Data Publishing with Dataverse

Mercè Crosas, Ph.D.

Twitter: @mercecrosas

Director of Data Science

Institute for Quantitative Social Science, Harvard University

ACRL Webinar, May 22, 2014

Introduction to our Team and Projects

Collaborations Resources Blog Team

Search

Data Science

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

IQSS
The Institute for Quantitative Social Science
HARVARD UNIVERSITY

Zelig Dataverse TwoRavens DataTags Consilience RBuild

http://datascience.iq.harvard.edu

CURRENT EFFORTS

Reproducible and Reusable Science
Connecting research results to the underlying data and analysis is central to the validation and extensibility of scientific discoveries. Our tools encourage open data and methodological transparency, when possible, and promote and enable data citation.

Computationally Assisted Exploration
We build analytical tools, such as Consilience and TwoRavens, that assist a

SOFTWARE PROJECTS

Zelig

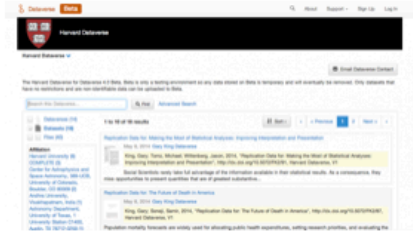
Zelig is an interface, that allows a large body of different statistical models in the R statistical language to be implemented and interpreted in a common framework and interface.

The Dataverse Network Project

DATA SCIENCE BLOG

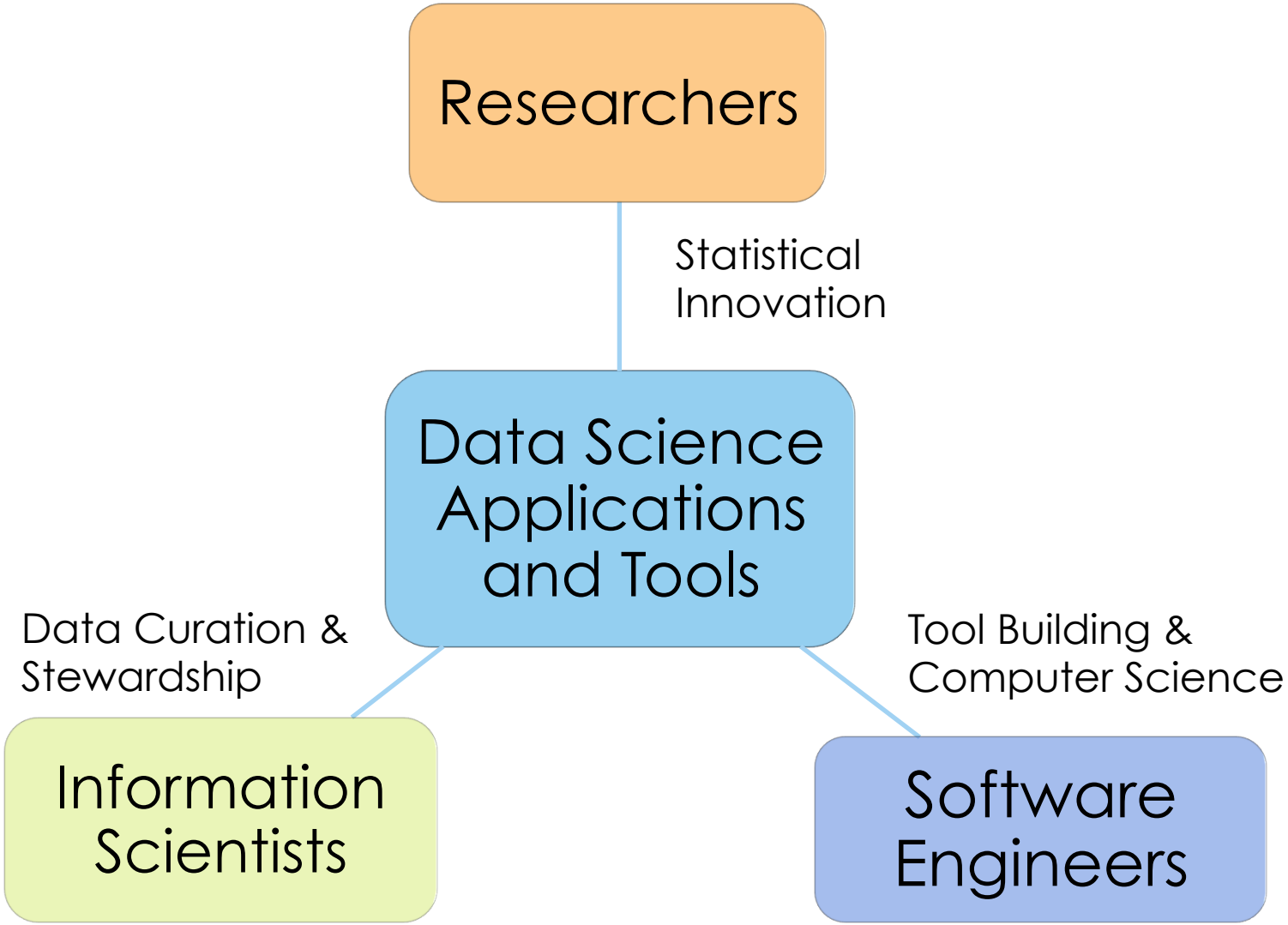
Dataverse 4.0 Beta

May 8, 2014



The Dataverse team has been hard at work on an extensive rewrite of the Dataverse application. Thanks to helpful feedback

