	Clear
Shame	Click Through	Password	Encrypted	Clear
Civil Penalties	Sign	Password	Encrypted	Encrypted
Criminal Penalties	Sign	Two Factor	Encrypted	Encrypted
Max Control	Sign	Two Factor	Double Encryption	Double Encryption

Currently supporting HIPAA and FERPA (and DUAs)

DataTags behind the scenes: A complex interview generation framework, which is automatically converted to a user-friendly interface



Interview Example: First question ...

Question: Please select one answer

Person-specific. Does your data include personal information?

Terms

personal information
as defined in HIPAA

data
0s and 1s in some structured way

Interview Example: After several questions ...

Question: Please select one answer

Were the data collected by a federal agency?

Answer Feed

Does the data being deposited directly relate to a student,	<input checked="" type="radio"/> no	<input type="button" value="Revisit"/>
For how long should we keep the data?	<input checked="" type="radio"/> 5 years	<input type="button" value="Revisit"/>
Covered. Are you an entity that is directly or indirectly	<input checked="" type="radio"/> yes	<input type="button" value="Revisit"/>
Business Associate. Did you acquire the data under a	<input checked="" type="radio"/> no	<input type="button" value="Revisit"/>
Limited Data Set. Did you acquire the data under a HIPAA	<input checked="" type="radio"/> no	<input type="button" value="Revisit"/>
Statistician Provision. Has an expert certified the data as	<input checked="" type="radio"/> no	<input type="button" value="Revisit"/>
HIPAA. Was the data received from a HIPAA covered	<input checked="" type="radio"/> no	<input type="button" value="Revisit"/>
Medical Records. Does the data contain personal health	<input checked="" type="radio"/> yes	<input type="button" value="Revisit"/>
Explicit Consent. Did each person whose information	<input checked="" type="radio"/> no	<input type="button" value="Revisit"/>
Person-specific. Does your data include personal	<input checked="" type="radio"/> yes	<input type="button" value="Revisit"/>

