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
Acrhitect, Senior Software Engineer,
Institute for Quantitative Social Science, Harvard University

http://datascience.iq.harvard.edu

Introduction to Dataverse

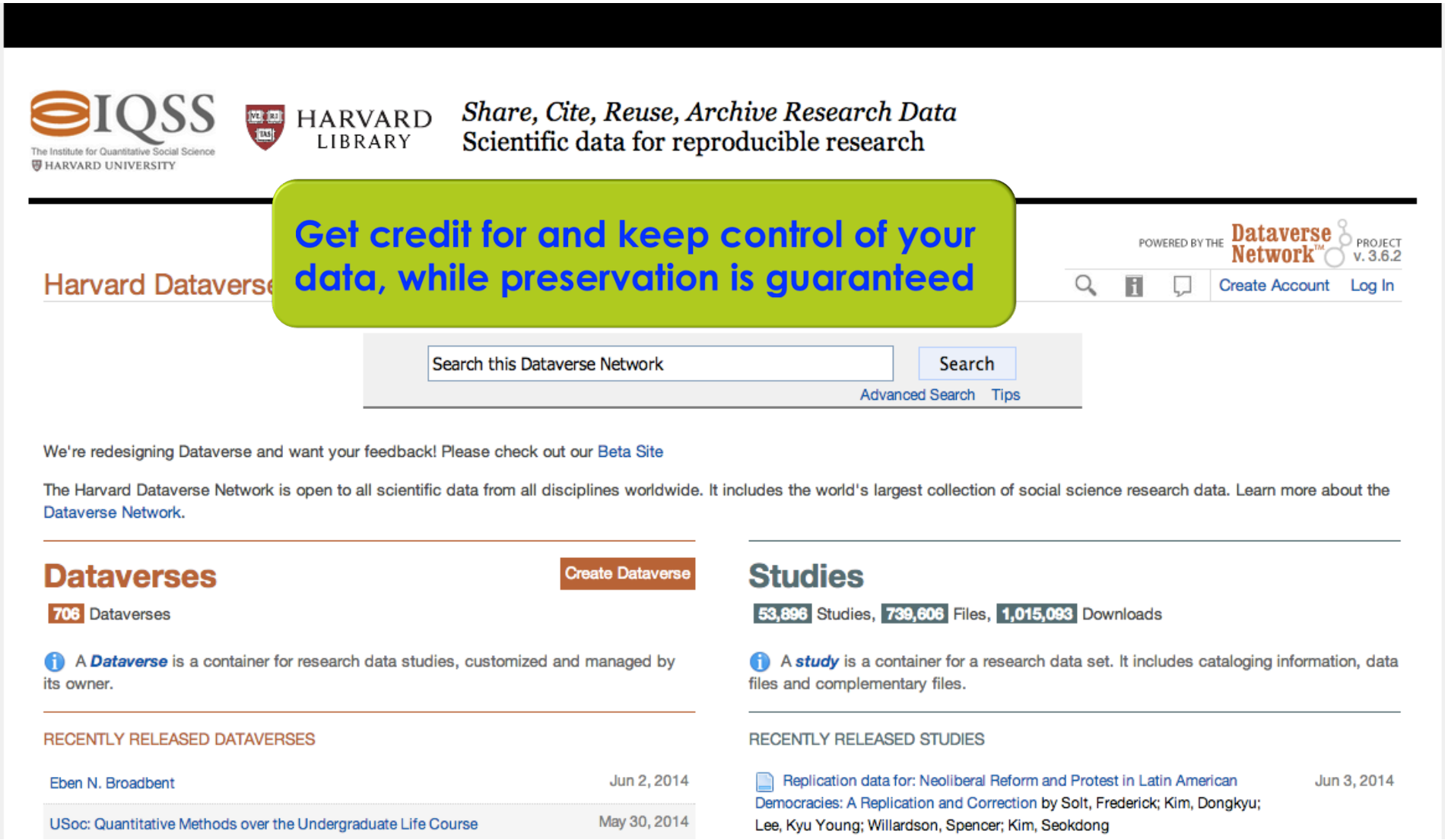
Dataverse Software

- 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 .

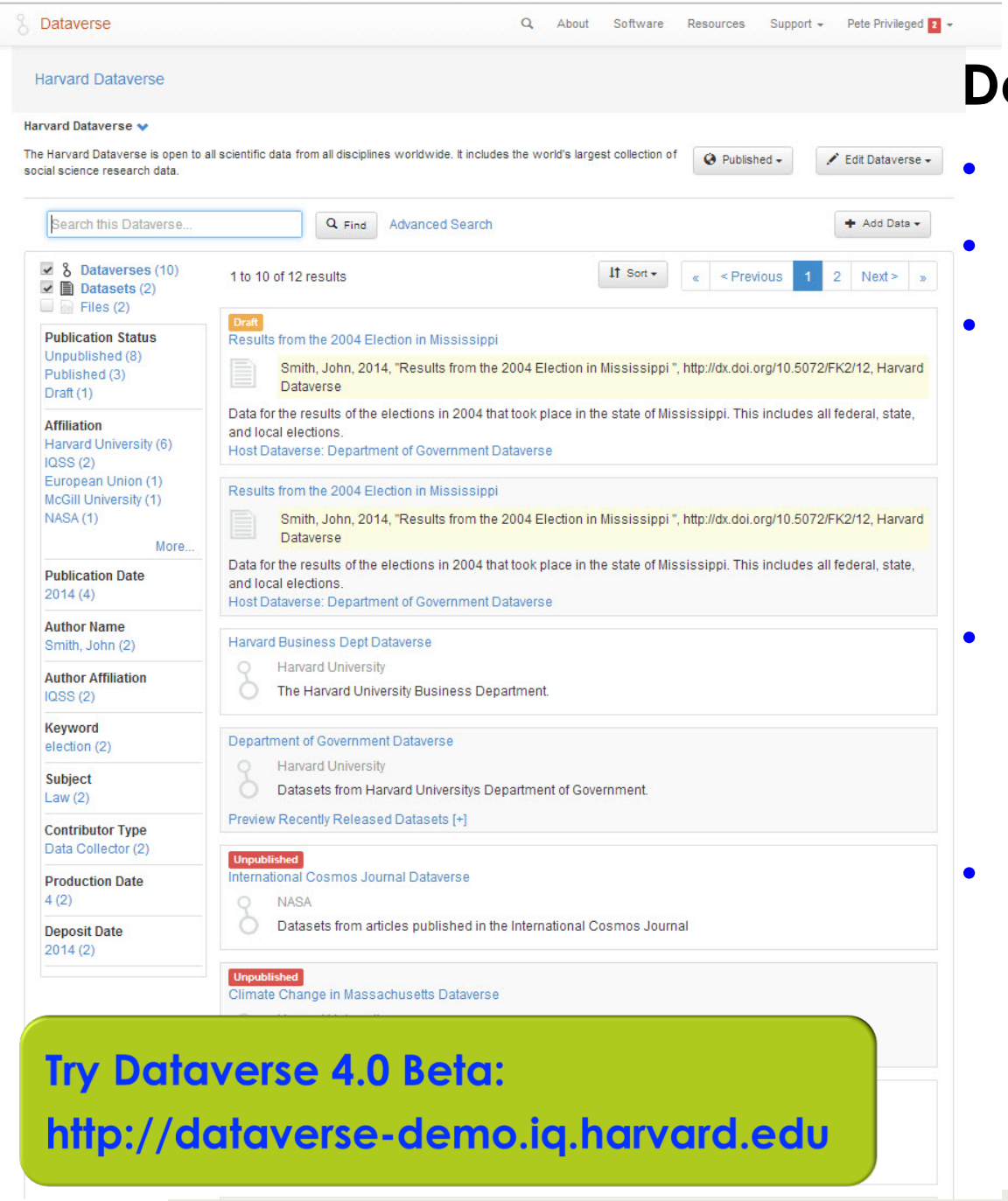
Find and publish data at: http://thedata.harvard.edu