Combines Expertise



With a Team of 20

Gary King,
Director of IQSS

Mercè Crosas,
Director of Data Science

Cris Rothfuss,
Excutive Director

Statistics and Analytics

James Honaker
Christine Choirat
Vito d'Orazio

QA and Tech support

Kevin Condon
Elda Sotiri

Software Development

Gustavo Durand
Robert Treacy
Ellen Kraffmiller
Michael Bar-Sinai
Leonid Andreev
Phil Durbin
Steve Kraffmiller
Xiangqing Yang
Raman Prasad (BARI)

Data Curation and Archiving

Sonia Barbosa
Eleni Castro
Dwayne Liburd

Usability and UI

Elizabeth Quigley
Michael Heppler

Two widely-Used Frameworks Developed in the last Decade

Zelig

A framework that allows analysts to use and interpret a large body of R statistical models from heterogeneous contributors through a common interface.

The Dataverse Network® Project

A data publishing framework that allows researchers to share, preserve, cite and analyze data, while keeping control and gaining credit for their data.

New Tools that Integrate with our Initial Work

An interactive web interface that allows users at all levels of statistical expertise to explore their data and appropriately construct statistical models.

Integrates with Zelig and Dataverse.

A framework that allows data contributors to set a level of sensitivity for their dataset based on legal regulations, which defines how the data can be stored and shared.

