Introduction to Data Sharing and Dataverse, and Dataverse integration with Zelig and DataTags

REU Orientation, June 10, 2014

Mercè Crosas

Director of Data Science, IQSS

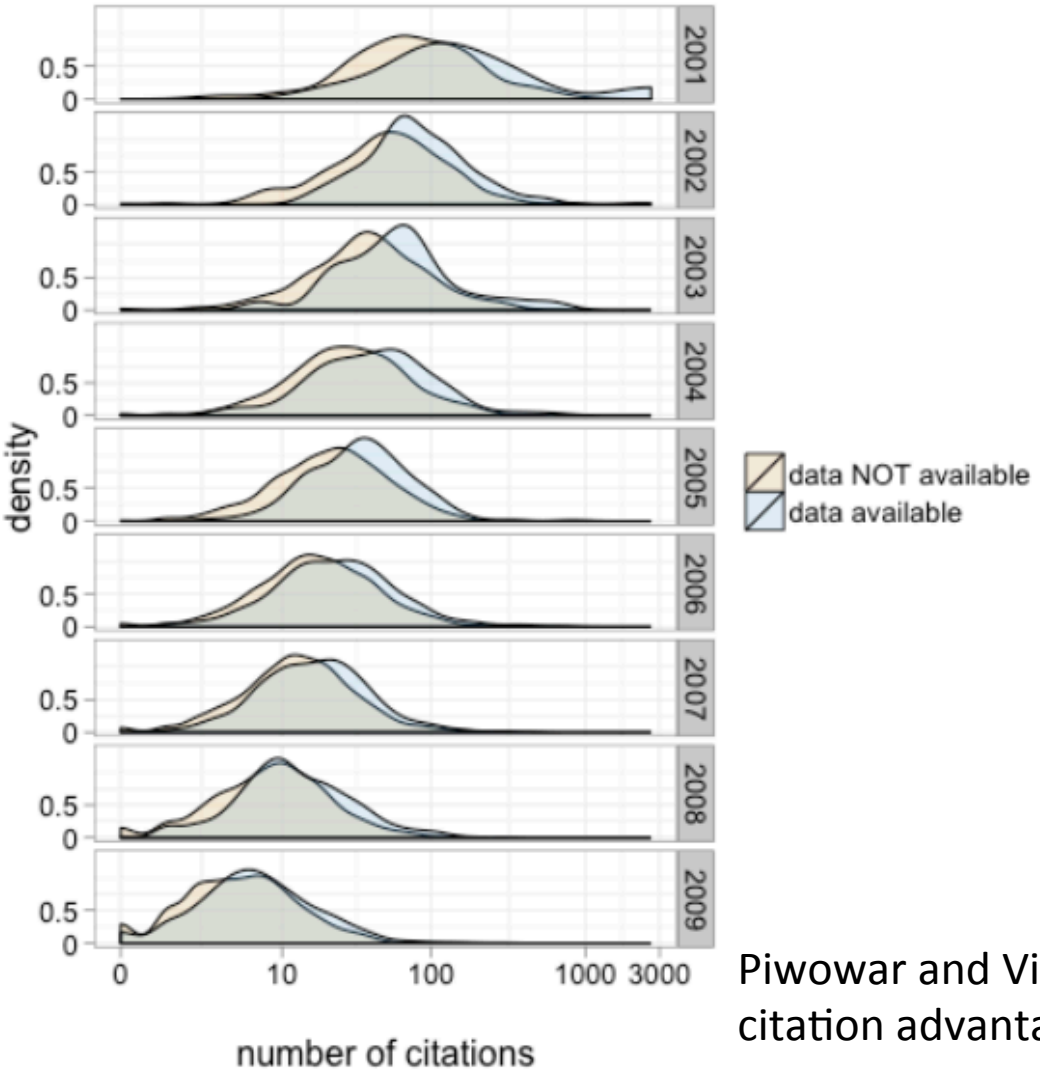
Data sharing is good for science

Making your research data accessible is important:

- To reproduce research
- To make public assets available to the public
- To leverage investments in research data
- To advance research and innovation

Borgman, Oct 2013, “Why you should care about open data” Open Access Week Talk

...and good for you

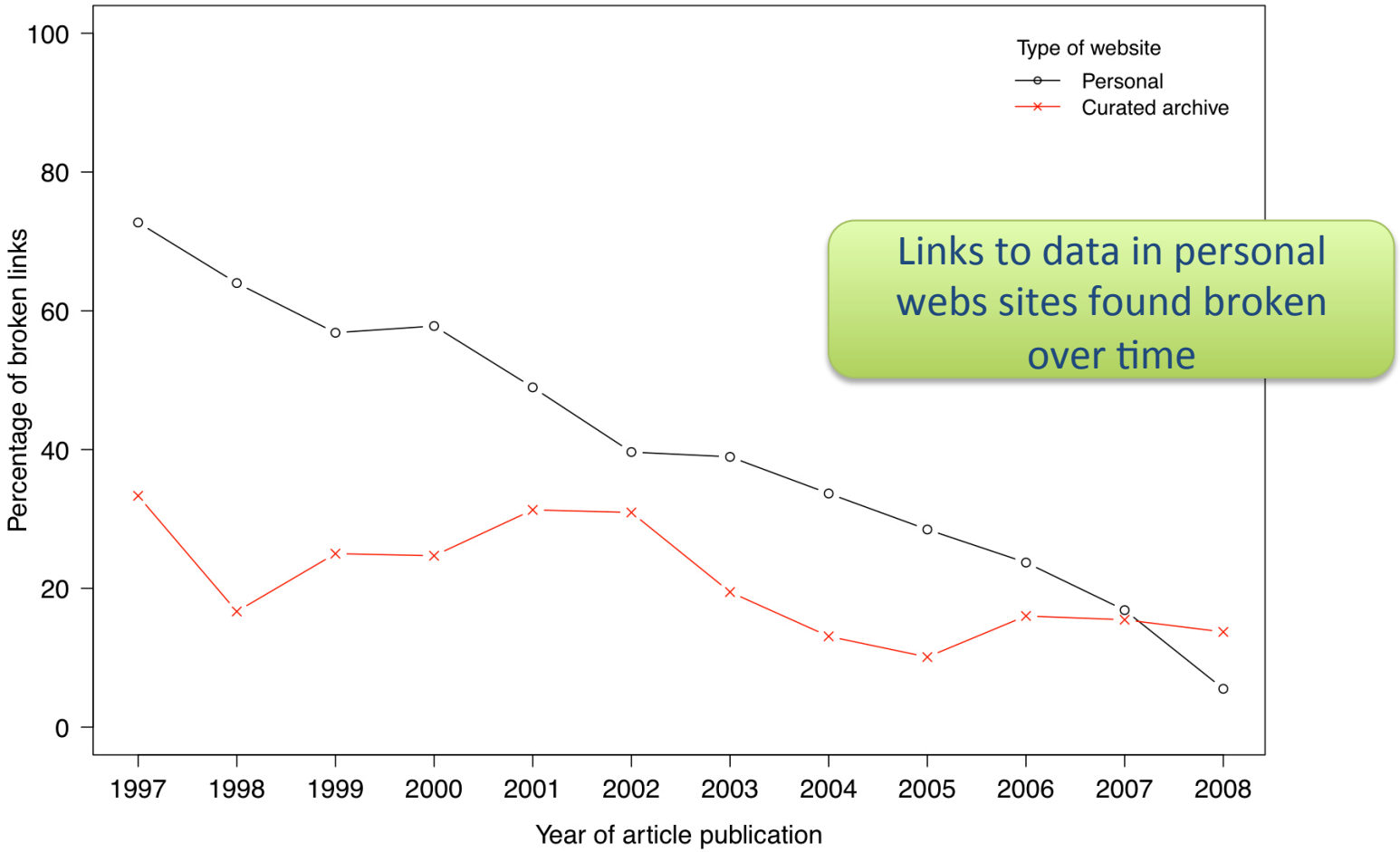


10,555 studies that created gene expression microarray data:

- Studies that share data received 9% **more citations**
- Authors published most papers using their own data within 2 years
- Data reuse papers by third-party investigators continued for 6 years

Piwowar and Vision (2013), Data reuse and the open data citation advantage. PeerJ 1:e175; DOI 10.7717/peerj.175

But data sharing must include long-term accessibility



Pepe, Goodman, Muench, Crosas, Erdmann, 2014 "Sharing, Archiving and Citing Data in Astronomy" *Forthcoming*

We hosted a workshop at Harvard University to address issues about data sharing and reuse, and the result was: "10 Rules"

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Editorial

Ten Simple Rules for the Care and Feeding of Scientific Data

Alyssa Goodman[1], Alberto Pepe[1]*, Alexander W. Blocker[1], Christine L. Borgman[2], Kyle Cranmer[3], Merce Crosas[1], Rosanne Di Stefano[1], Yolanda Gil[4], Paul Groth[5], Margaret Hedstrom[6], David W. Hogg[3], Vinay Kashyap[1], Ashish Mahabal[7], Aneta Siemiginowska[1], Aleksandra Slavkovic[8]

1 Harvard University, Cambridge, Massachusetts, United States of America, **2** University of California, Los Angeles, Los Angeles, California, United States of America, **3** New York University, New York, New York, United States of America, **4** University of Southern California, Los Angeles, Los Angeles, California, United States of America, **5** Vrije Universiteit Amsterdam, Amsterdam, The Netherlands, **6** University of Michigan, Ann Arbor, Michigan, United States of America, **7** California Institute of Technology, Pasadena, California, United States of America, **8** Pennsylvania State University, State College, Pennsylvania, United States of America

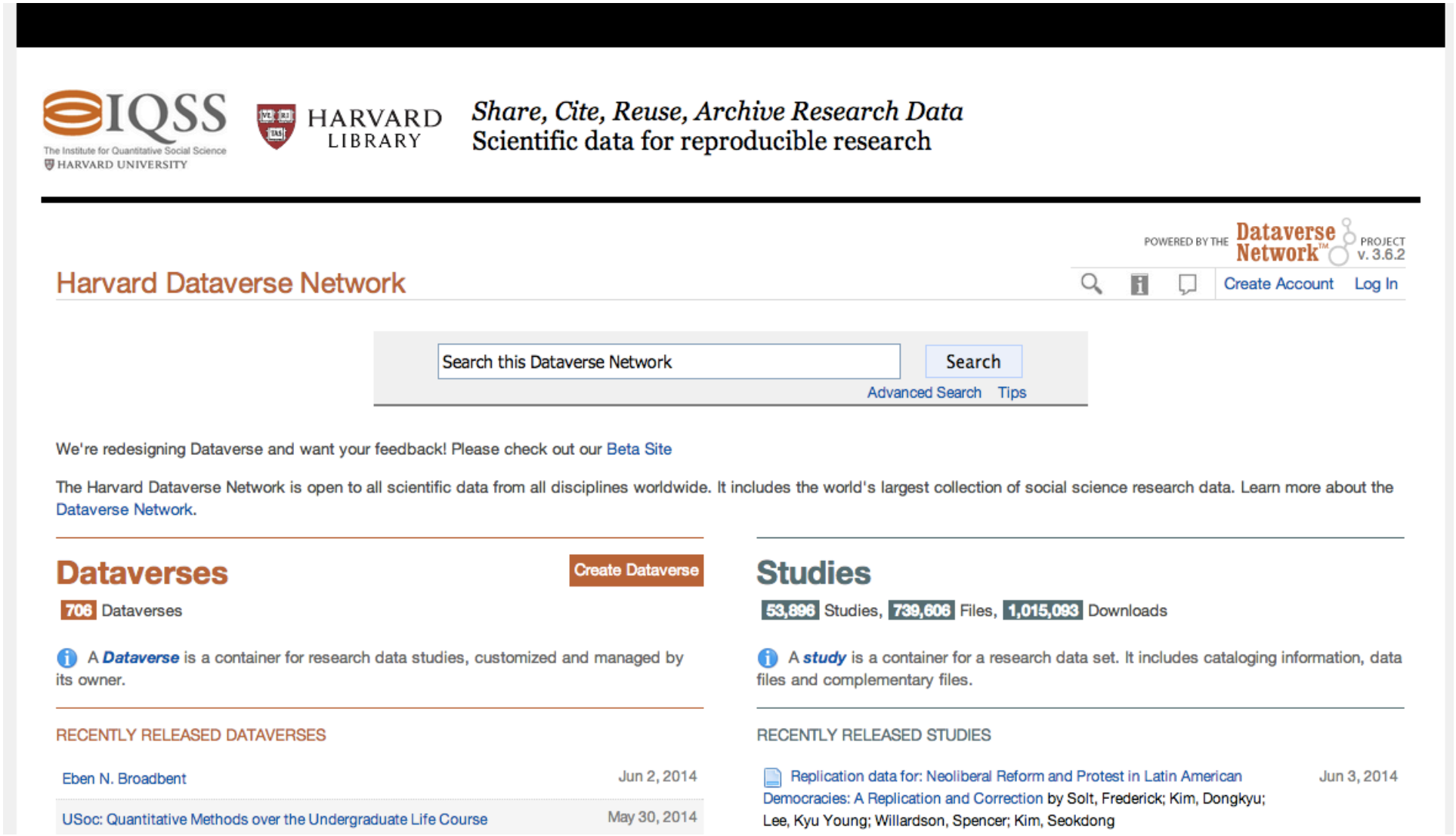
Ten Simple Rules

- Rule 1: Love your data, and let others love them too
- Rule 2: Share your data online, with a permanent identifier
- Rule 3: Conduct science with data reuse in mind
- Rule 4: Publish workflow as context
- Rule 5: Link your data to your publications as early as possible
- Rule 6: Publish your code
- Rule 7: Say how you want to get credit for your data
- Rule 8: Foster and use data repositories
- Rule 9: Reward colleagues who share their data properly
- Rule 10: Help establish data science and data scientists as vital

The Dataverse repository as a solution for data sharing

- The Dataverse hosted at Harvard is open and free to all researchers worldwide.
- Serves as a solution to help you follow the 10 Rules.
- Contains already > 53, 000 data sets, the largest general-purpose data repositories in the world.
- The Dataverse open-source software is developed at Harvard's IQSS, by our data science team plus contributors.

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



Benefits of publishing data with Dataverse

What you contribute

- Sufficient information accompanying the data
- Data files with rich metadata

What Dataverse gives you

- Credit for your data through data citation
- Control on how to share your data
- Data exploration and analysis for tabular data
- Long-term data preservation

Sufficient information with the data

The **replication standard** holds that:

Sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party could replicate the results without any additional information from the author.