Interview Example: ... and a Final Tag

Your dataset is tagged as



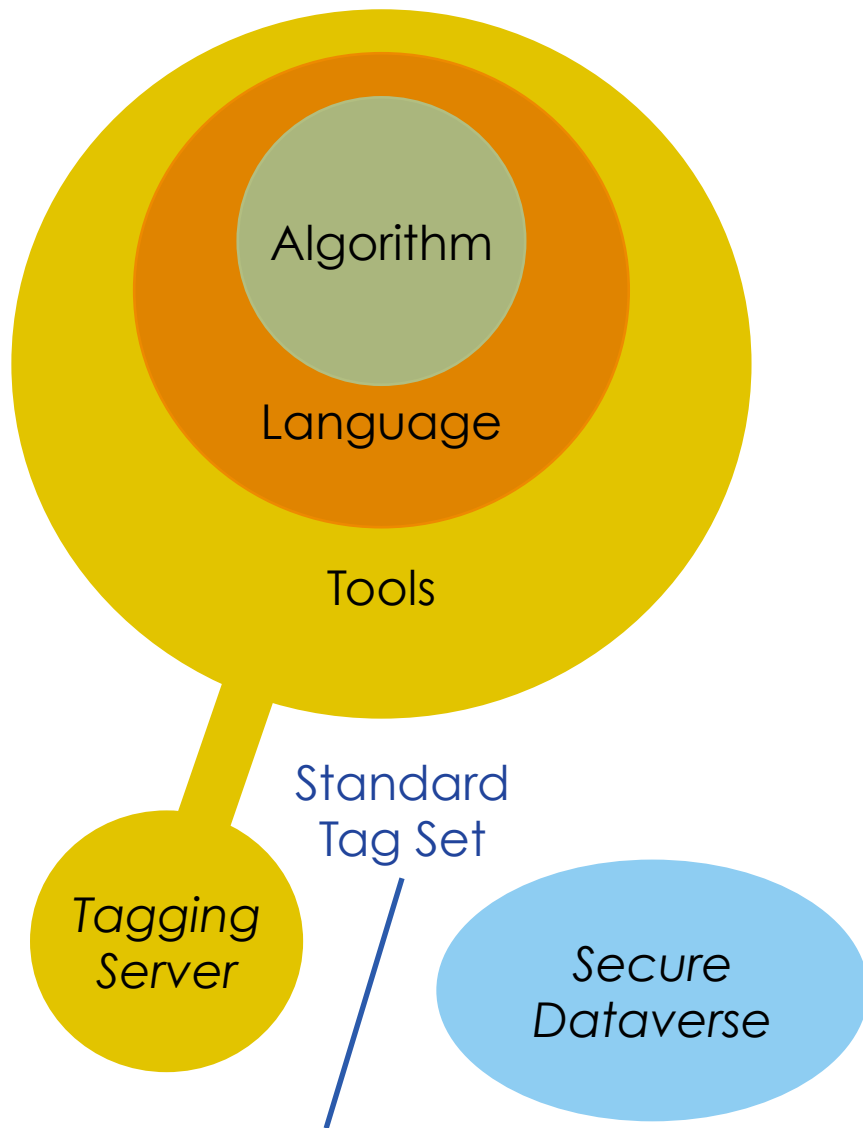
red

Very sensitive identifiable personal information, shared with strong verification of approved recipients under signed agreement.

Full Tags

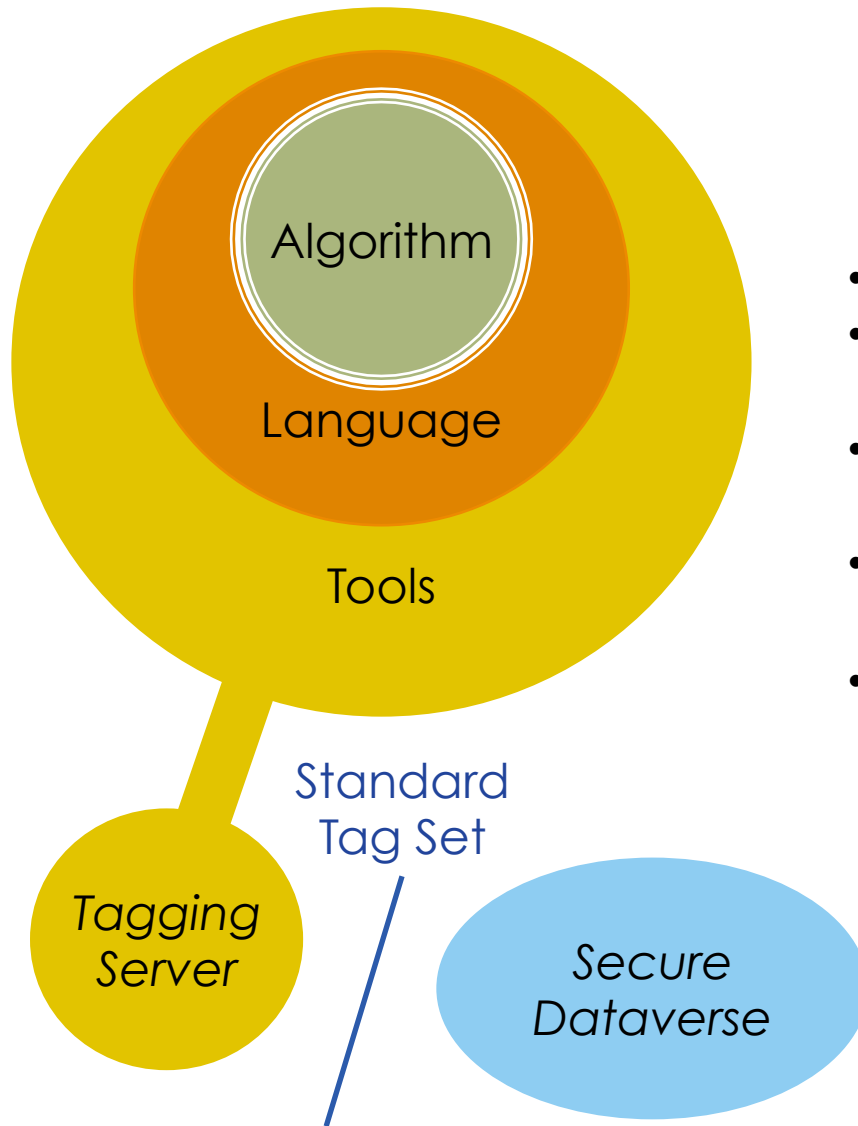
DataTags	
code	red
DataType	
harm	criminal
effort	identifiable
standards	HIPAA
Handling	
storage	encrypt
auth	approval
transit	encrypt
basis	HIPAABusinessAssociate
DUA	
timeLimit	_5yr