Integrates with Dataverse.
In collaboration with NSF Privacy Tools project

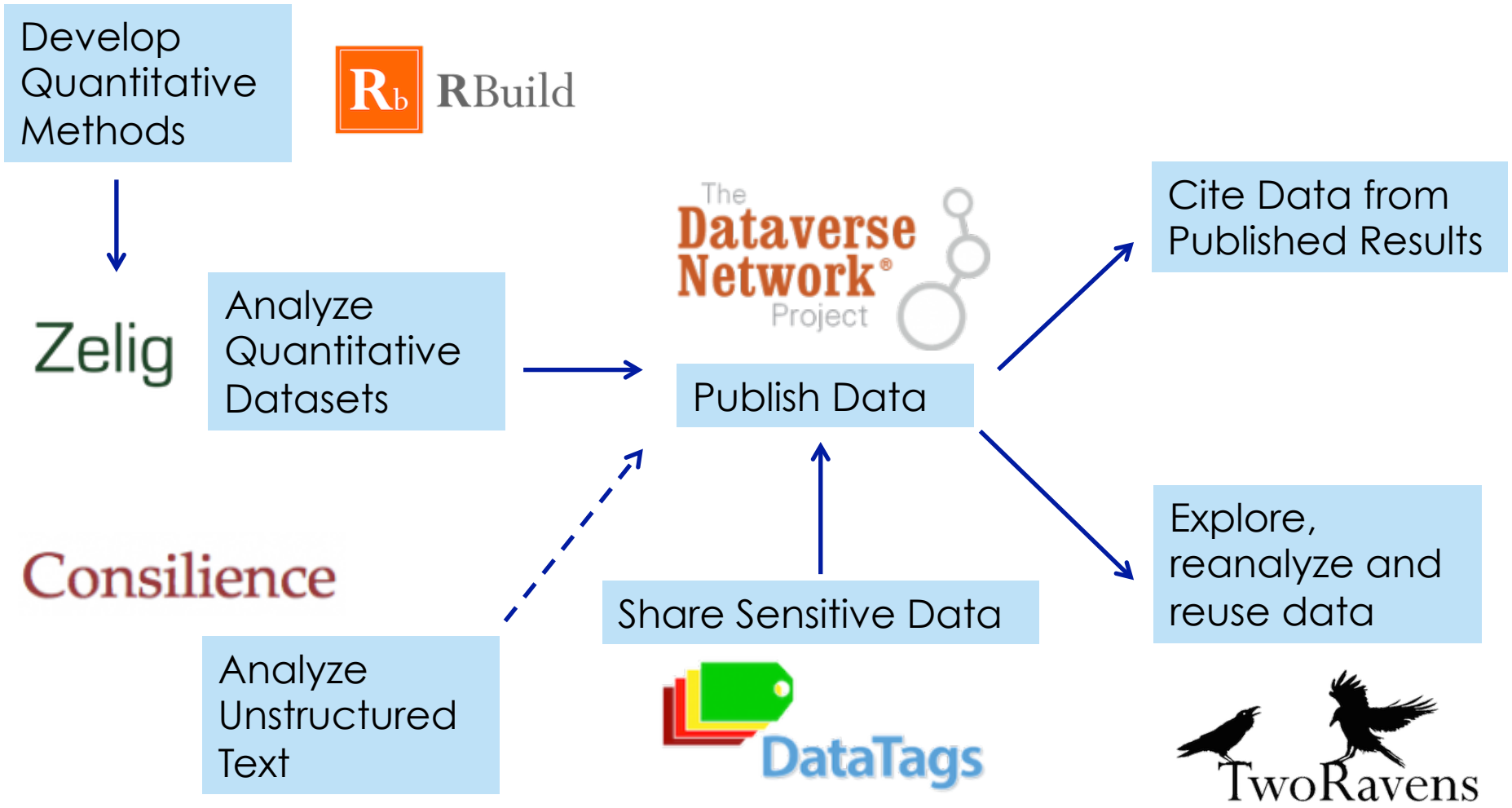
Expanding in other Areas

Consilience

A web application that assists researchers to discover new clusters to categorize large document sets, leveraging all the clustering methods in the literature.

An application that provides a continuous integration build solution for R packages shared in Git to archived published code in CRAN.

Support Throughout the Research Cycle



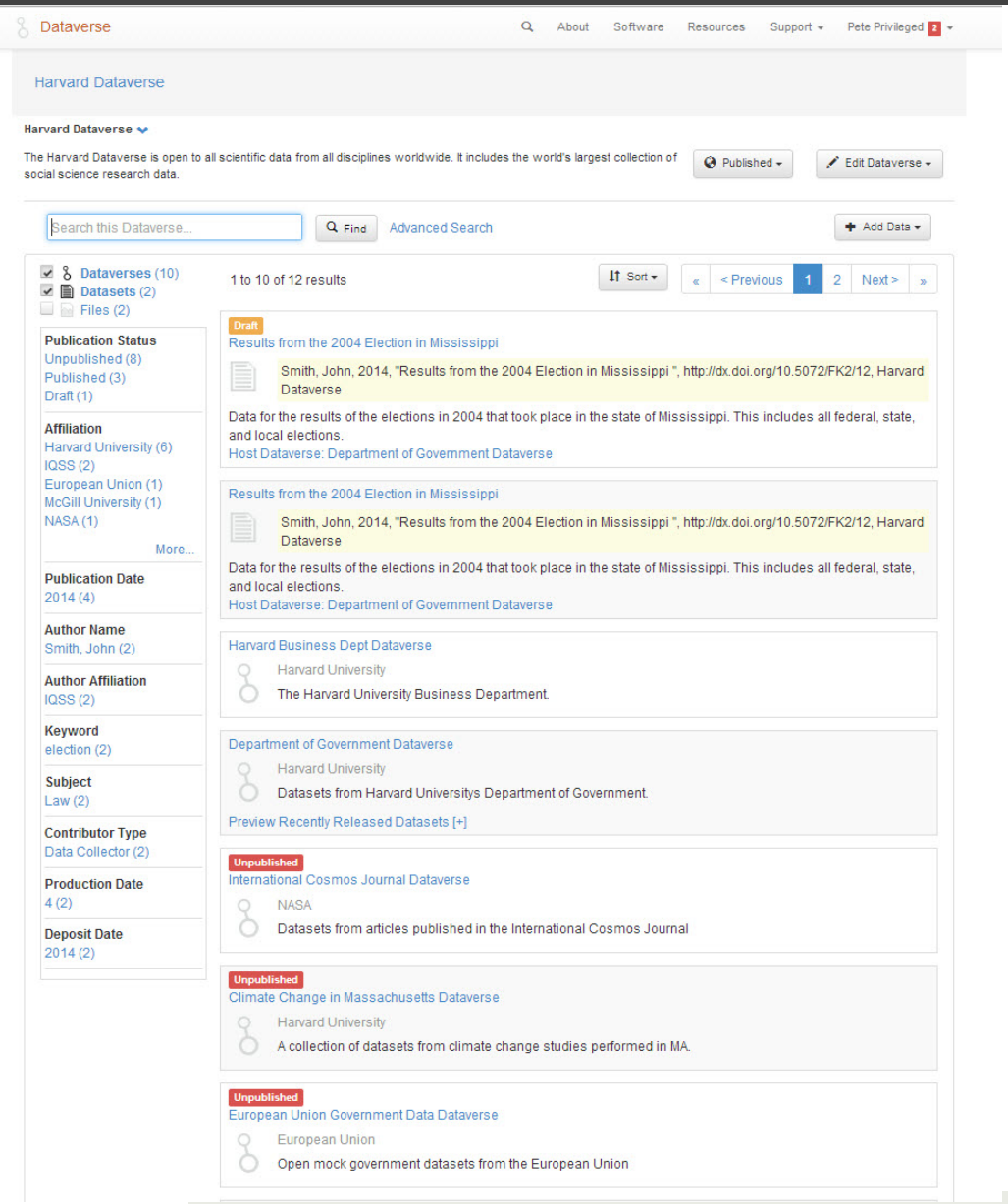
Develop > Analyze > Share > Explore > Validate & Reuse