King, Gary. 1995. Replication, Replication. PS: Political Science and Politics 28: 443–499.

“sufficient information”?

How were the respondents selected? Who did the interviewing? What was the question order? How did you decide which informants to interview or villages to visit? How long did you spend in each community? Did you speak to people in their language or through an interpreter? Which version of the ICPSR file did you extract information from? How knowledgeable were the coders? How frequently did the coders agree? Exactly what codes were originally generated and what were all the recodes performed? Precisely which measure of unemployment was used? What were the exact rules used for conducting the content analysis? When did the time series begin and end? What countries were included in your study and how were they chosen? What statistical procedures were used? What method of numerical optimization did you choose? Which computer program was used? How did you fill in or delete missing data?

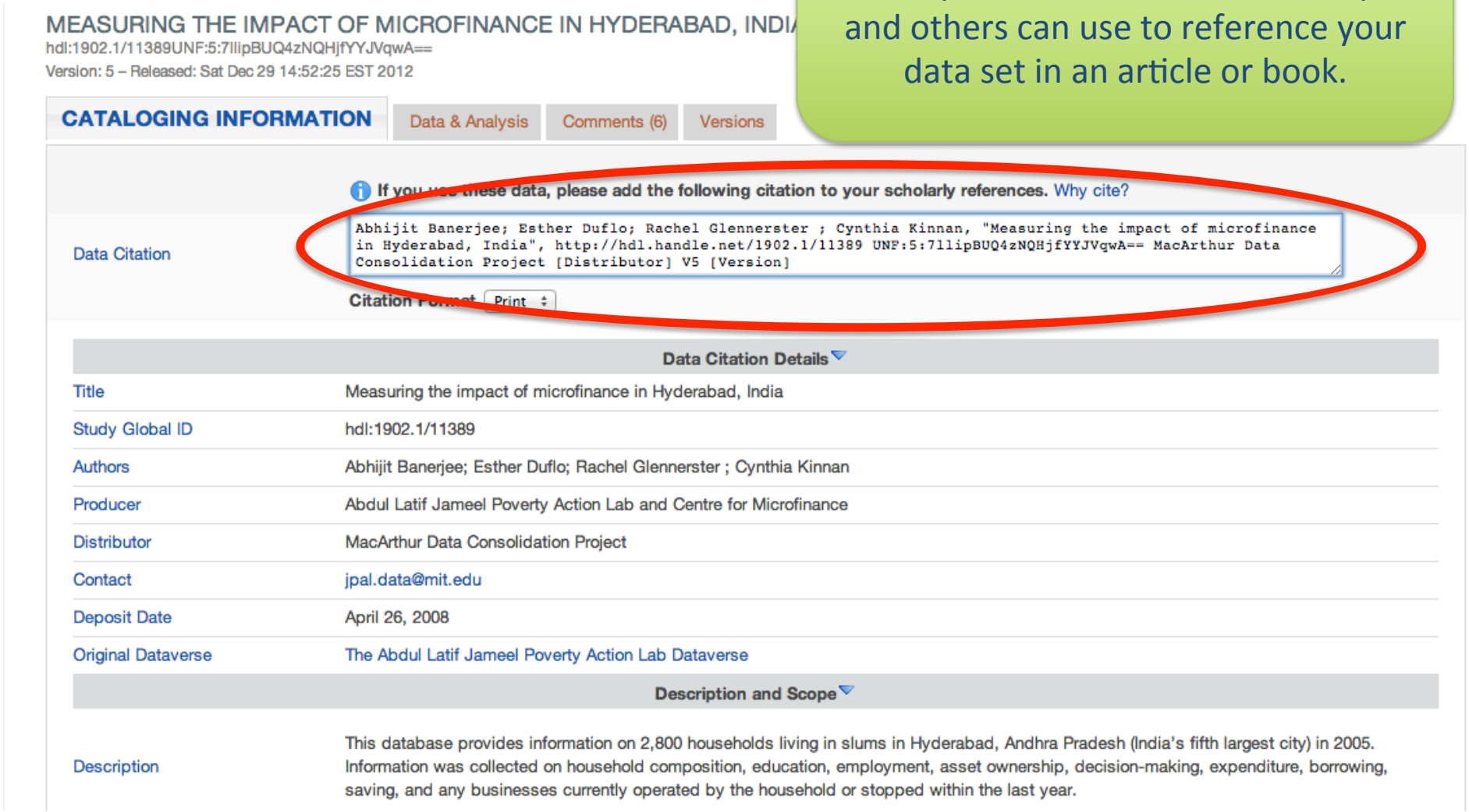
King, Gary. 1995. Replication, Replication. PS: Political Science and Politics 28: 443–499.

Metadata rich data files

Consider using the following files for tabular data sets:

- R Data: R is open-source, with a growing community
- SPSS, STATA: Also commonly used in social sciences
- Add full variable metadata
- Indicate properly missing data

Dataverse generates a **data citation** with a persistent identifier, which you and others can use to reference your data set in an article or book.



MEASURING THE IMPACT OF MICROFINANCE IN HYDERABAD, INDI

hdl:1902.1/11389UNF:5:7llipBUQ4zNQHjfYYJVqwA==

Version: 5 – Released: Sat Dec 29 14:52:25 EST 2012

CATALOGING INFORMATION | Data & Analysis | Comments (6) | Versions

If you use these data, please add the following citation to your scholarly references. Why cite?

Data Citation

Abhijit Banerjee; Esther Duflo; Rachel Glennerster ; Cynthia Kinnan, "Measuring the impact of microfinance in Hyderabad, India", http://hdl.handle.net/1902.1/11389 UNF:5:7llipBUQ4zNQHjfYYJVqwA== MacArthur Data Consolidation Project [Distributor] V5 [Version]

Citation Format: Print

Data Citation Details

Title	Measuring the impact of microfinance in Hyderabad, India
Study Global ID	hdl:1902.1/11389
Authors	Abhijit Banerjee; Esther Duflo; Rachel Glennerster ; Cynthia Kinnan
Producer	Abdul Latif Jameel Poverty Action Lab and Centre for Microfinance
Distributor	MacArthur Data Consolidation Project
Contact	jpal.data@mit.edu
Deposit Date	April 26, 2008
Original Dataverse	The Abdul Latif Jameel Poverty Action Lab Dataverse

Description and Scope

Description	This database provides information on 2,800 households living in slums in Hyderabad, Andhra Pradesh (India's fifth largest city) in 2005. Information was collected on household composition, education, employment, asset ownership, decision-making, expenditure, borrowing, saving, and any businesses currently operated by the household or stopped within the last year.