Dataverse Features

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



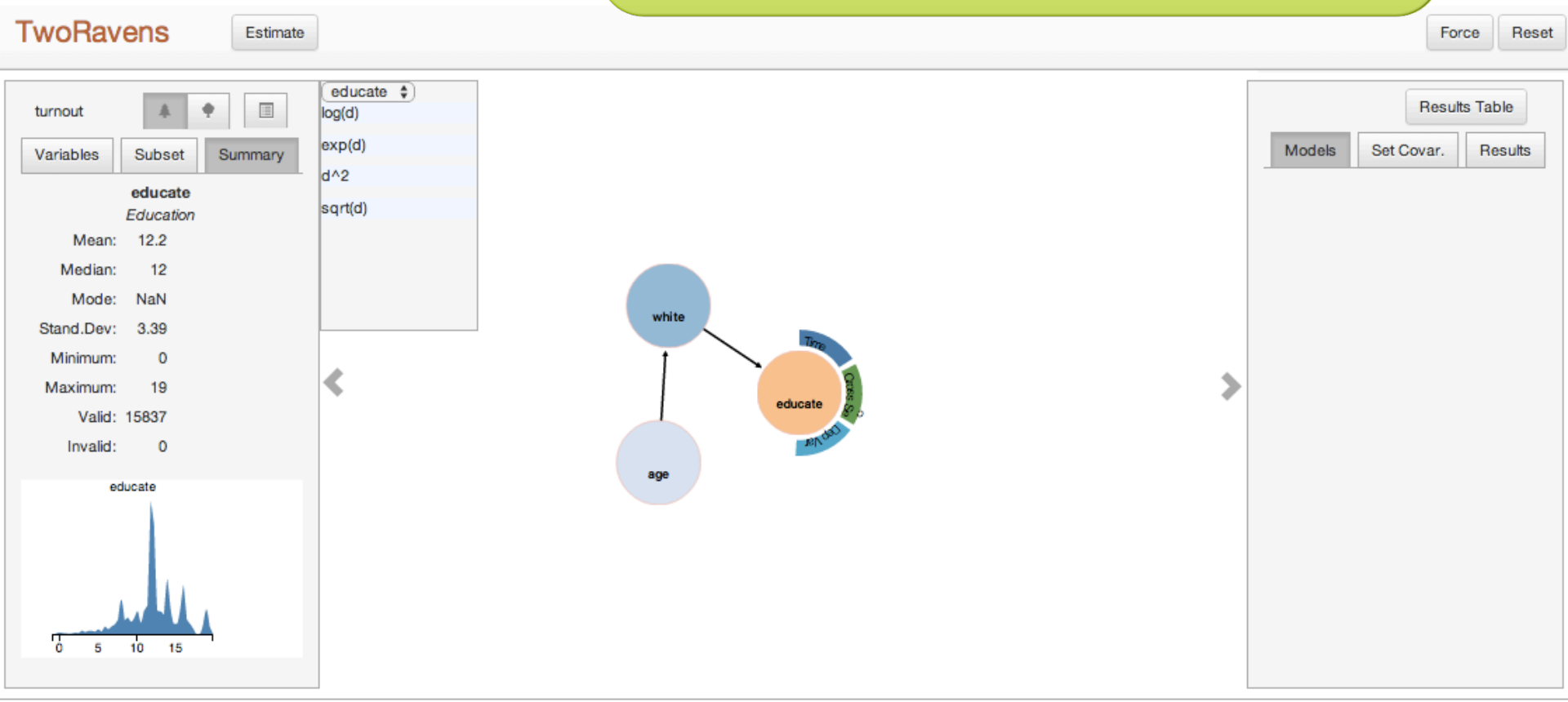
Try Dataverse 4.0 Beta:
http://dataverse-demo.iq.harvard.edu

Dataverse 4.0 (Fall 2014)

- 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**

TwoRavens

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



Citation Metadata

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.

Keyword data publication

Subject *
- [ ] Mathematical Sciences
- [ ] Physics
- [x] Social Sciences
- [ ] Other

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

Social Science and Humanities Metadata

Topic Classification
Term | Vocabulary | URL

Software
Name | Version

Series
Name | Information

Time Period Covered
Start (YYYY-MM-DD) | End (YYYY-MM-DD)

Date of Collection
Start (YYYY-MM-DD) | End (YYYY-MM-DD)

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

Astronomy and Astrophysics Metadata

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** (YYYY-MM-DD) — **End** (YYYY-MM-DD)