Harvard Dataverse

The Harvard Dataverse Repository

- In collaboration with the Harvard Library, Harvard hosts a Dataverse instance free and open to all researchers across all disciplines.
- It currently holds > 53,000 datasets, with 735,000 files.
- Find or deposit data at: http://thedata.harvard.edu

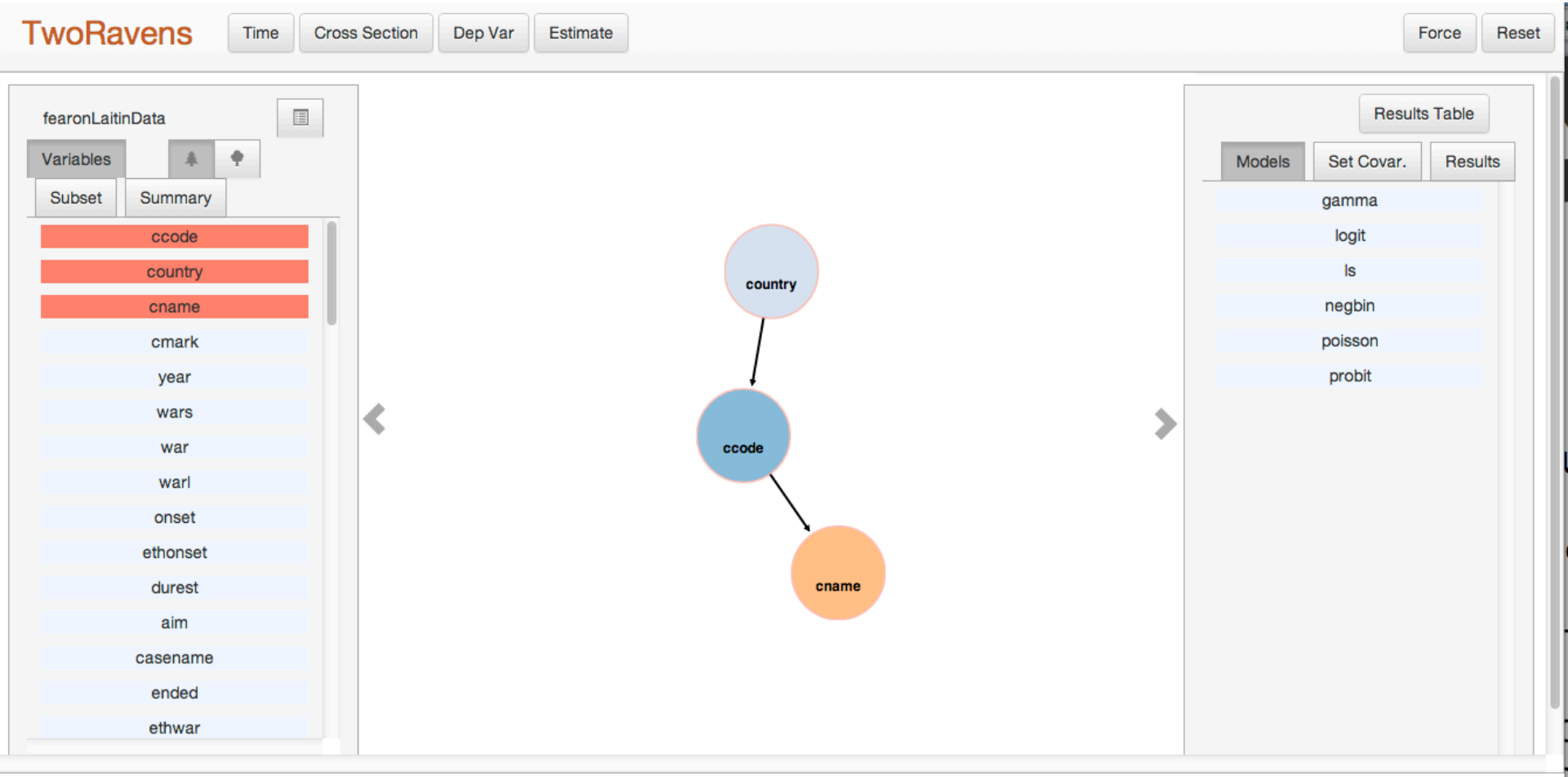
Dataverse 4.0



This summer:

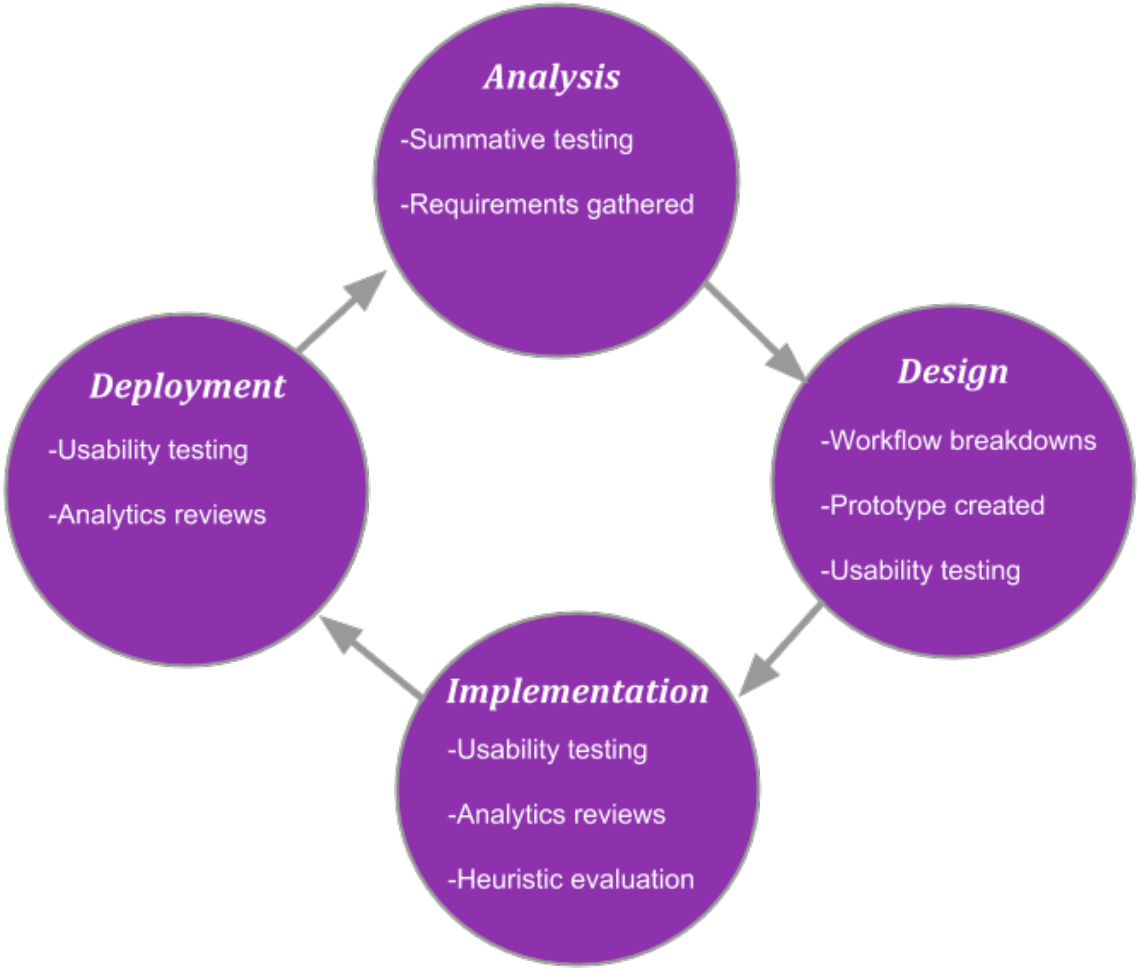
- New UI
- New rich, faceted search
- New data file ingest (excel, CSV, R, Stata, SPSS)
- New metadata for social sciences, astronomy, biomedical sciences.
- Integration with **TwoRavens**.

Integration with TwoRavens



- Users can explore, get summary statistics, and analyze tabular data
- It has access to statistical models in Zelig

User Feedback in Every Step



Dataverse 4.0 Beta available now for Testing:

http://dataverse-demo.iq.harvard.edu

Data Publishing Workflows

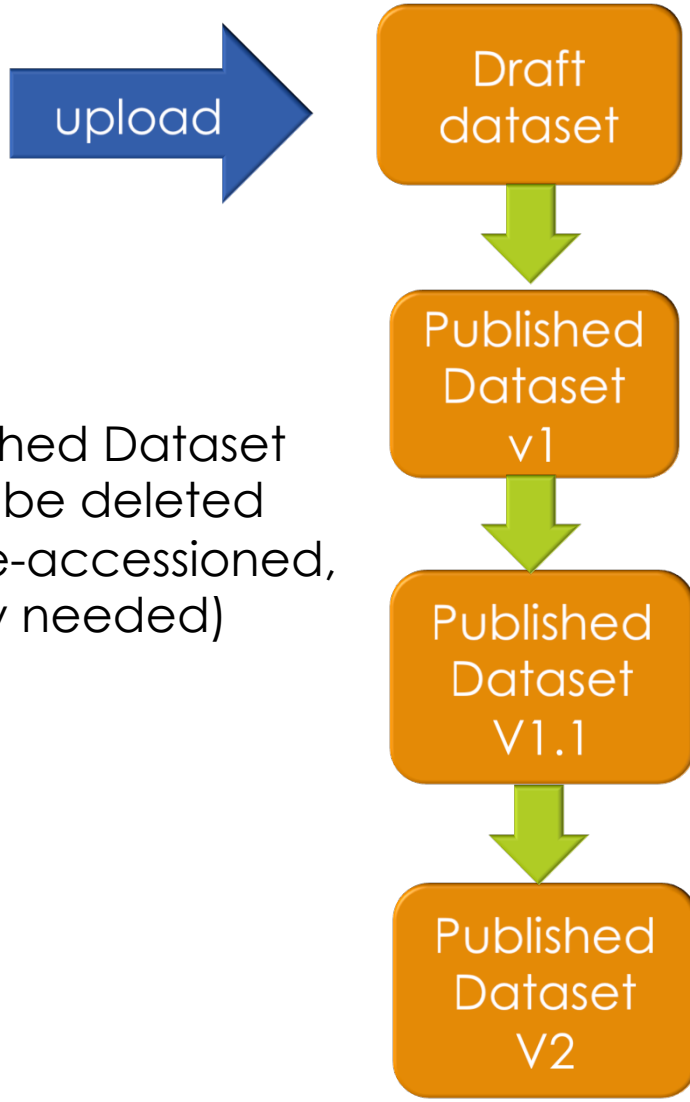
Data Publishing Guidelines

Three pillars to Data Publishing:

- A trusted data repository to guarantee long-term access
- A formal data citation*
- Sufficient information to understand and reuse the data (metadata, documentation, code)

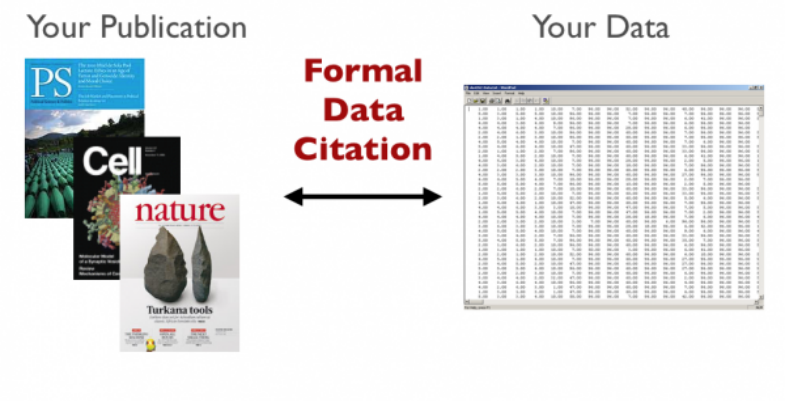
* Data Citation Principles: https://www.force11.org/datacitation

A Rigorous Data Publishing Workflow



A Published Dataset cannot be deleted (only de-accessioned, if legally needed)

Workflows that Integrate with Journals



Option A. Publish a dataset to your Dataverse, then provide the Data Citation to the journal.

Option B. Contribute to a journal Dataverse:

1. Add dataset to Journal Dataverse as a draft.
2. Journal Editor reviews it, and approves it for release.
3. Dataset is published with Data Citation and link from journal article to the data.

Option C. Seamless Integration between journal system and Dataverse.