Importance of data citation

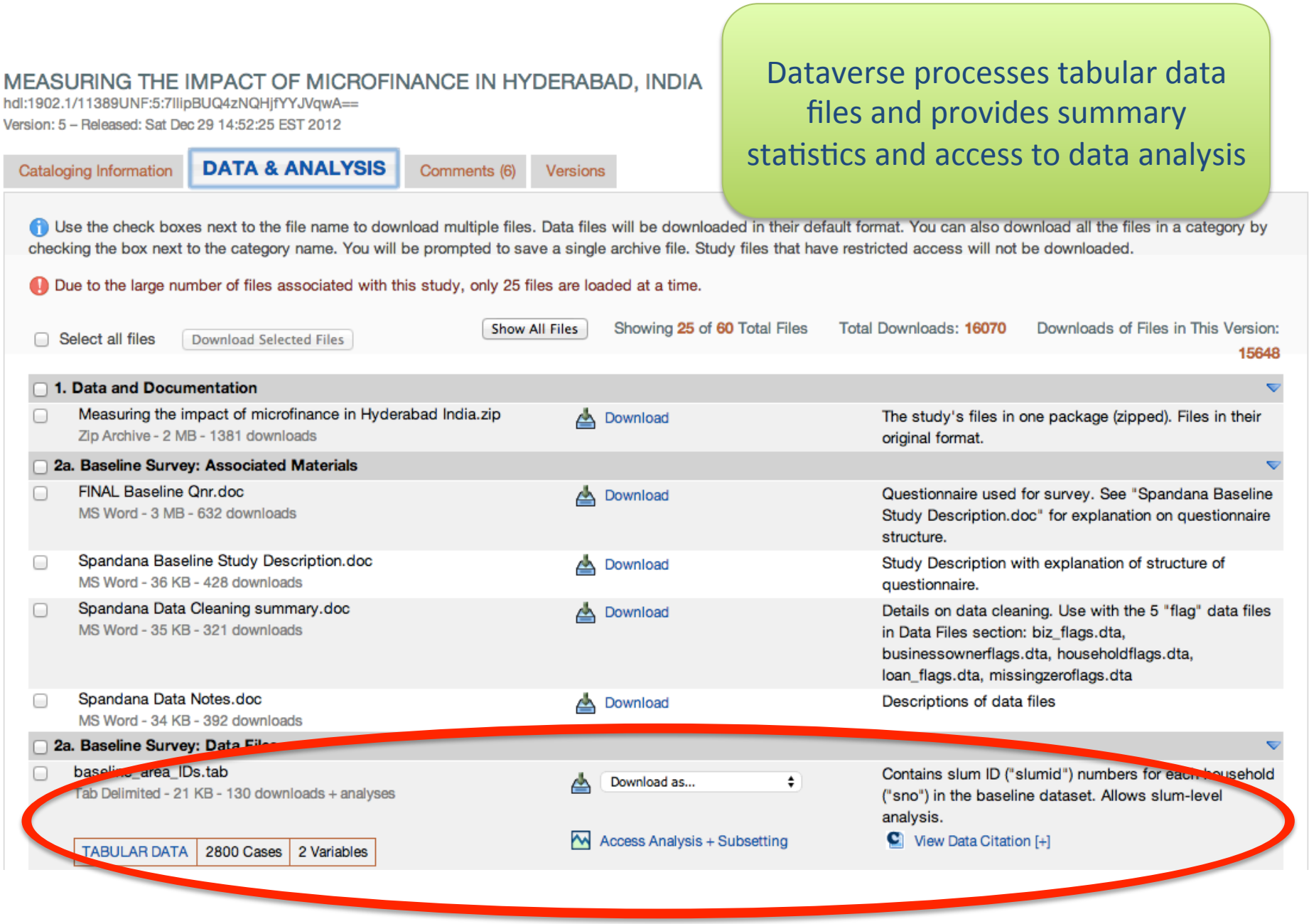
Dataverse data citation is compliant with the Joint Declaration of Data Citation Principles, which states that:

Sound, reproducible scholarship rests upon a foundation of robust, accessible data. For this to be so in practice as well as theory, data must be accorded due importance in the practice of scholarship and in the enduring scholarly record. In other words, data should be considered legitimate, citable products of research. Data citation, like the citation of other evidence and sources, is good research practice and is part of the scholarly ecosystem supporting data reuse.

DC1
Data Citation Principles

To learn more and endorse the principles:
https://www.force11.org/datacitation

Dataverse processes tabular data files and provides summary statistics and access to data analysis



Data Analysis with Zelig

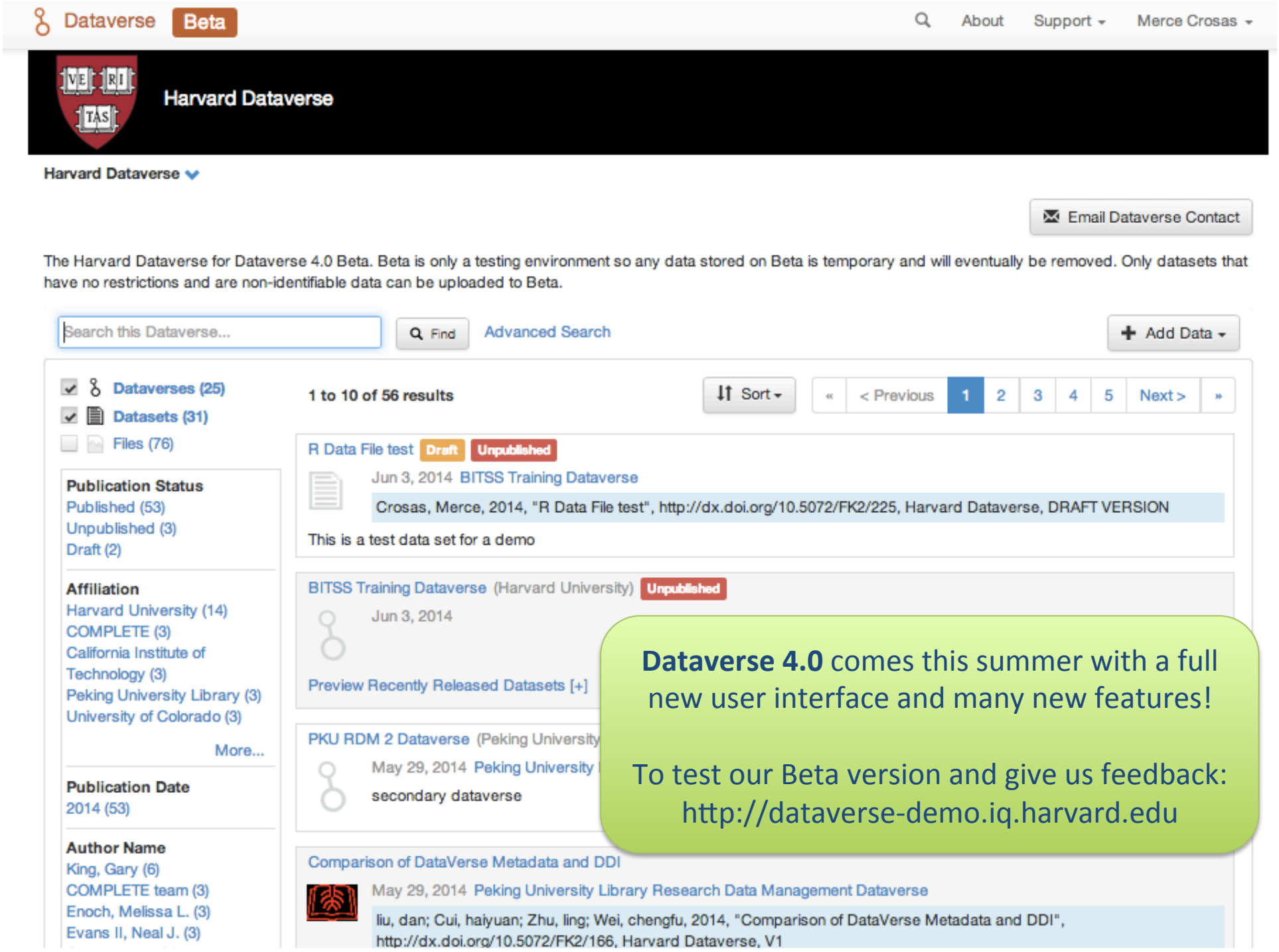
Dataverse integrates with **Zelig**:

- Zelig is an R package that provides a common interface to a large set of statistical models
- It is also developed at Harvard's IQSS, by our data science team plus contributors
- An enhanced version (Zelig 5) will be available this summer
- More information at: http://datascience.iq.harvard.edu/zelig

Additional Dataverse Features

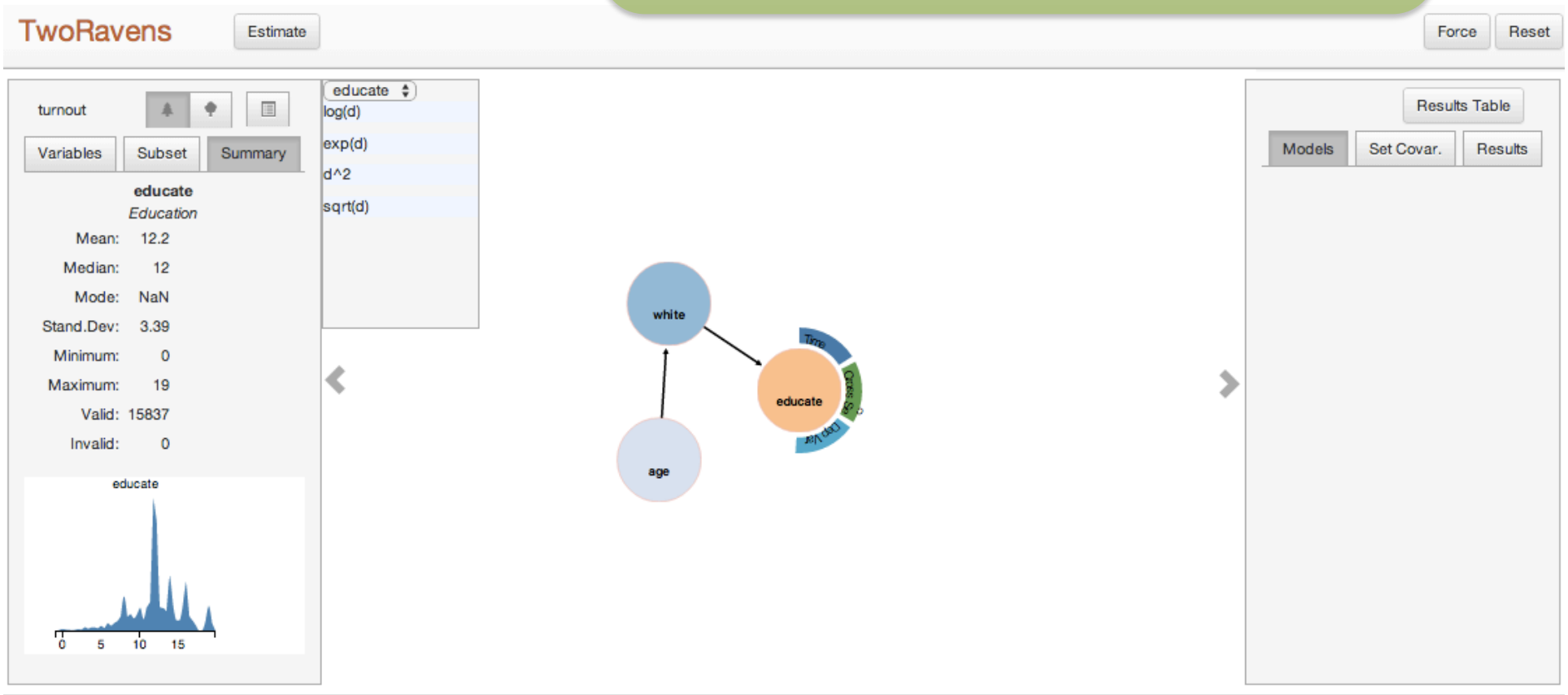
Dataverse also allows you to:

- 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

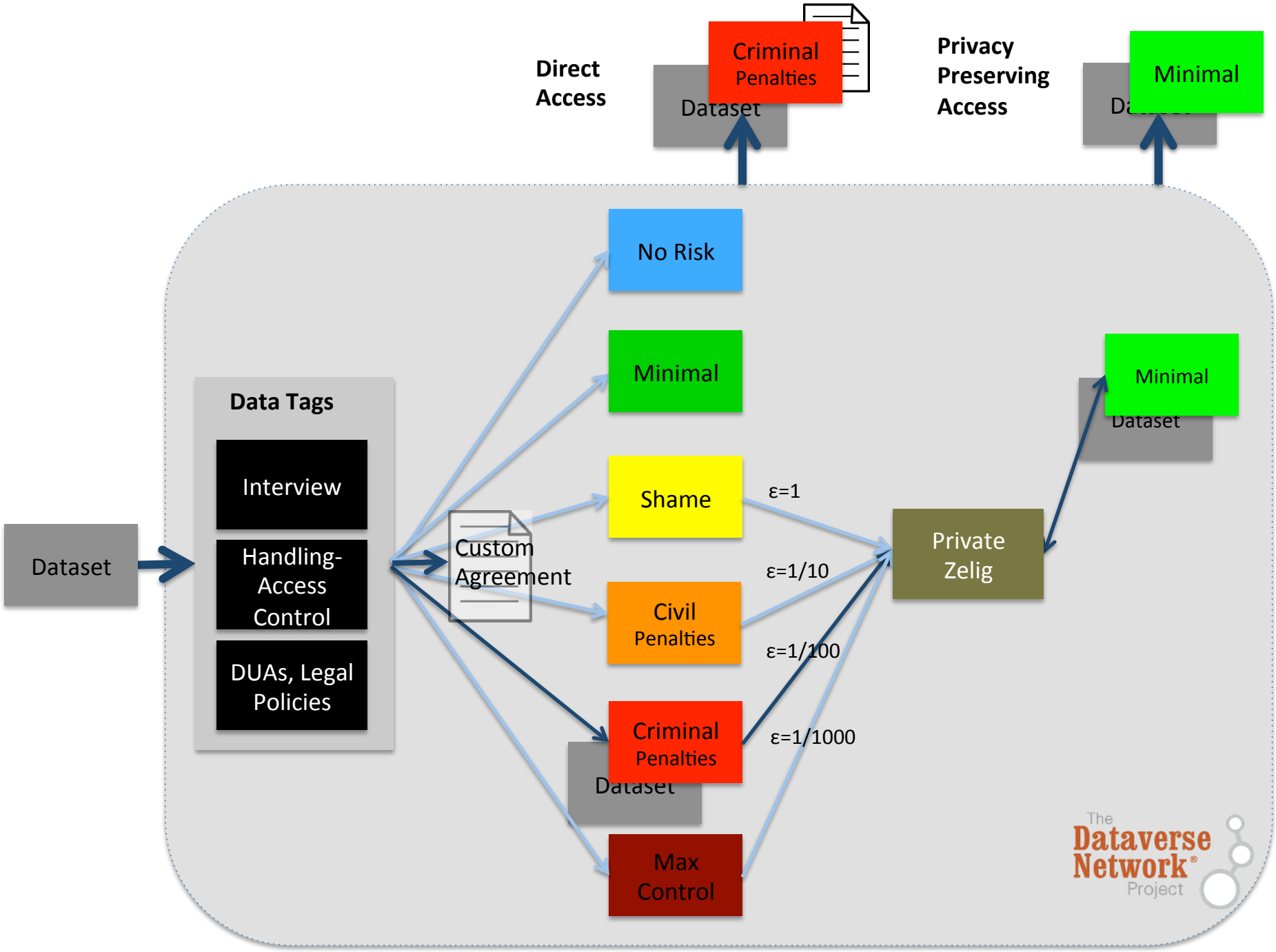

Dataverse Beta
About
Support
Merce Crosas
Harvard Dataverse
Harvard Dataverse
Email Dataverse Contact
The Harvard Dataverse for Dataverse 4.0 Beta. Beta is only a testing environment so any data stored on Beta is temporary and will eventually be removed. Only datasets that have no restrictions and are non-identifiable data can be uploaded to Beta.
Search this Dataverse...
Find
Advanced Search
Add Data
Dataverses (25)
Datasets (31)
Files (76)
1 to 10 of 56 results
Sort
« < Previous 1 2 3 4 5 Next > »
R Data File test Draft Unpublished
Jun 3, 2014 BITSS Training Dataverse
Crosas, Merce, 2014, "R Data File test", http://dx.doi.org/10.5072/FK2/225, Harvard Dataverse, DRAFT VERSION
This is a test data set for a demo
BITSS Training Dataverse (Harvard University) Unpublished
Jun 3, 2014
Preview Recently Released Datasets [+]
PKU RDM 2 Dataverse (Peking University
May 29, 2014 Peking University
secondary dataverse
Comparison of DataVerse Metadata and DDI
May 29, 2014 Peking University Library Research Data Management Dataverse
liu, dan; Cui, haiyuan; Zhu, ling; Wei, chengfu, 2014, "Comparison of DataVerse Metadata and DDI", http://dx.doi.org/10.5072/FK2/166, Harvard Dataverse, V1
Publication Status
Published (53)
Unpublished (3)
Draft (2)
Affiliation
Harvard University (14)
COMPLETE (3)
California Institute of Technology (3)
Peking University Library (3)
University of Colorado (3)
More...
Publication Date
2014 (53)
Author Name
King, Gary (6)
COMPLETE team (3)
Enoch, Melissa L. (3)
Evans II, Neal J. (3)
Dataverse 4.0 comes this summer with a full new user interface and many new features!
To test our Beta version and give us feedback: http://dataverse-demo.iq.harvard.edu

TwoRavens

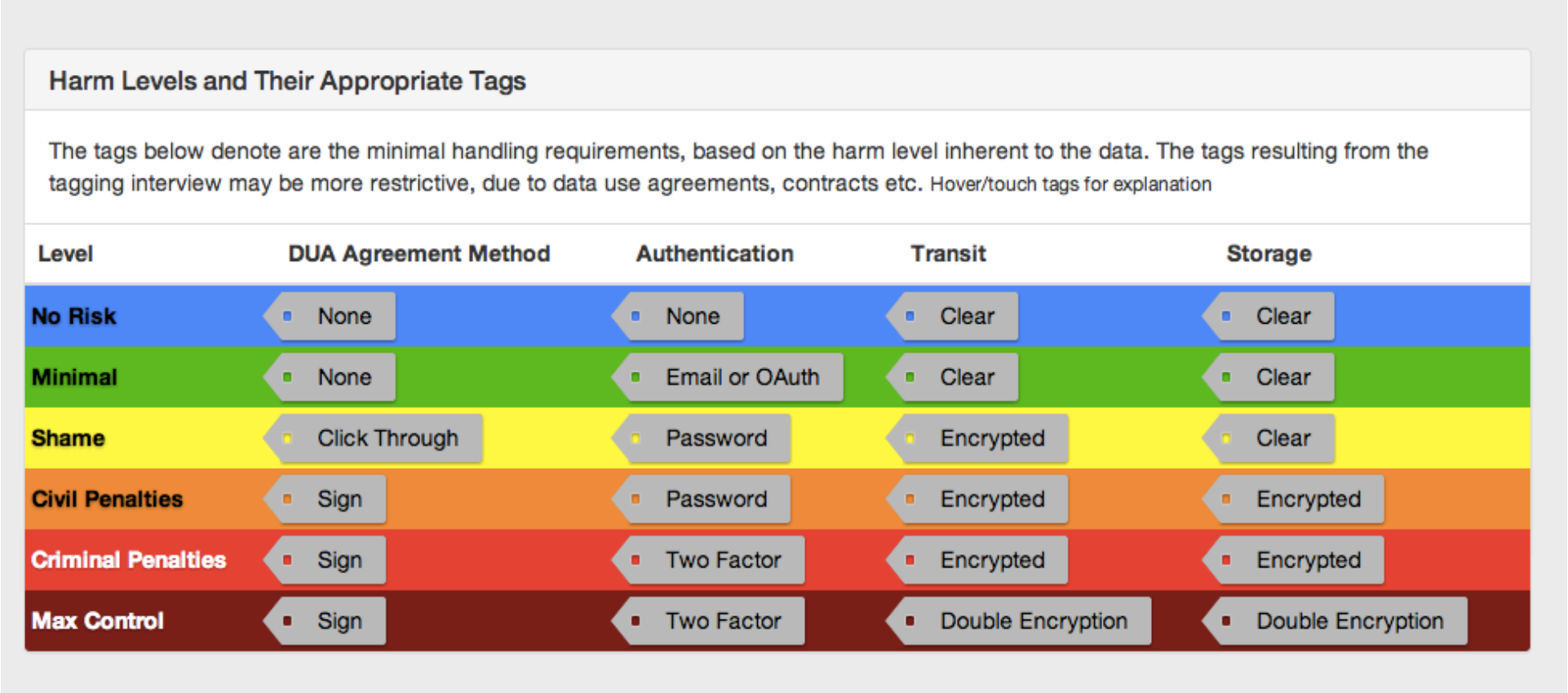
Dataverse 4.0 will include a new interactive data exploration and analysis tool, **TwoRavens,** which integrates with **Zelig statistical framework**



DataTags will provide data owners with simple data handling prescriptions that comply with the numerous regulations that apply to datasets, as well as with data use agreements that may apply to them.


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
Dataset
Max Control
ε=1
ε=1/10
ε=1/100
ε=1/1000
Private Zelig
Minimal
Dataset
The Dataverse Network® Project