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

Organism

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

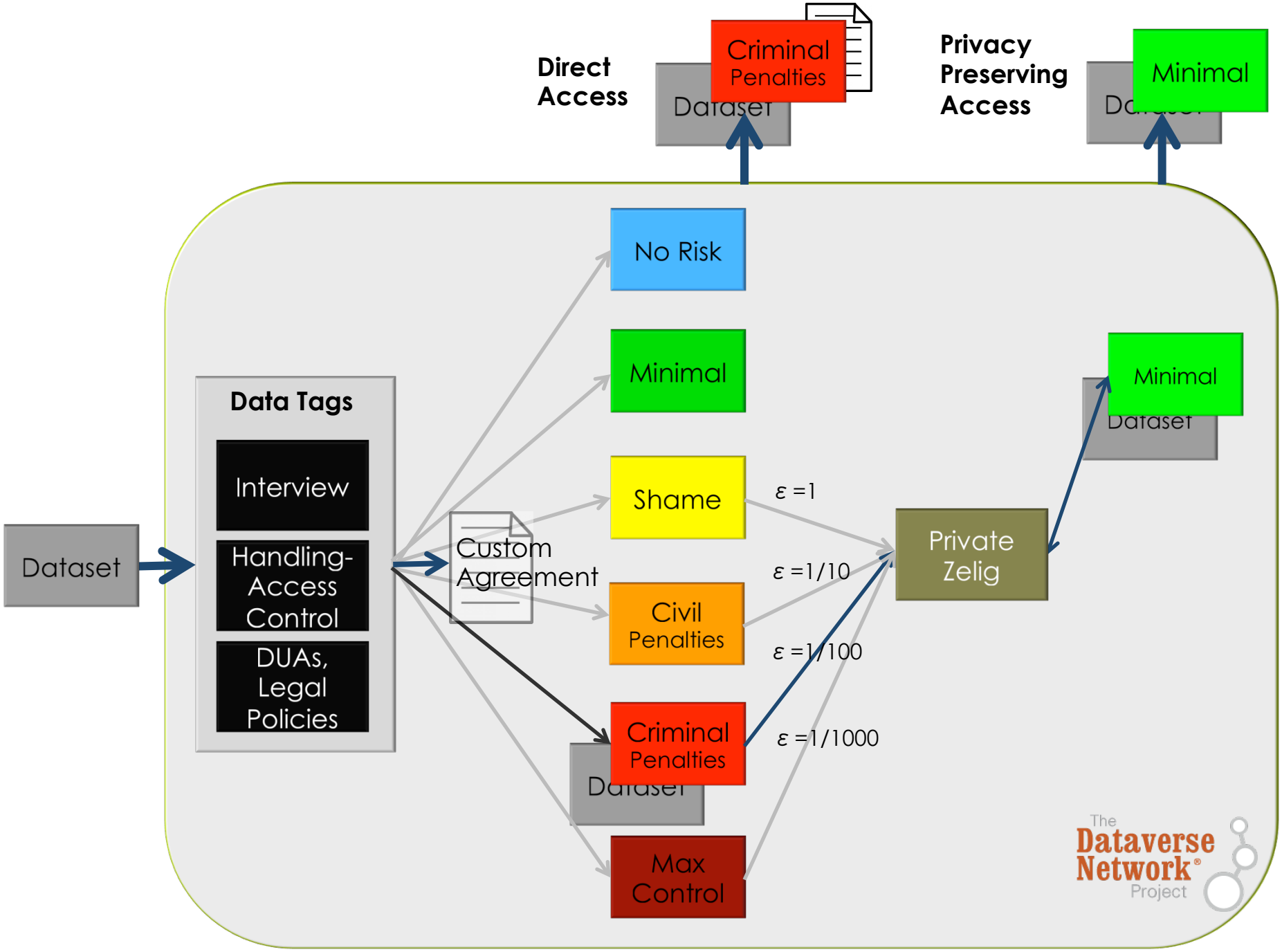
Cell Type

+

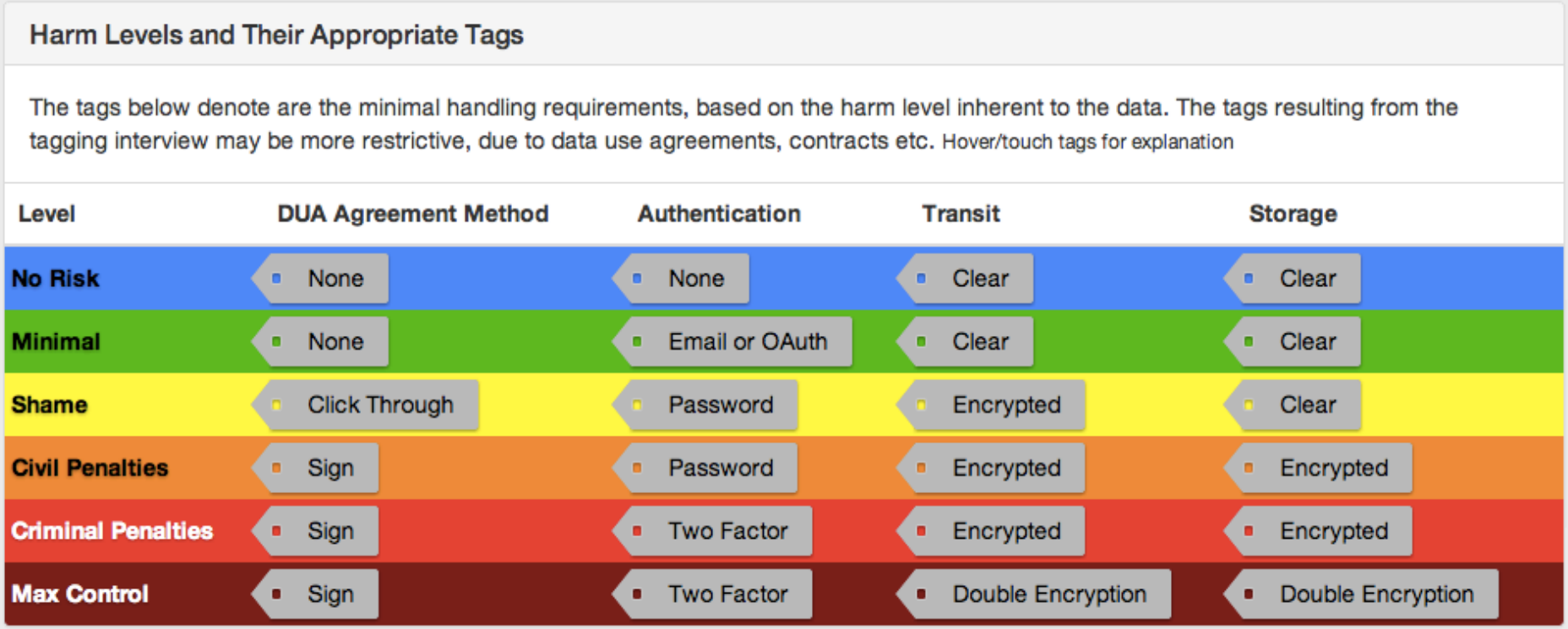
Bio Metadata:
Compliant with ISA-Tab schema, plus biomedical ontologies

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


Direct Access
Criminal Penalties
Dataset
Privacy Preserving Access
Minimal
Data Tags
Interview
Handling-Access Control
DUAs, Legal Policies
Dataset
Custom Agreement
No Risk
Minimal
Shame
Civil Penalties
Criminal Penalties
Max Control
ε =1
ε =1/10
ε =1/100
ε =1/1000
Private Zelig
Minimal
The Dataverse Network® Project

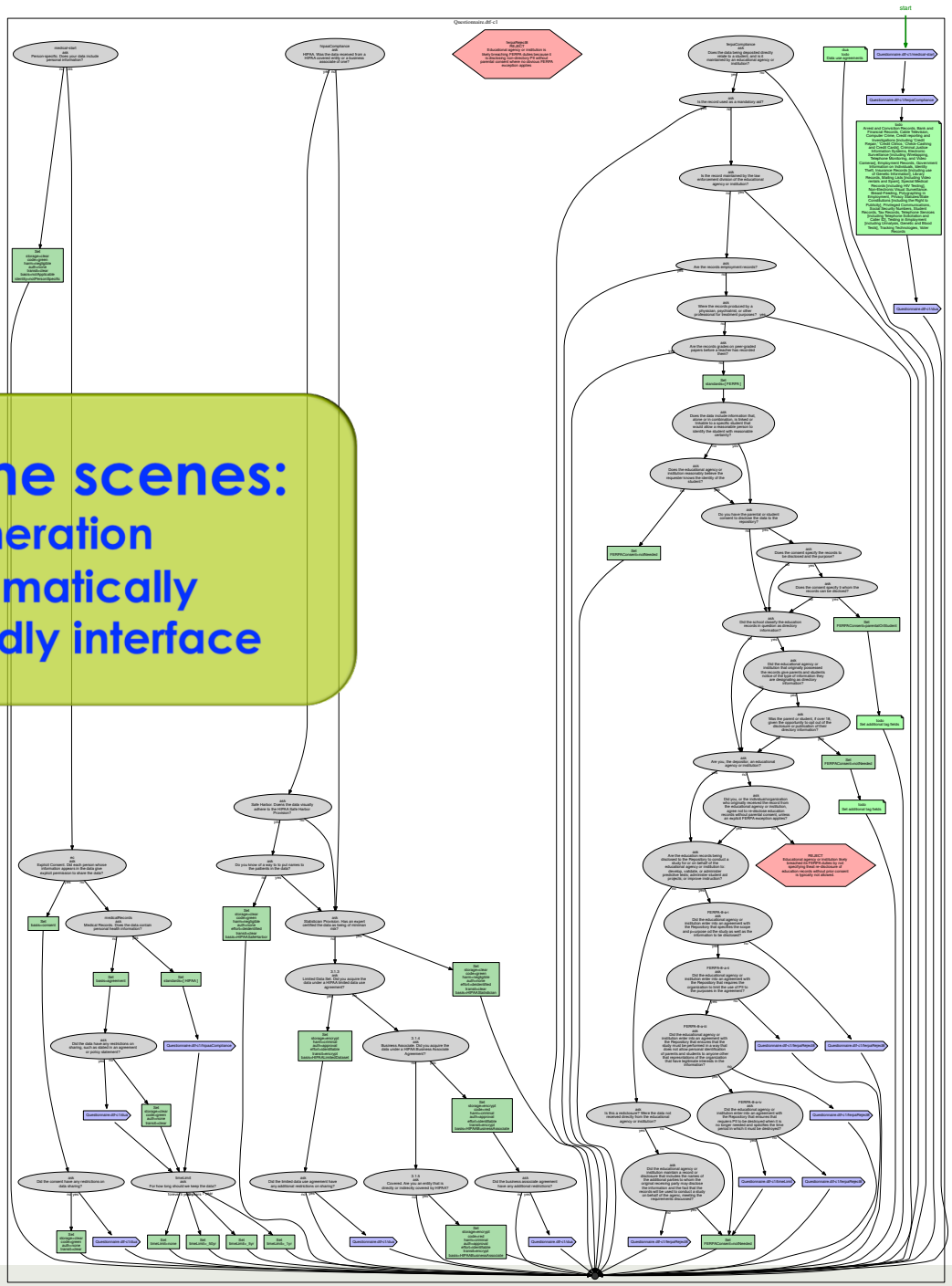
Try our new Beta version: http://datatags.org