Project Structure



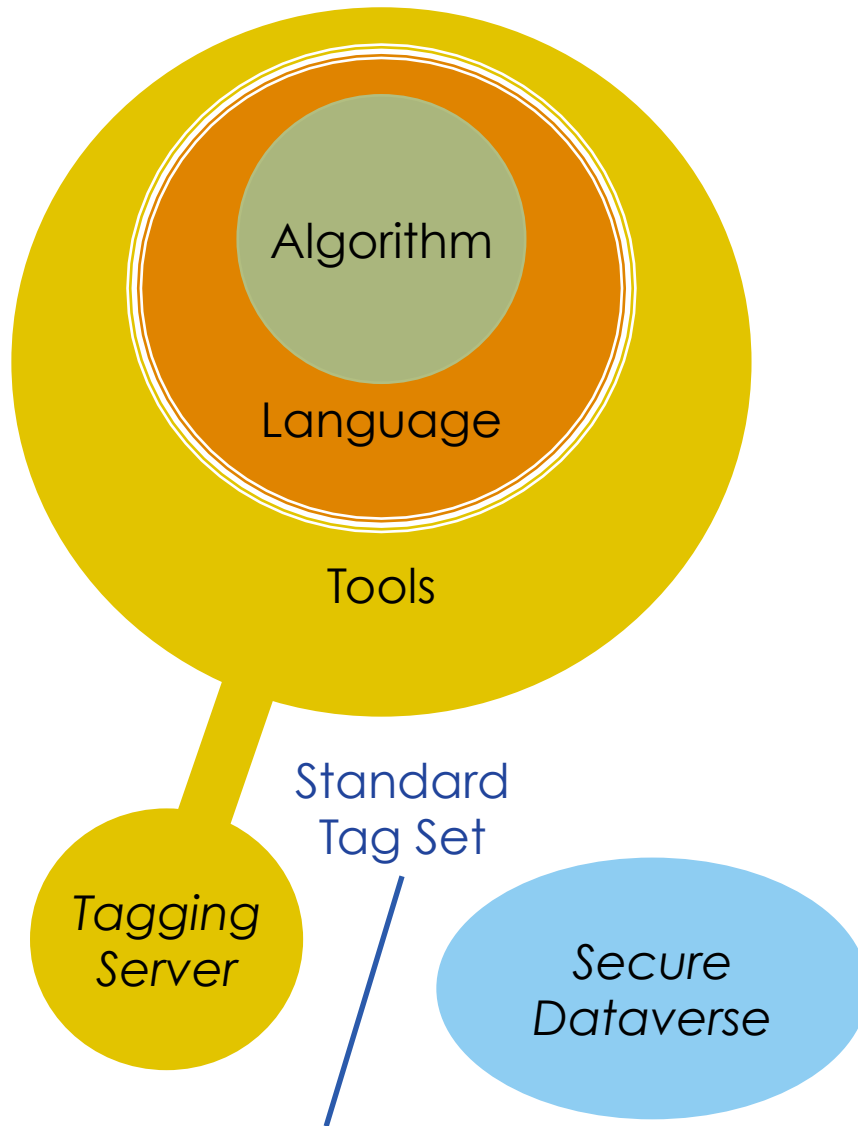
The DataTags project consists of several distinct components.