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



Harm Levels and Their Appropriate Tags

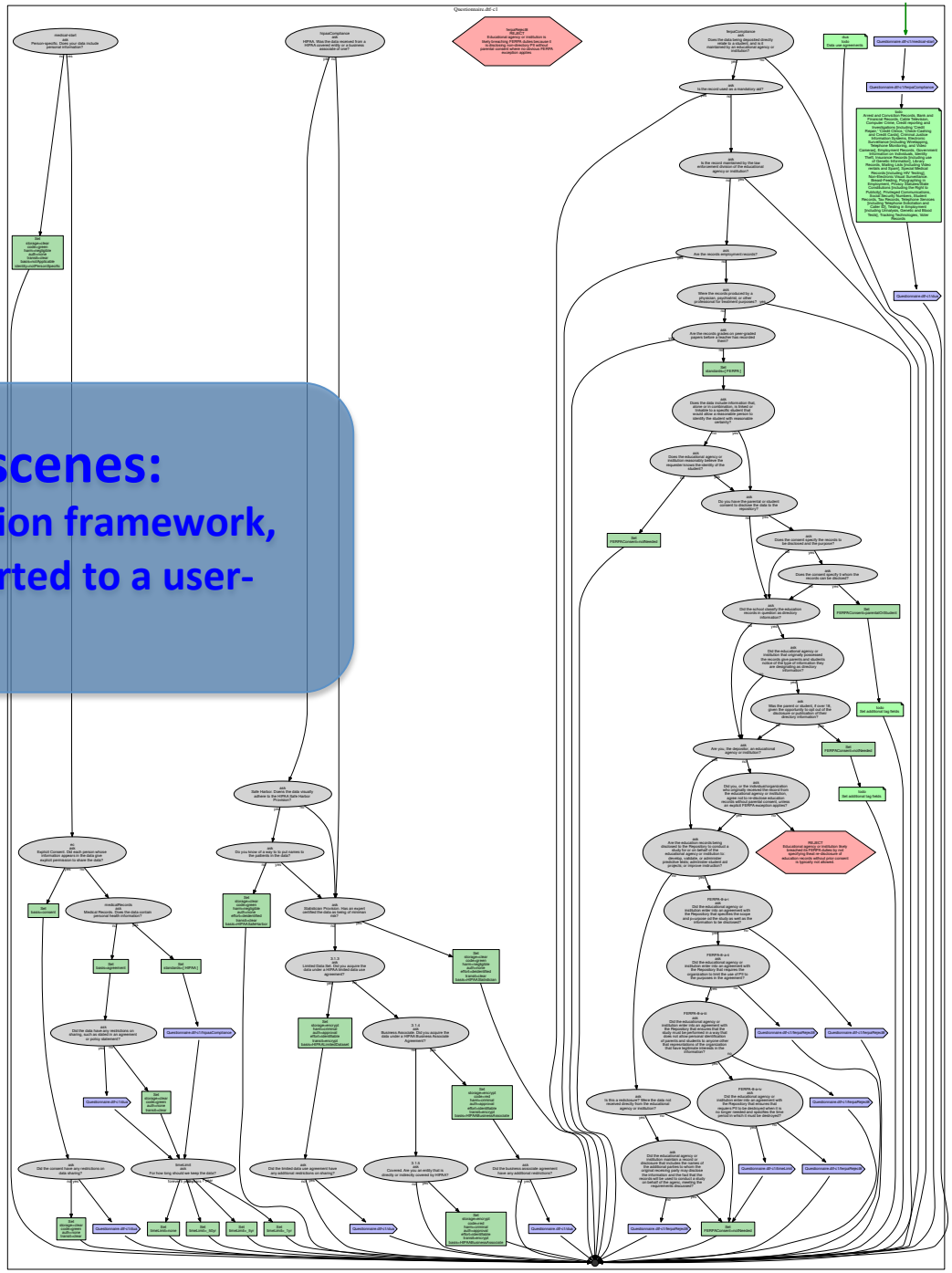
The tags below denote are the minimal handling requirements, based on the harm level inherent to the data. The tags resulting from the tagging interview may be more restrictive, due to data use agreements, contracts etc. Hover/touch tags for explanation

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

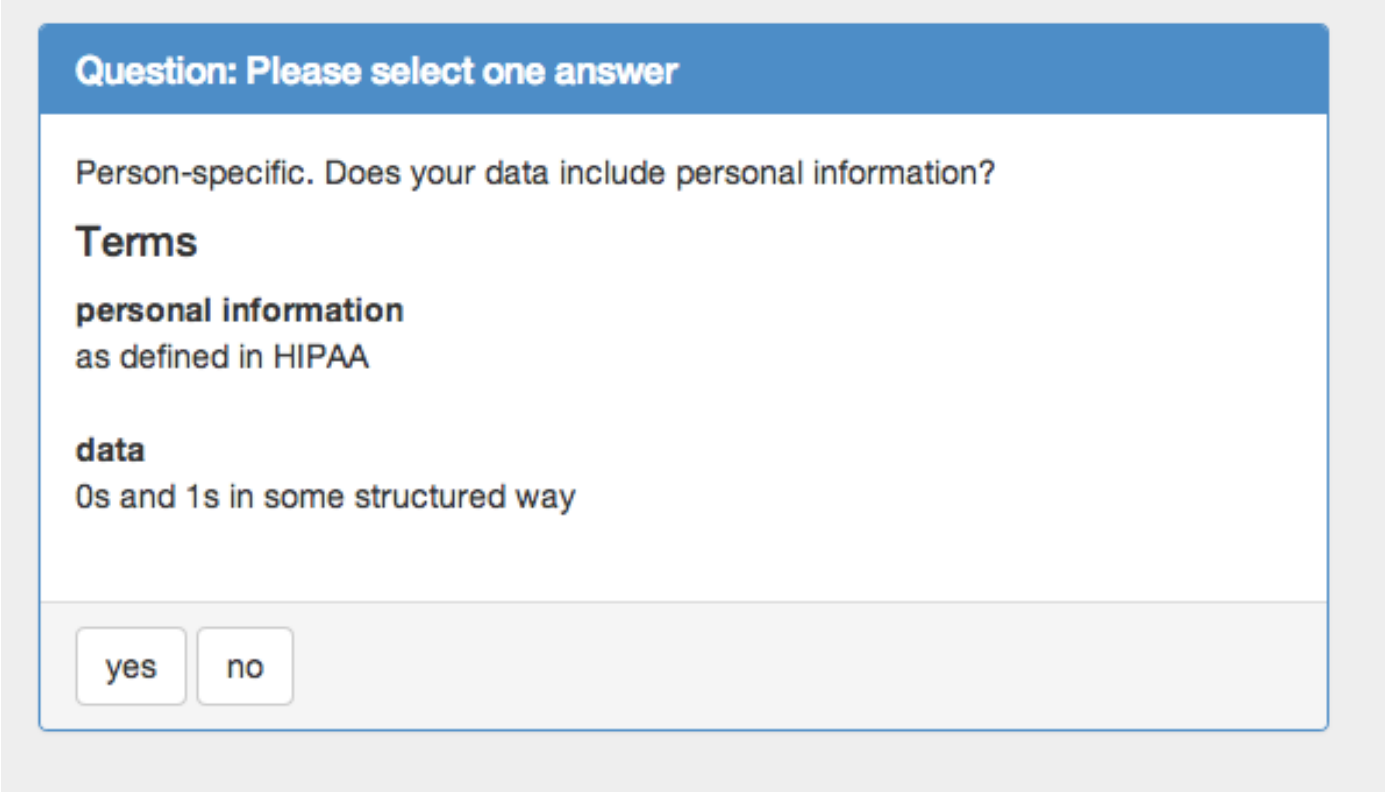
Currently supporting HIPAA and FERPA (and DUAs)

DataTags behind the scenes:

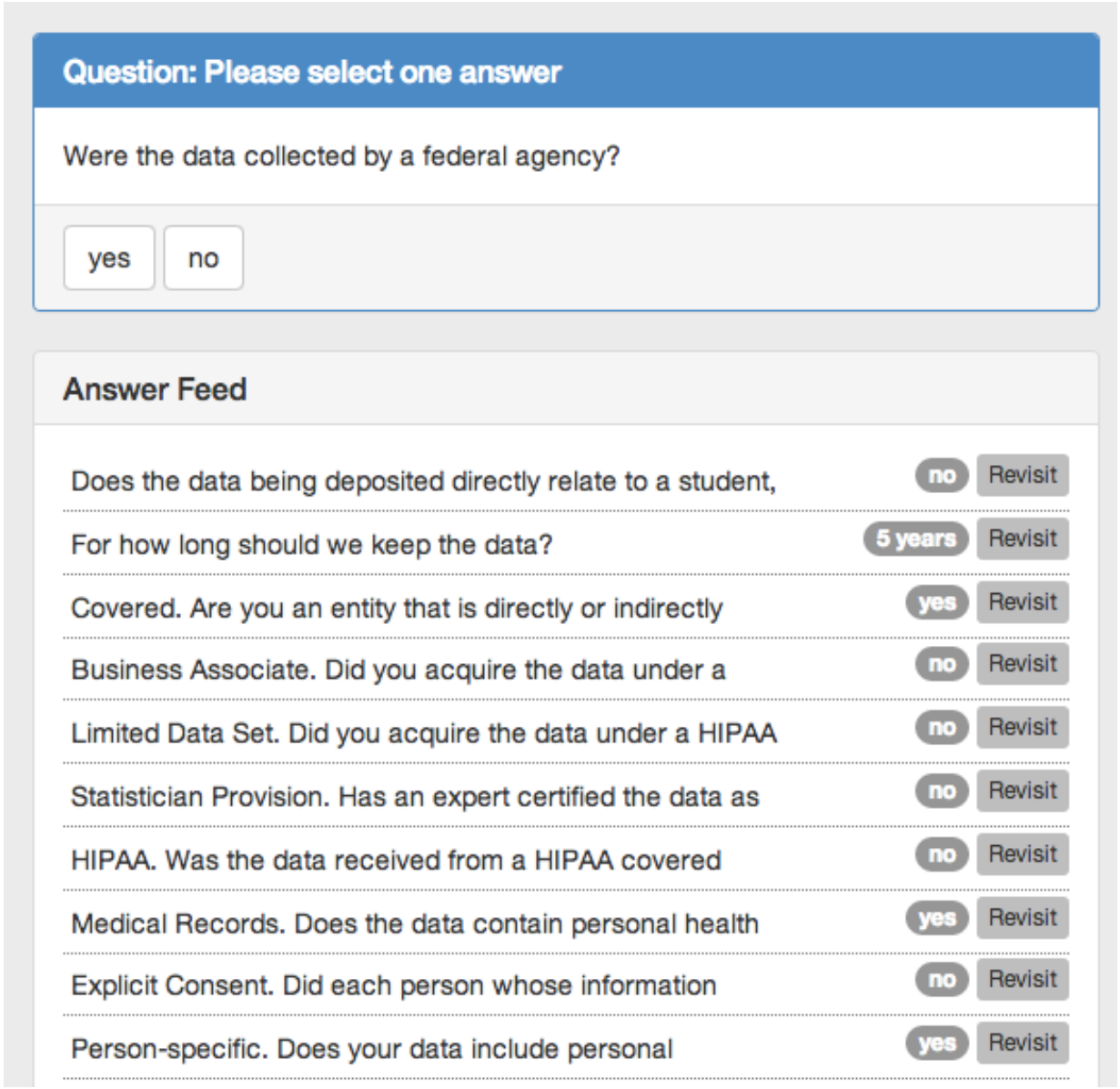
A complex interview generation framework, which is automatically converted to a user-friendly interface



Interview Example: First question …



Interview Example: After several questions …



Interview Example: … and a Final Tag

Your dataset is tagged as

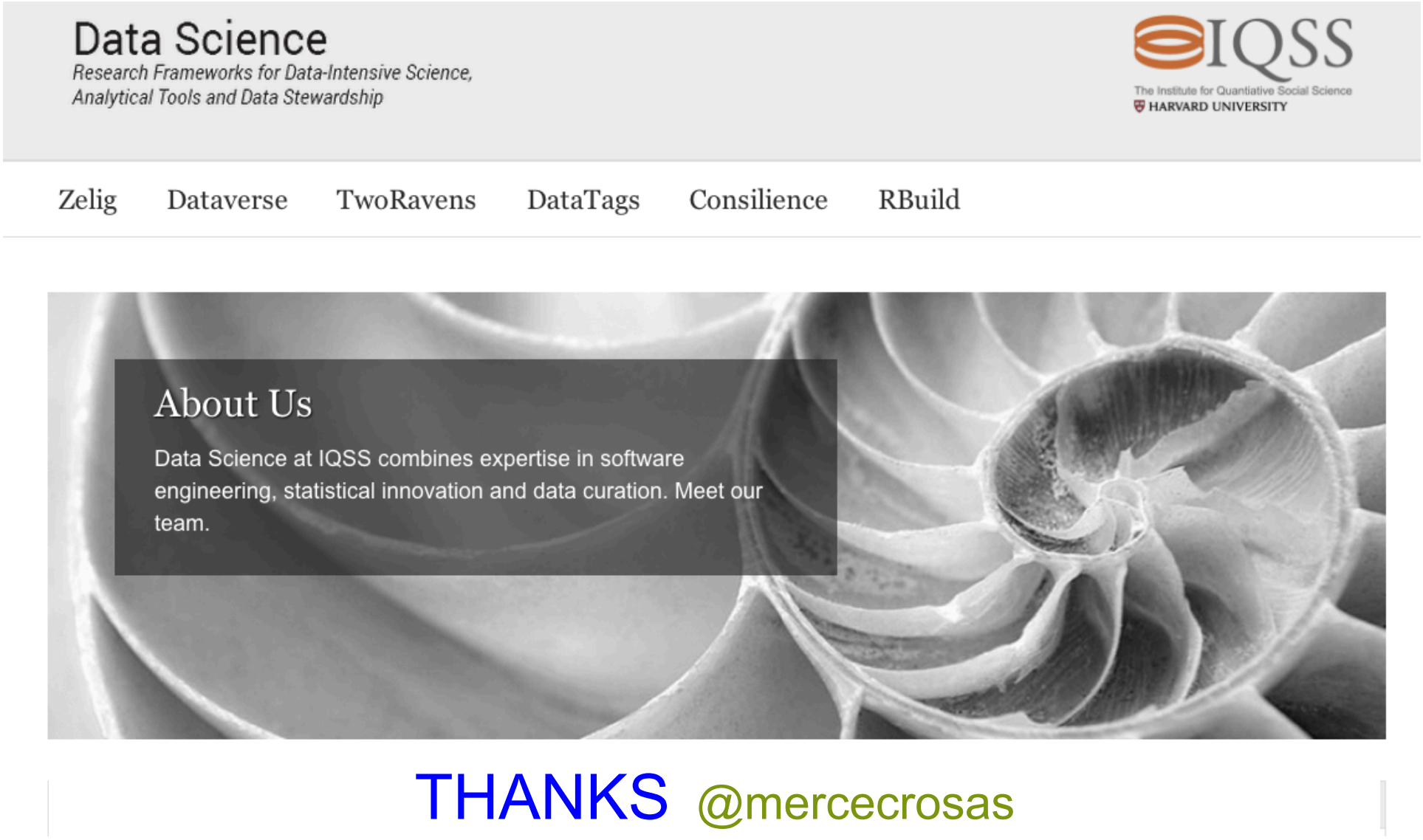
red

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

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



THANKS @mercecrosas