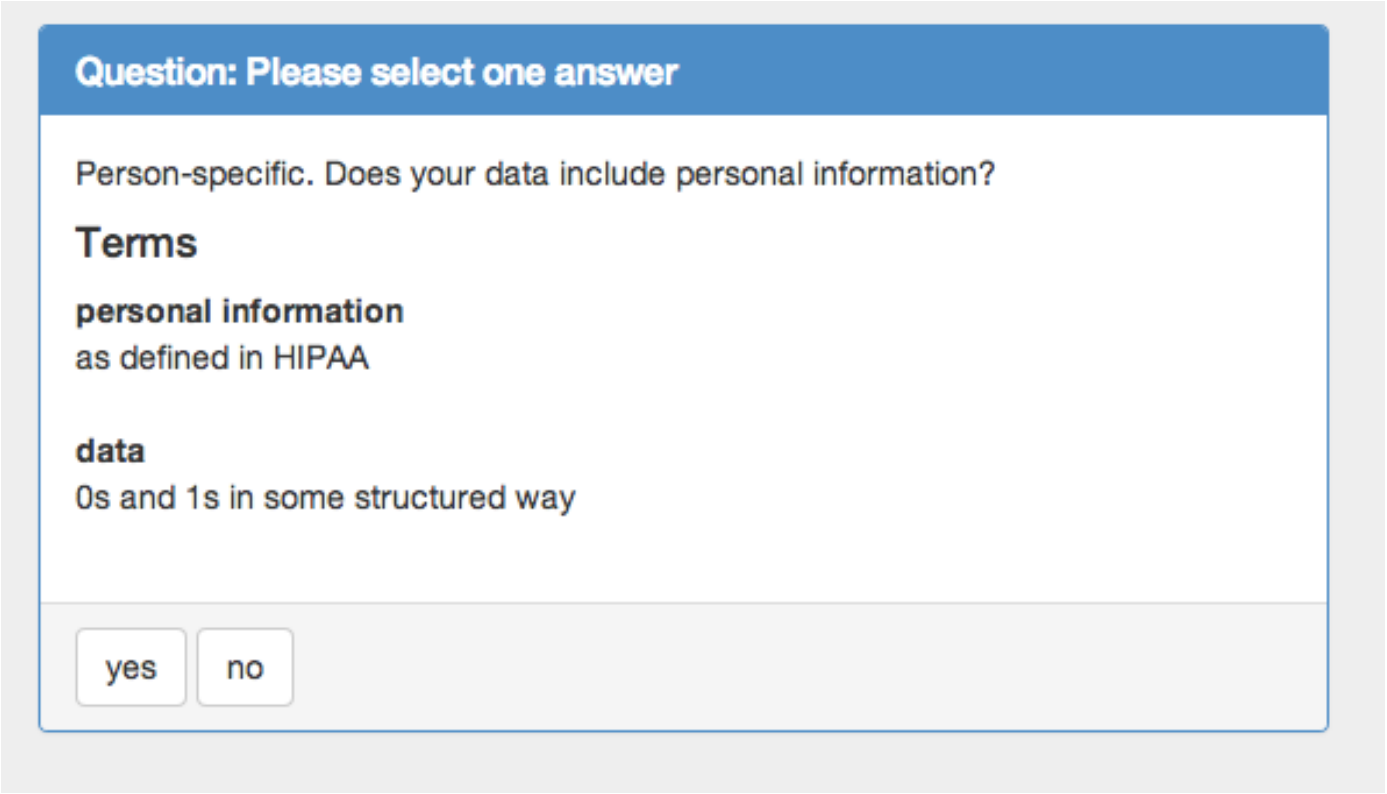
Currently supporting HIPAA and FERPA (and DUAs)

DataTags behind the scenes:

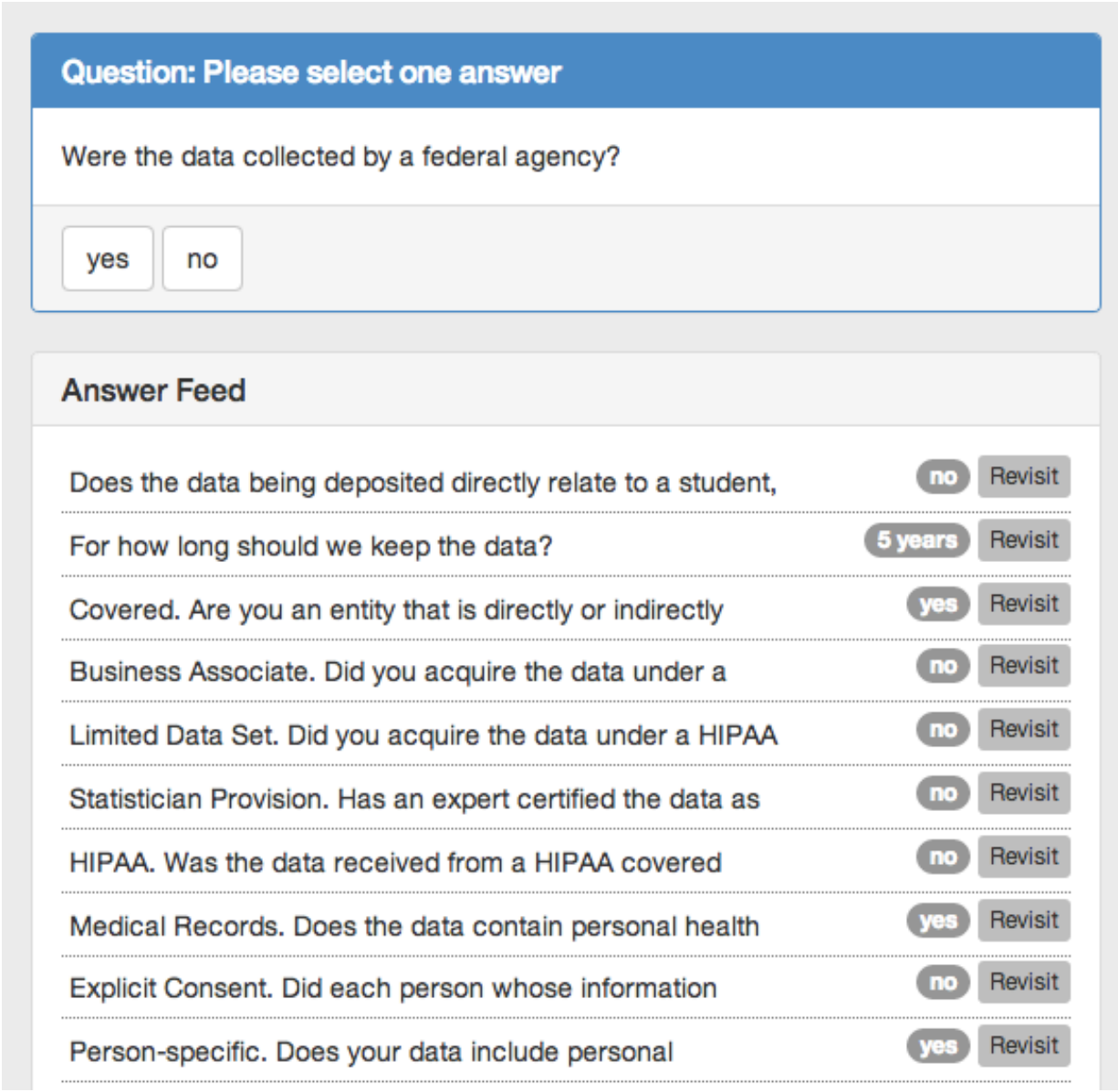
A complex interview generation framework, which is automatically converted to a user-friendly interface



Interview Example: First question …



Interview Example: After several questions …



Interview Example: … and a Final Tag

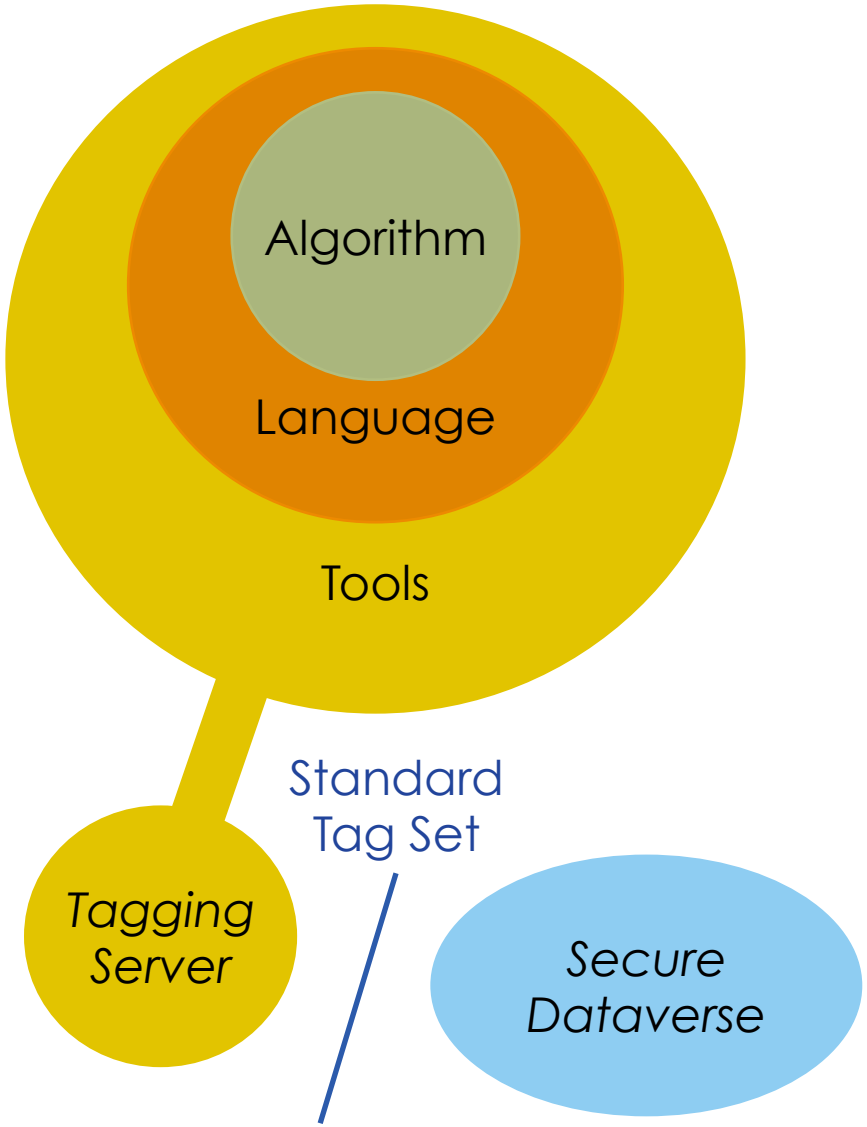
Your dataset is tagged as

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

Full Tags

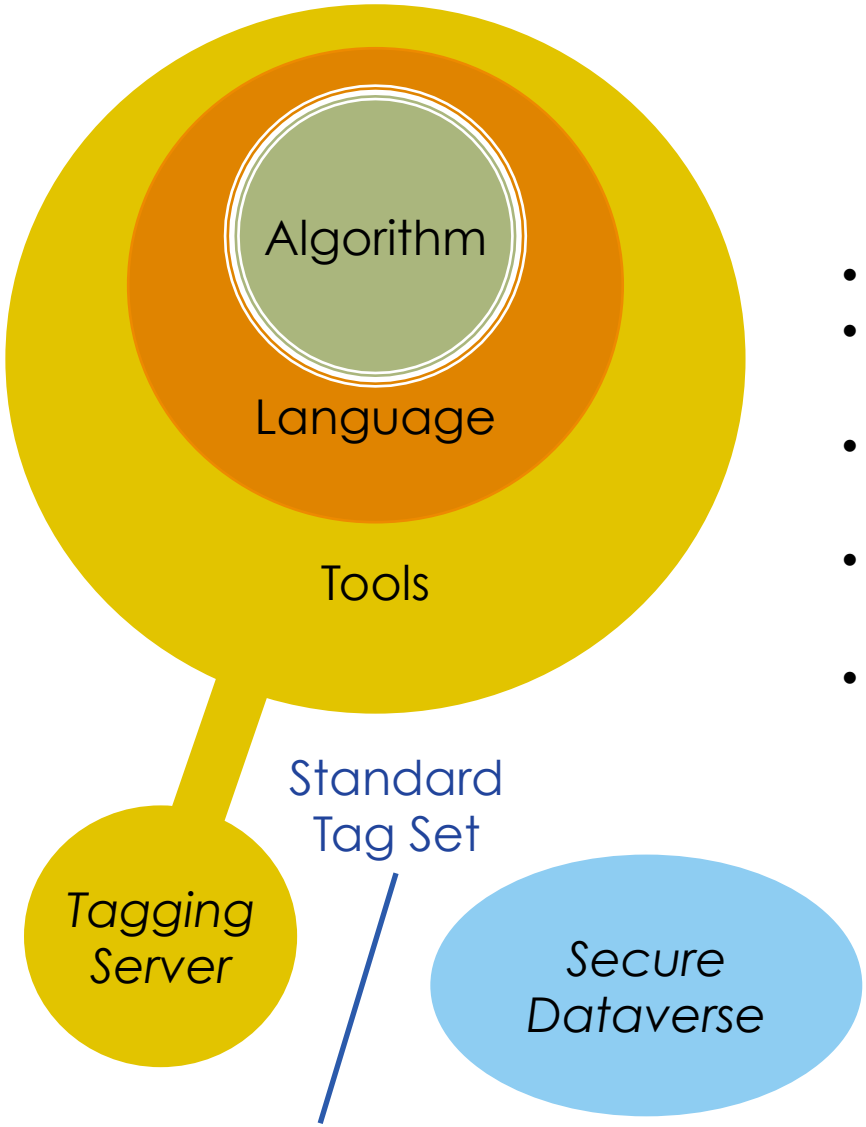
- DataTags
  - code: red
  - DataType
    - harm: criminal
    - effort: identifiable
    - standards: HIPAA
  - Handling
    - storage: encrypt
    - auth: approval
    - transit: encrypt
  - basis: HIPAABusinessAssociate
  - DUA
    - timeLimit: _5yr