Algorithm



- “Harmonizes law and technology”
- Consists of a tag ontology and an interview process
- Created by legal and technological experts
- Currently Supports HIPAA, FERPA, CIPSEA and Privacy Act
- Developed by Berkman, DPL and IQSS

Language



Ontology definition language

- Define an interview and coding process: ask Questions, Set values to the tags
- Allows localization and extension
- Supports any closed-ended questionnaire. DataTags is a private case of this.

Interview and coding language

- Defines tagging ontologies
- Allows atomic (simple), aggregate and compound values

Tag Definition

DataTags: code, basis, Handling, DataType, DUA, IP, identity, FERPA, CIPSEA.

TODO: IP.

code: one of

blue (Non-confidential information),
green (Potentially identifiable but not...),
yellow (Potentially harmful personal information...),
orange (May include sensitive, identifiable information...),
red (Very sensitive identifiable personal information...),
crimson (Requires explicit permission for each transaction...)

▪

Handling: storage, transit, authentication, auth.

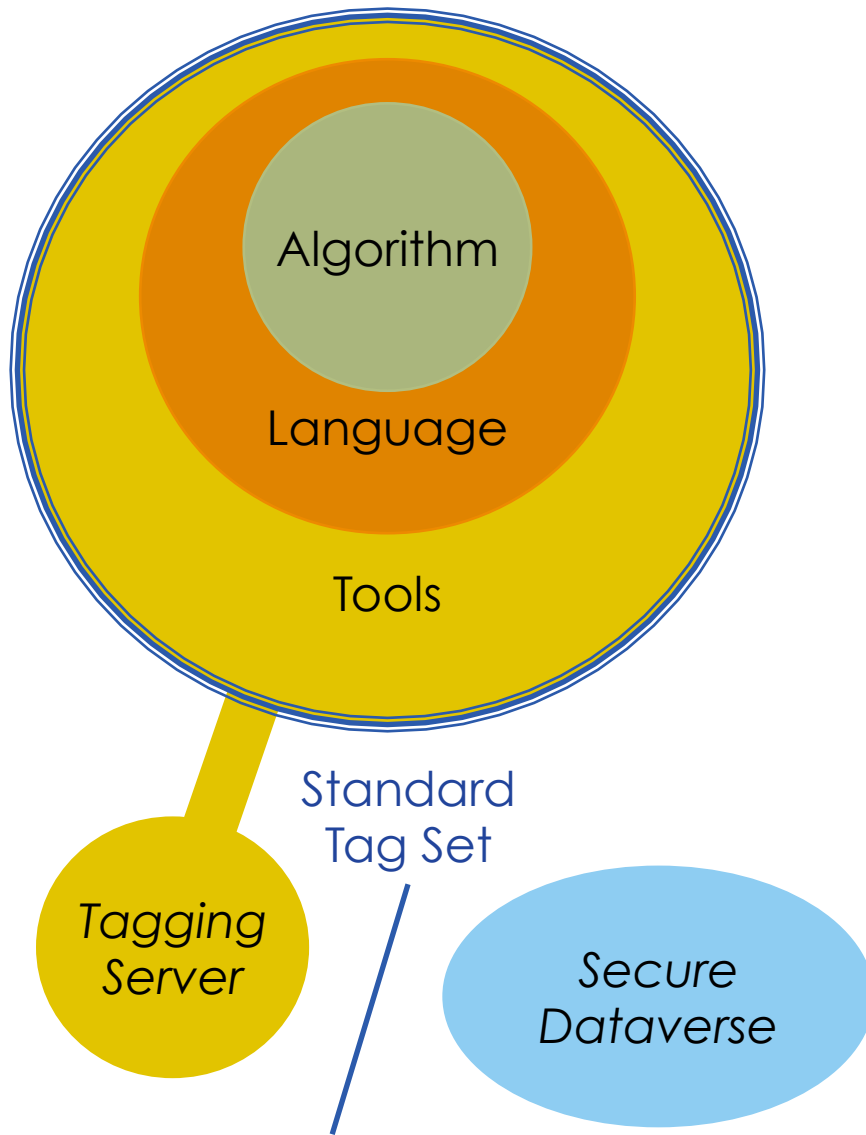
storage: one of clear, encrypt, doubleEncrypt.

standards: some of HIPAA, FERPA, ElectronicWiretapping, CommonRule, CIPSEA.