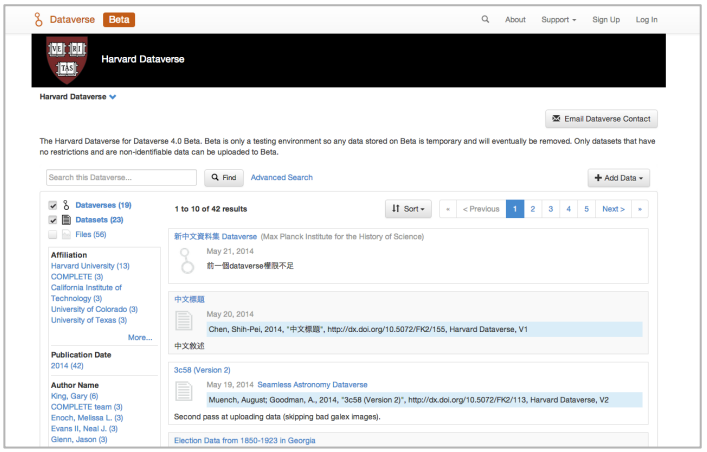
OJS and Dataverse Integration

OJS Journal

Citation to Data

Citation to Article

Harvard Dataverse

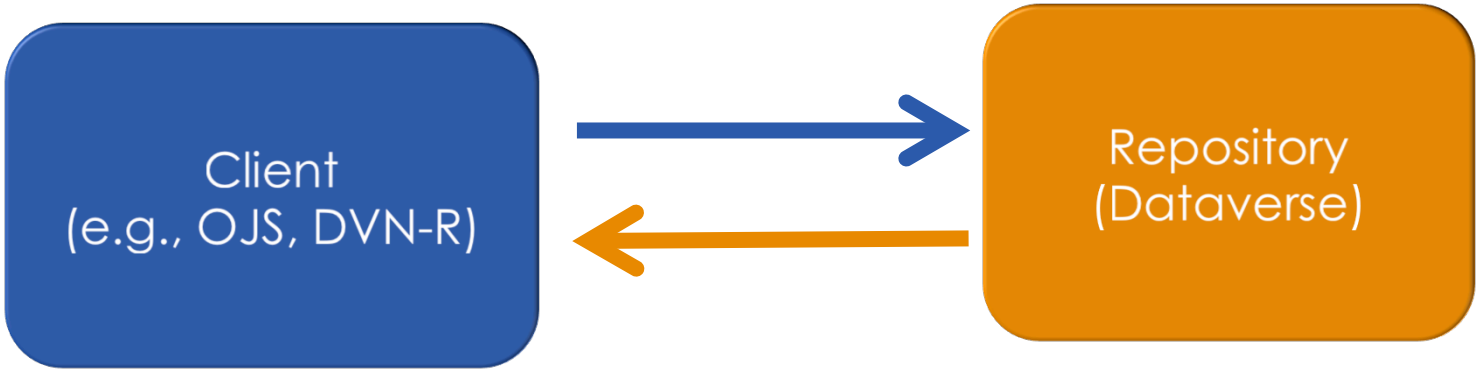


- Sloan funded project to integrate PKP's Open Journal System with the Dataverse software.
- Pilot with ~ 50 journals
- OJS Dataverse plugin now available with latest OJS release
- http://projects.iq.harvard.edu/ojs-dvn

Seamless System Integration

For Option C:

- ✓ XML file: AtomPub "entry" with Dublin Core Terms (e.g., title, creator)
- ✓ Zip file: All data files associated with that dataset.
- ✓ HTTP header "In-Progress: false" to publish datasets.



- ✓ XML file: "Deposit Receipt"

Client can query repository (server) any time to get status

Deposit API based on SWORD

- Follows SWORD2 specifications
- SWORD is supported within academic publishing; based on the web standard Atom Publishing Protocol.
- The SWORD project provides client libraries for Python, Java, Ruby, and PHP:
  - OJS uses the PHP client library
  - OSF uses the Python client library
  - DataUp and DVN-R built a custom Dataverse client

How it differs from SWORD

- Dataverse does not use SWORD download API:
  - Instead, Dataverse uses own Data API
  - Plan to support SWORD download in the future
- Added XML attribute to pass article citation from client:
  - Allow DCterms:isReferencedby to contain attributes (HoldingsURI) to link back to article from Dataverse
  - This is now part of the SWORD PHP client library

Support for Metadata Standards

- **Citation metadata:** Applies to all datasets – Supported currently by Data Deposit API
- Extensible metadata blocks for specific domains (in 4.0):
  - **Social sciences:**
    - Maps to DDI schema;
    - File metadata extracted from tabular data file
  - **Astronomy:**
    - Maps to VO schema;
    - Partially extracted from FITS file
  - **Biomedical sciences:**
    - Maps to ISA-tab schema
    - Controlled vocabularies maps to EFO, OBI, and Ontology of Clinical Research
    - Extended and managed using SKOS (support taxonomies within the framework of the semantic web

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 an analysis of current data publication practices.

Compliant with DataCite, Dublin Core, DDI study description

Keyword data publication

Subject *

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

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

Compliant with DDI for Social Sciences

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)

Compliant Virtual Observatory (VO) schema

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)