Project Structure



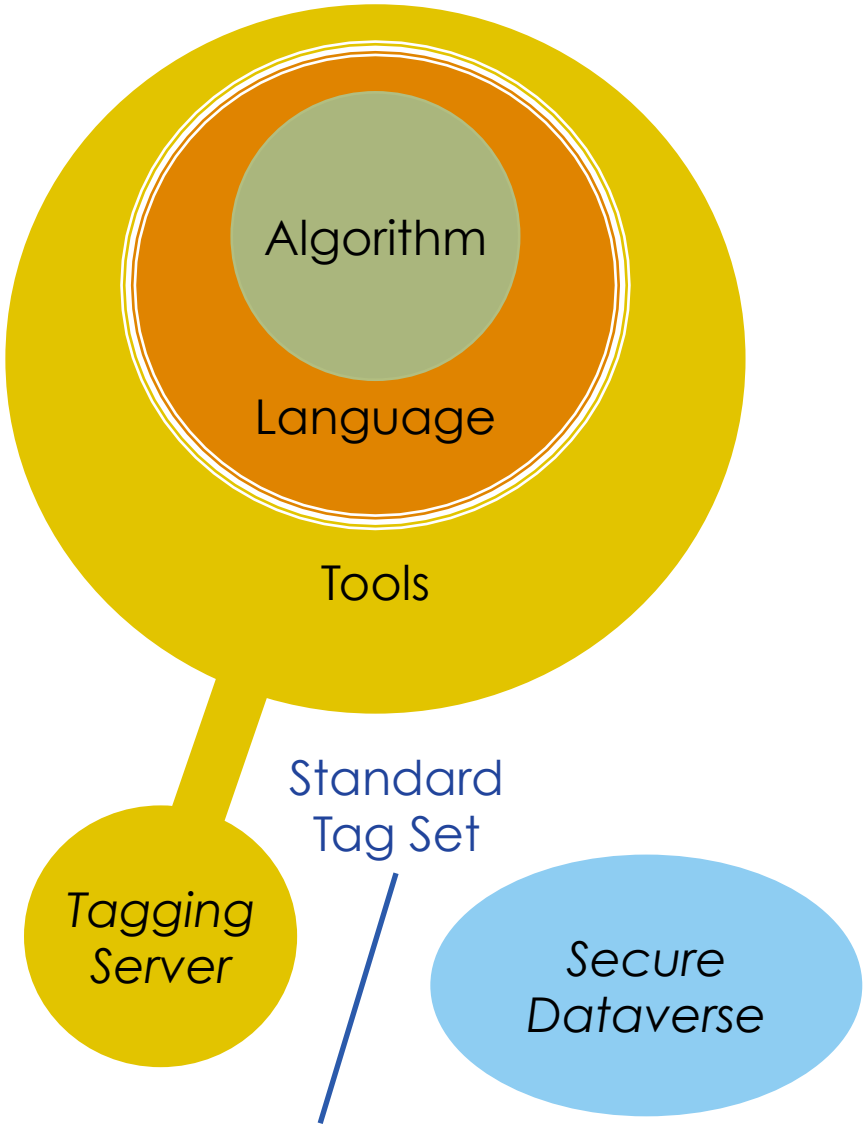
The DataTags project consists of several distinct components.