Questionnaire Definition

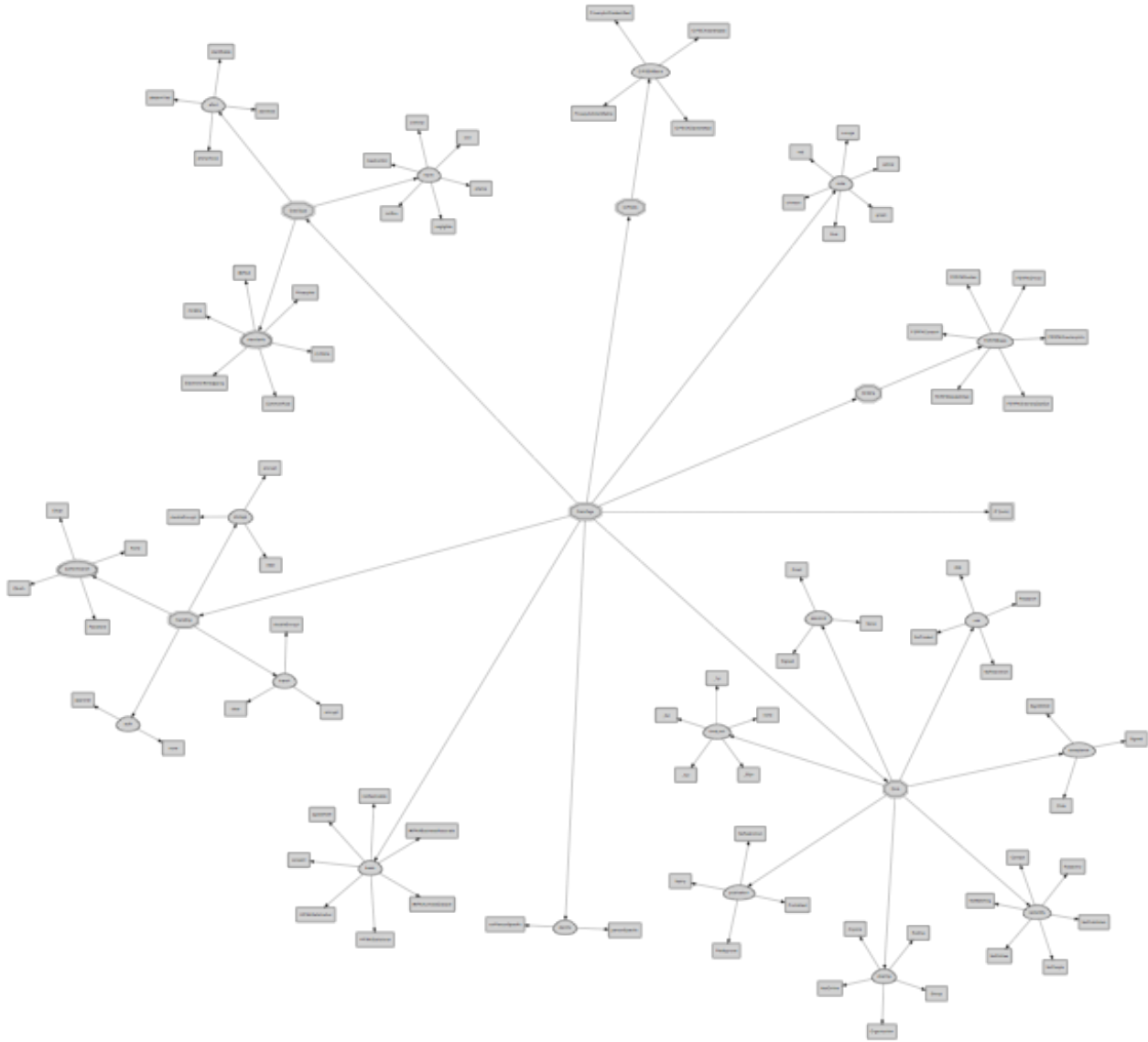
```
(>medical-start< ask:
  (text: Person-specific. Does your data include personal information?)
  (terms:
    (data: 0s and 1s in some structured way)
    (personal information: as defined in HIPAA))
  (no:
    (set: code=green, storage=clear, transit=clear, auth=none,
      basis=notApplicable, identity=notPersonSpecific,
      harm=negligible)
    (end)
  ))
(>ec< ask:
  (text: Explicit Consent. Did each person whose information appears in the
    data give explicit permission to share the data?)
  (yes:
    (set: basis=consent)
    (ask:
      (text: Did the consent have any restrictions on data sharing?)
      (no: (set: code=green, storage=clear, transit=clear, auth=none))
      (yes: (call: dua)))
    (end)
  ))
))
```