Compliant ISA-Tab schema, plus biomedical ontologies

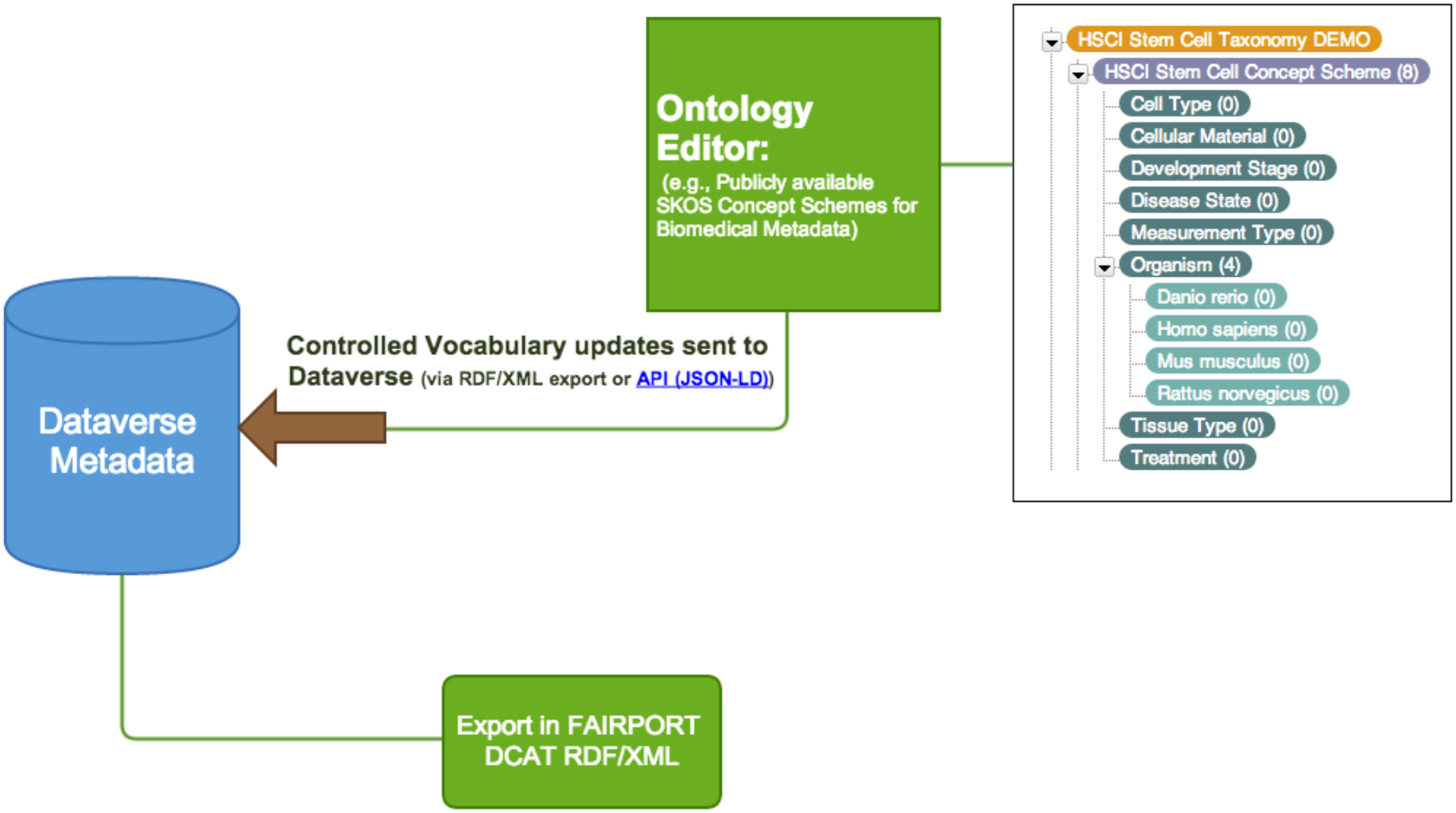
Organism

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

Cell Type

+

A workflow to access and update Controlled Vocabularies



Future Projects

Expanding to support more Data

- Sharing sensitive data with DataTags and Secure Dataverse
- Integration with other systems:
  - Open Science Framework
  - DataUp
  - WorldMap
  - DataBridge
  - ORCID
  - …
- Expand to Large-scale datasets with efficient data storages

DataTags: For Sharing Sensitive Data

Data Tags Sharing data with confidence

Start Tagging

Harm Levels, and Their Appropriate Tags

Level	D.U.A. Agreement Method	Authentication	Transit Encryption	Storage Encryption
NoRisk	None	None	Clear	Clear
Minimal	None	Email_or_OAuth	Clear	Clear
Shame	ClickThrough	Password	Encrypted	Encrypted
CivilPenalties	Sign	Password	Encrypted	Encrypted
CriminalPenalties	Sign	TwoFactor	Encrypted	Encrypted
MaxControl	Sign	TwoFactor	DoubleEncryption	DoubleEncryption

Final tags may not match the tags of a specific harm level. Hover over the terms to view an explanation.

Dataverse Network PROJECT | Data Privacy Lab | IQSS

Data Tags Sharing data with confidence

Person-specific

Does your data include personal information?

YES NO

Data Tags

DUA Agreement Method	n/a
Authentication Type	n/a
Transit Encryption Type	n/a
Storage Encryption Type	n/a

Full Interview

Tagging Complete!

Direct Data Access

CriminalPenalties

DUA Agreement Method	Sign
Authentication Type	TwoFactor
Transit Encryption Type	Encrypted
Storage Encryption Type	Encrypted

Try Dataverse 4.0 Beta and give us feedback:

http://dataverse-demo.iq.harvard.edu

Learn more about our projects at:

http://datascience.iq.harvard.edu

THANKS

mcrosas@iq.harvard.edu Twitter: mercecrosas