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

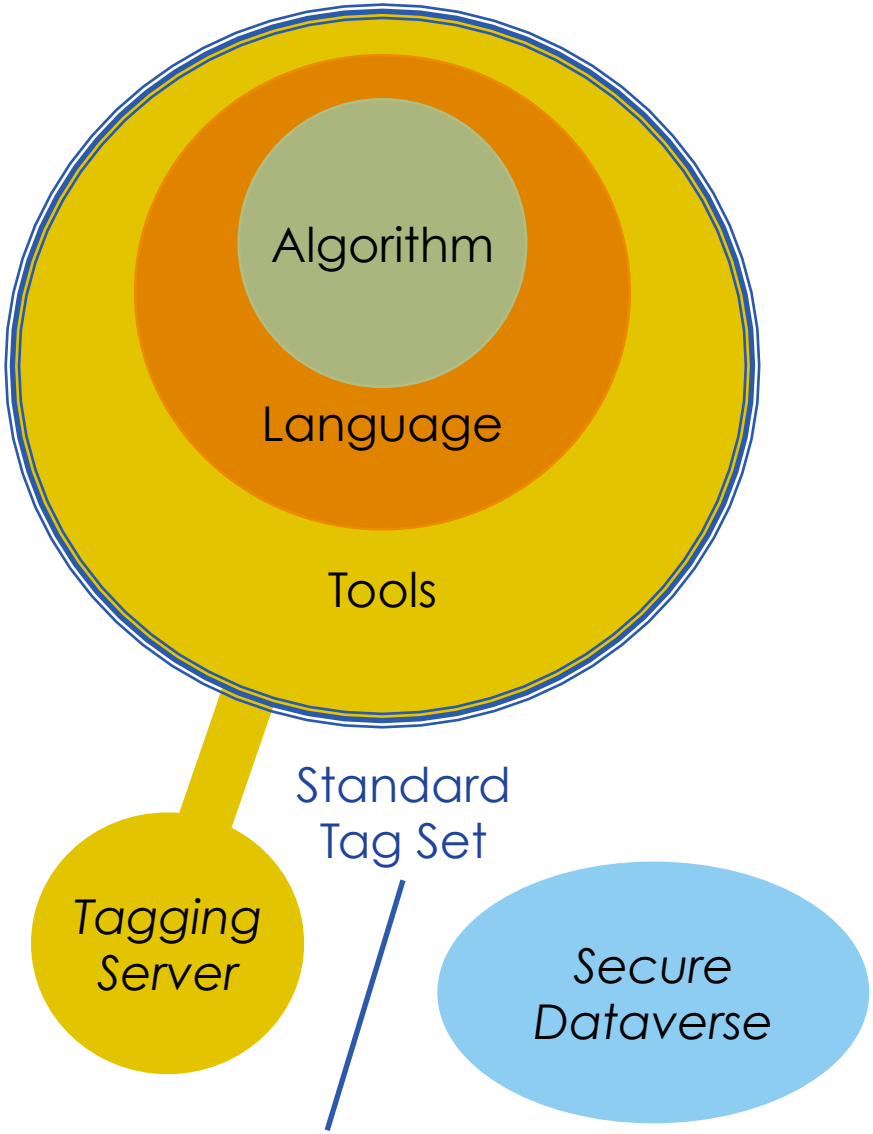
Tag Definition

```
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

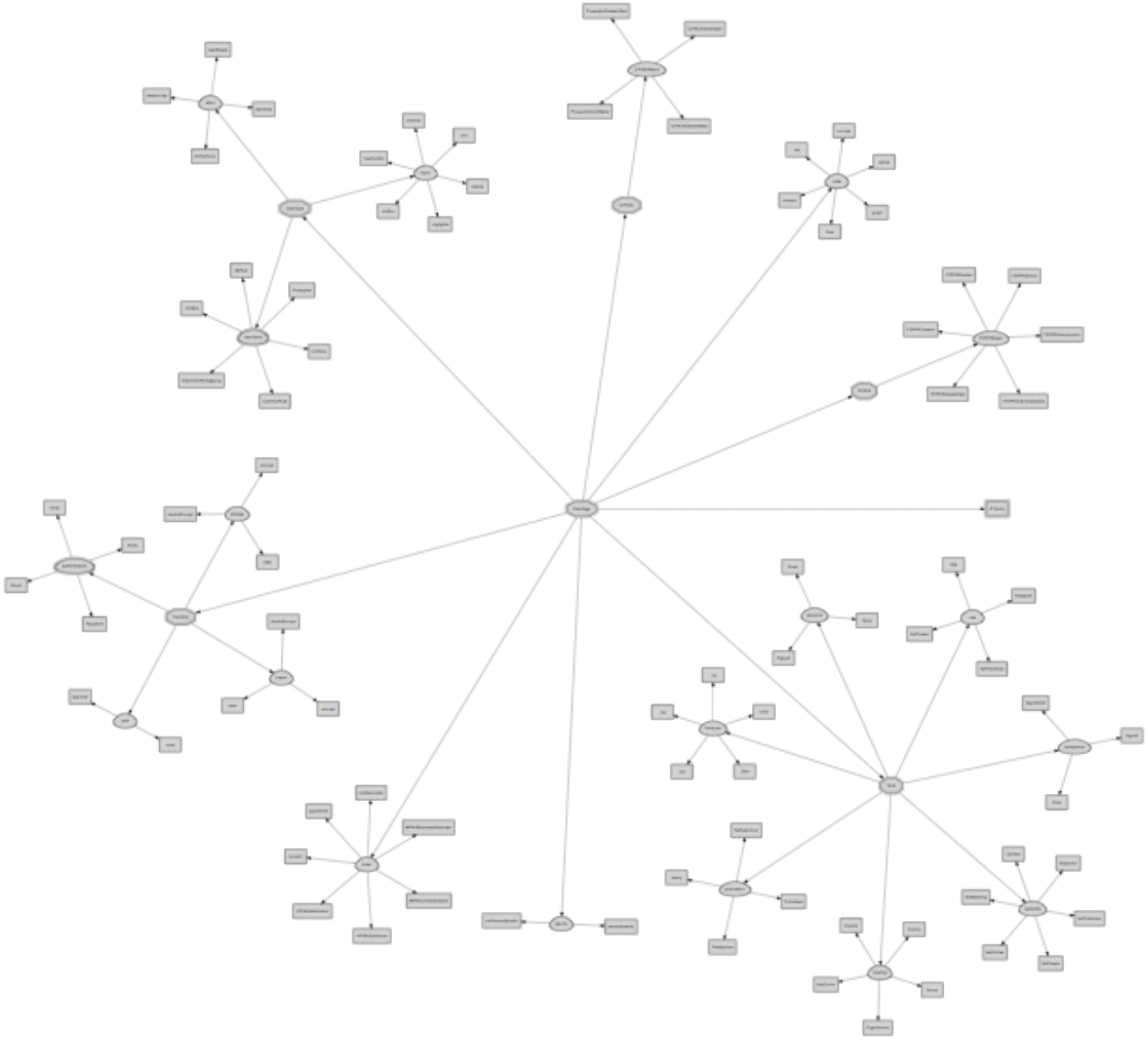
```
(>medical-start< ask:
    (text: Person-specific. Does your data include personal information?)
    (terms:
        (data: 0s and 1s in some structured way)
        (personal information: as defined in HIPAA))
    (no:
        (set: code=green, storage=clear, transit=clear, auth=none,
            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

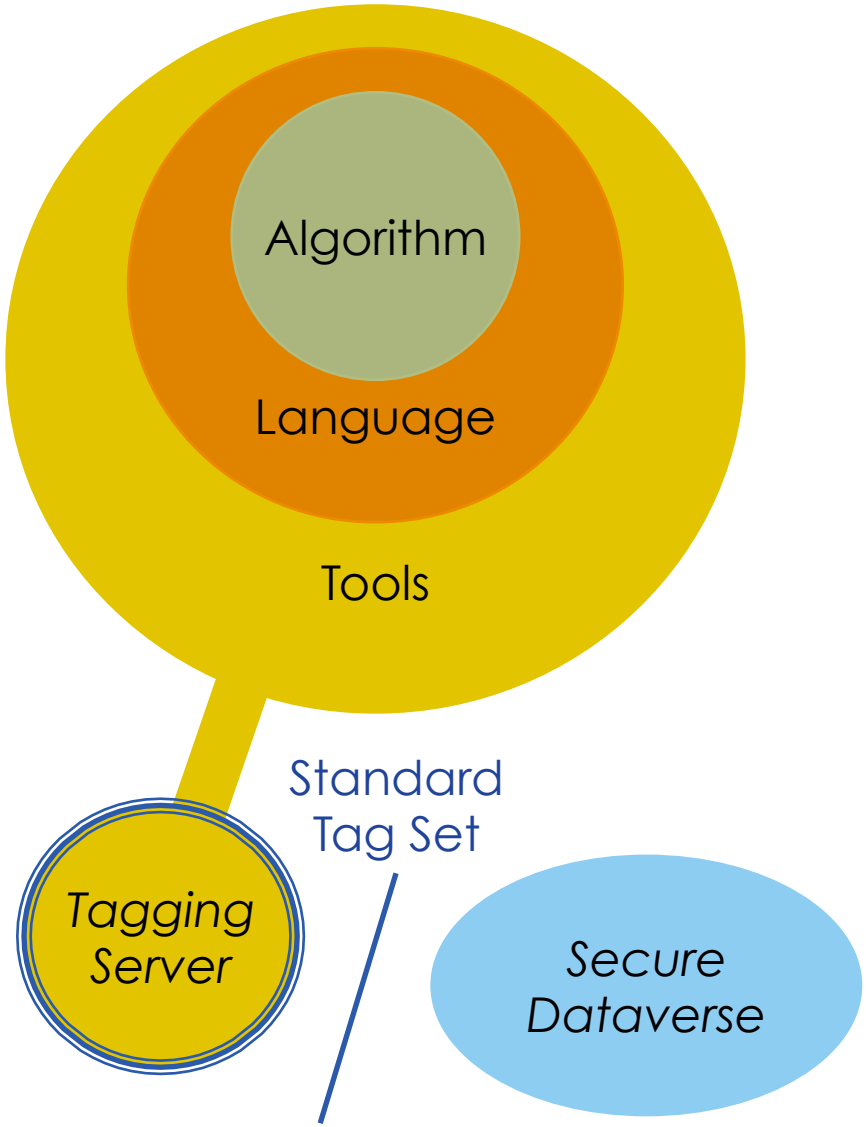


- Editing: Any text editor
- Compiler
- Visualizers
- Runtime Engine
- Java library
- Command-line Runner