Tools

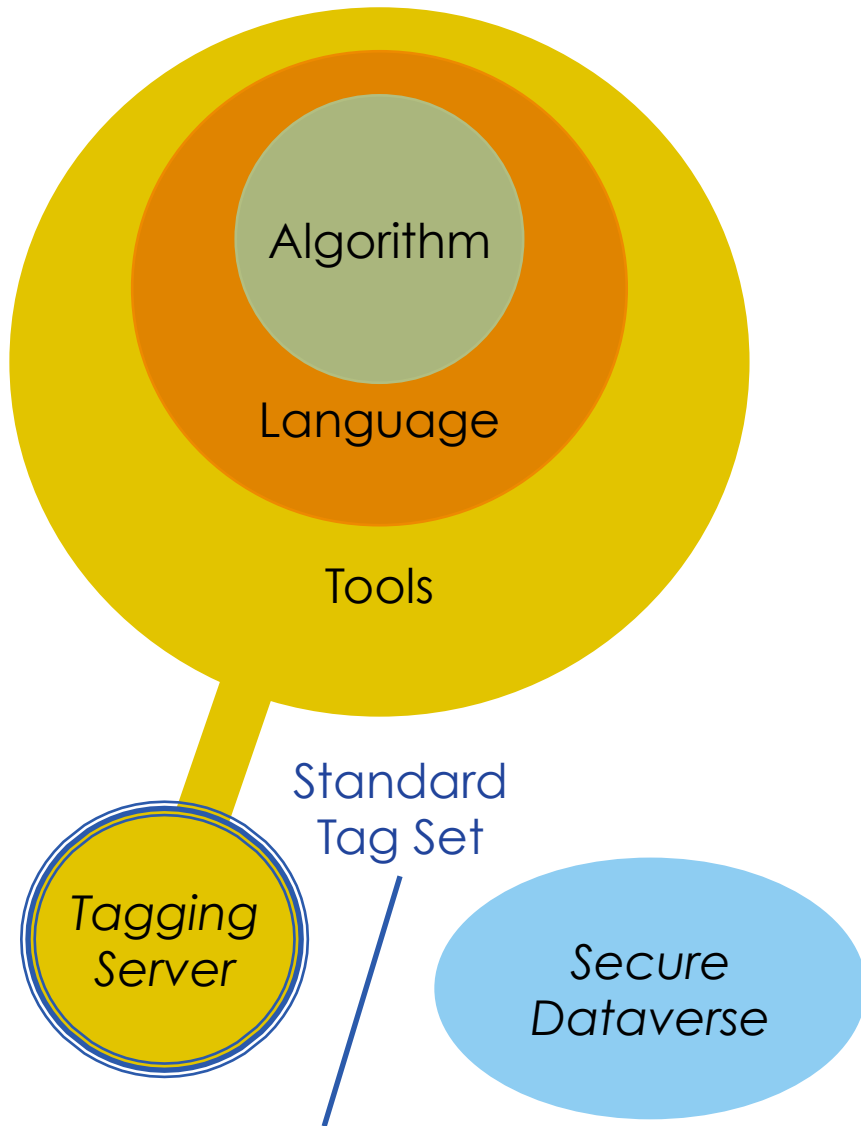


- Editing: Any text editor
- Compiler
- Visualizers
- Runtime Engine
- Java library
- Command-line Runner