Tools: Visualizations



Tagging Server

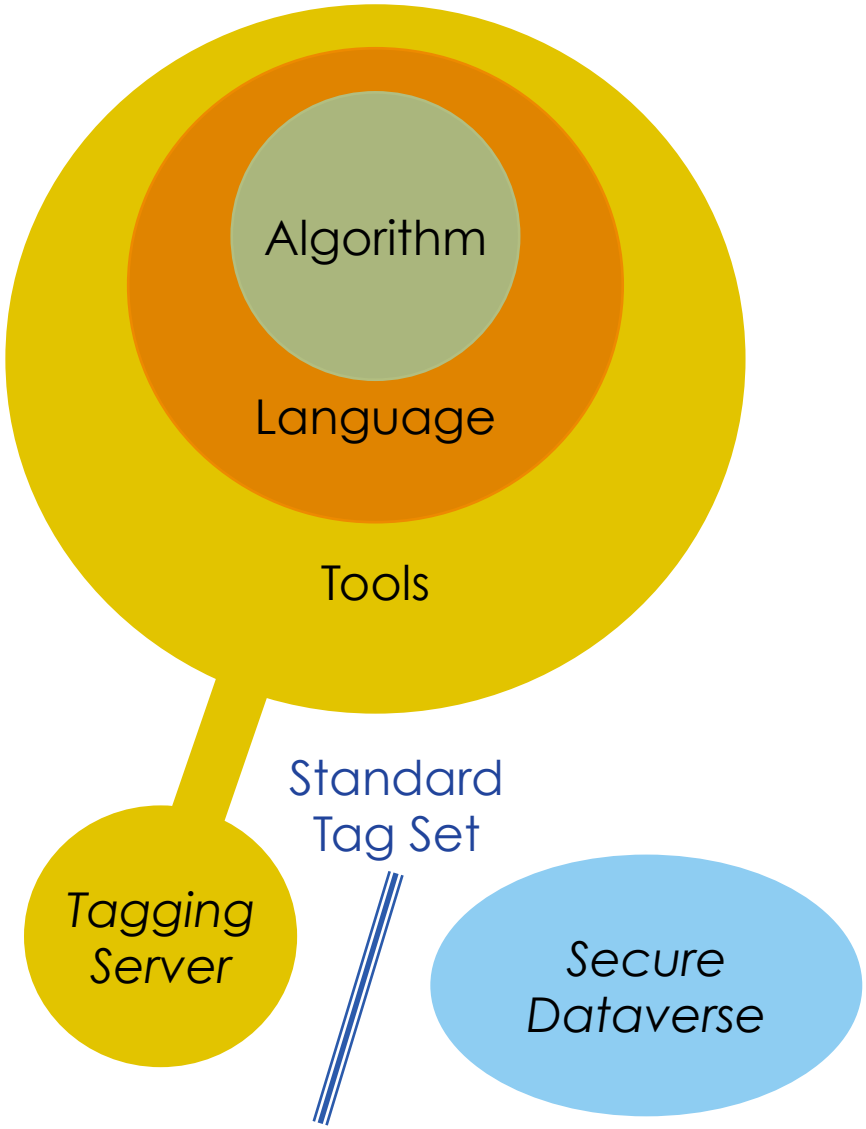


- Web-based GUI for the runtime engine
- Focus on usability
- Integration with other systems, most notably data repositories such as Dataverse, via API
- Will allow other teams to develop tagging interviews

Tagging Server Demo

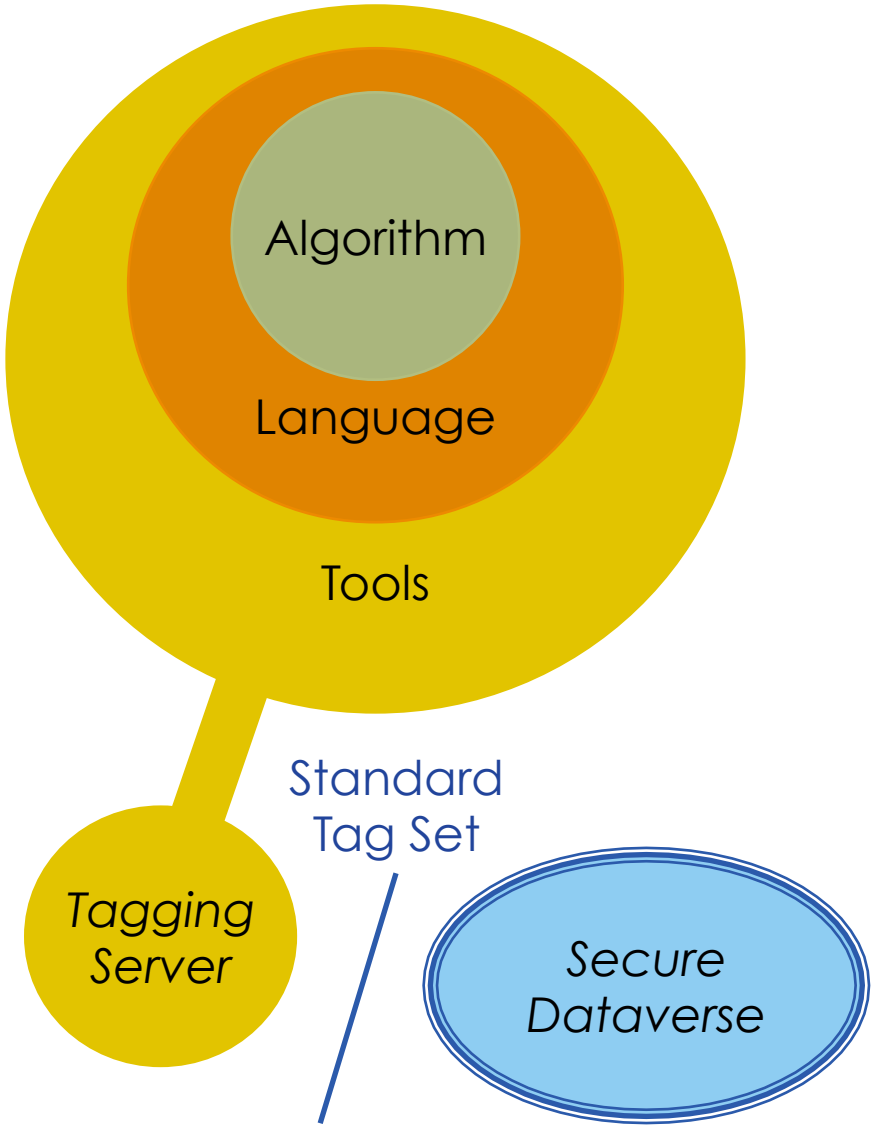
http://www.datatags.org

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

IQSS
The Institute for Quantiative Social Science
HARVARD UNIVERSITY

Zelig Dataverse TwoRavens DataTags Consilience RBuild

THANKS @mercecrosas @michbarsinai