Tools: Visualizations



Tagging Server

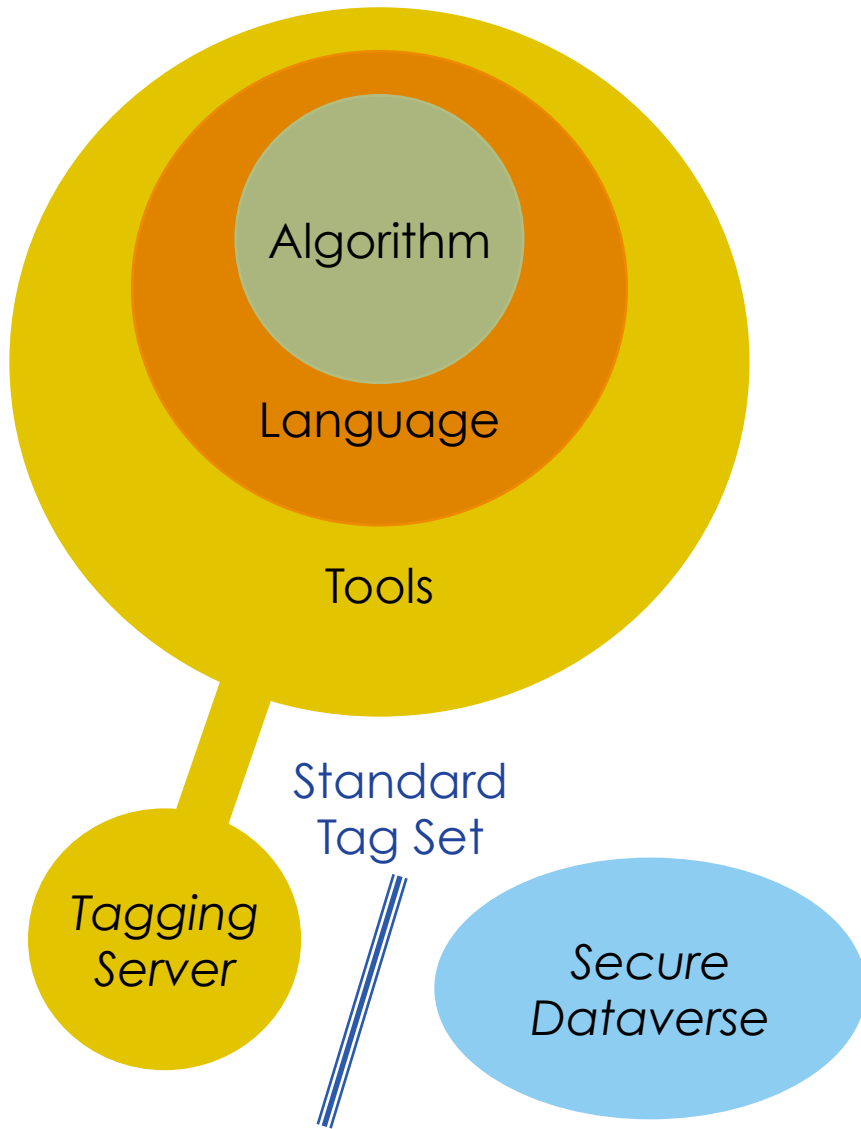


- Web-based GUI for the runtime engine
- Focus on usability
- Integration with other systems, most notably data repositories such as Dataverse, via API
- Will allow other teams to develop tagging interviews

Tagging Server Demo

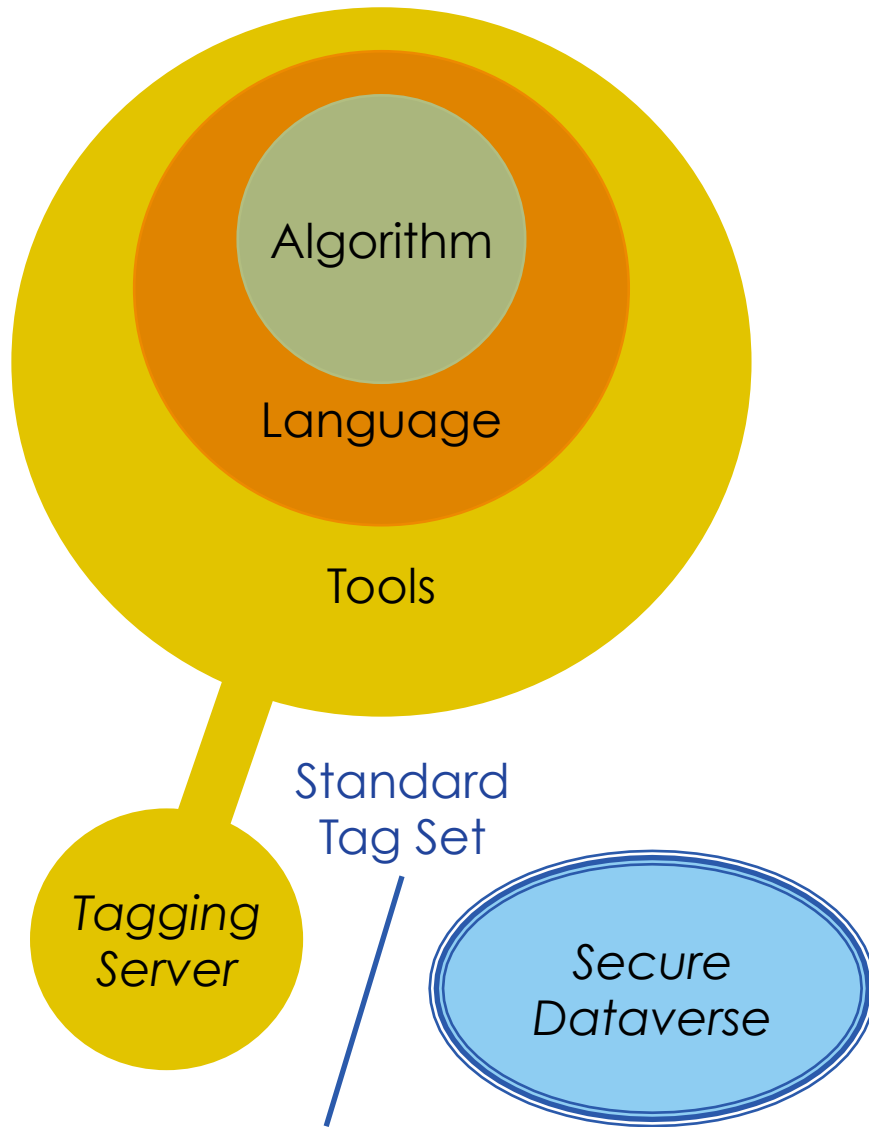
<http://www.datatags.org>

Standard Tag Set



- Allows the tagging process to be machine-actionable
- Data repositories will recognize the set, and will know how to operate according its possible tagging values

Secure Dataverse



- A data repository that can interpret a standard set of data tag, and handle datasets accordingly
- Tagging the data is part of the data ingest process

Learn more at: <http://datascience.iq.harvard.edu>

Data Science

*Research Frameworks for Data-Intensive Science,
Analytical Tools and Data Stewardship*



Zelig Dataverse TwoRavens DataTags Consilience RBuild

About Us

Data Science at IQSS combines expertise in software engineering, statistical innovation and data curation. Meet our team.

THANKS [@mercecrosas](#) [@michbarsinai](#)