Data Publishing Workflows with Dataverse

Mercè Crosas, Ph.D.

Twitter: @mercecrosas

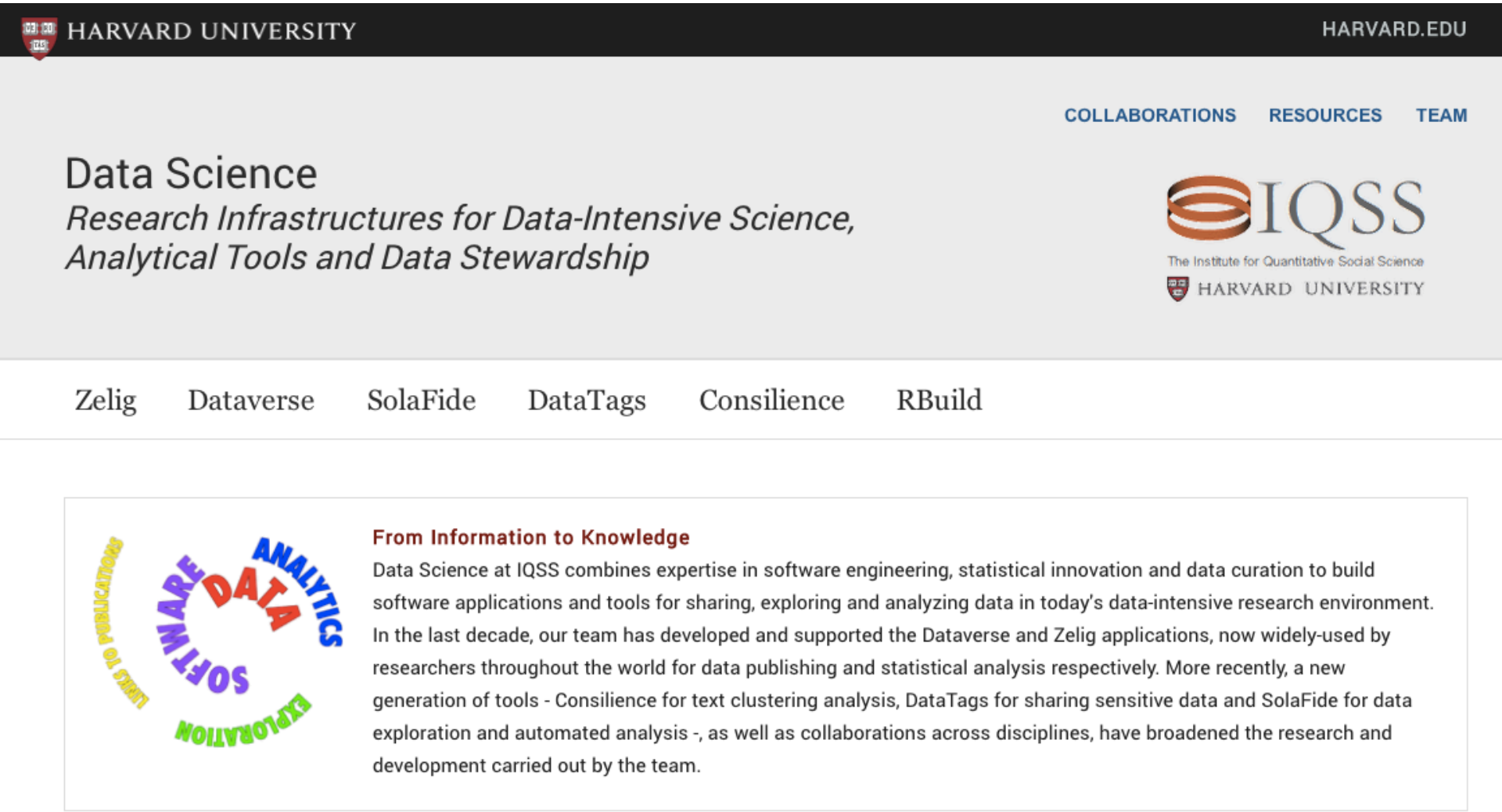
Director of Data Science

Institute for Quantitative Social Science, Harvard University

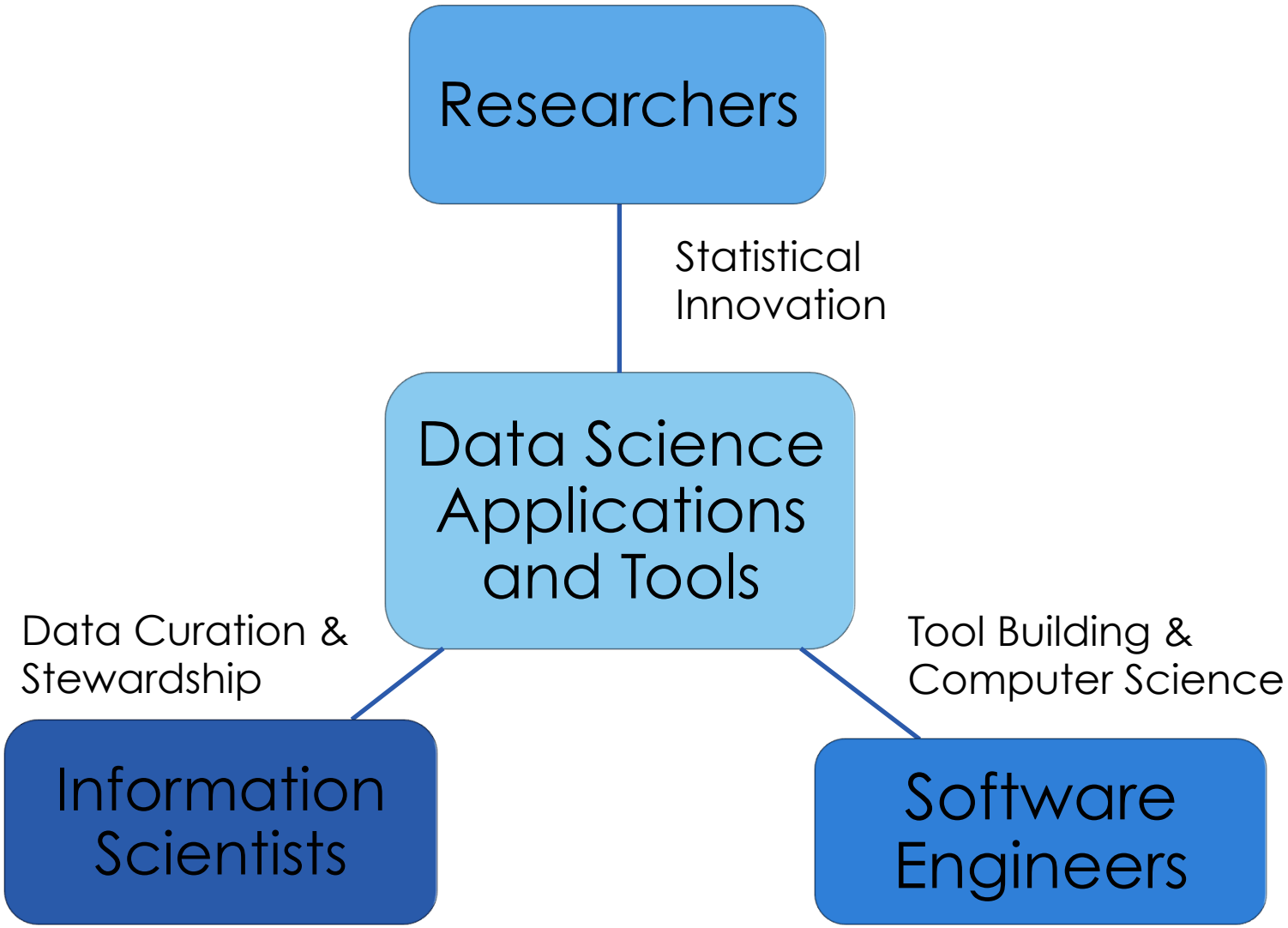
MIT, May 6, 2014

Intro to our Data Science Team and Projects

Data Science at the Institute for Quantitative Social Science
http://datascience.iq.harvard.edu



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

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

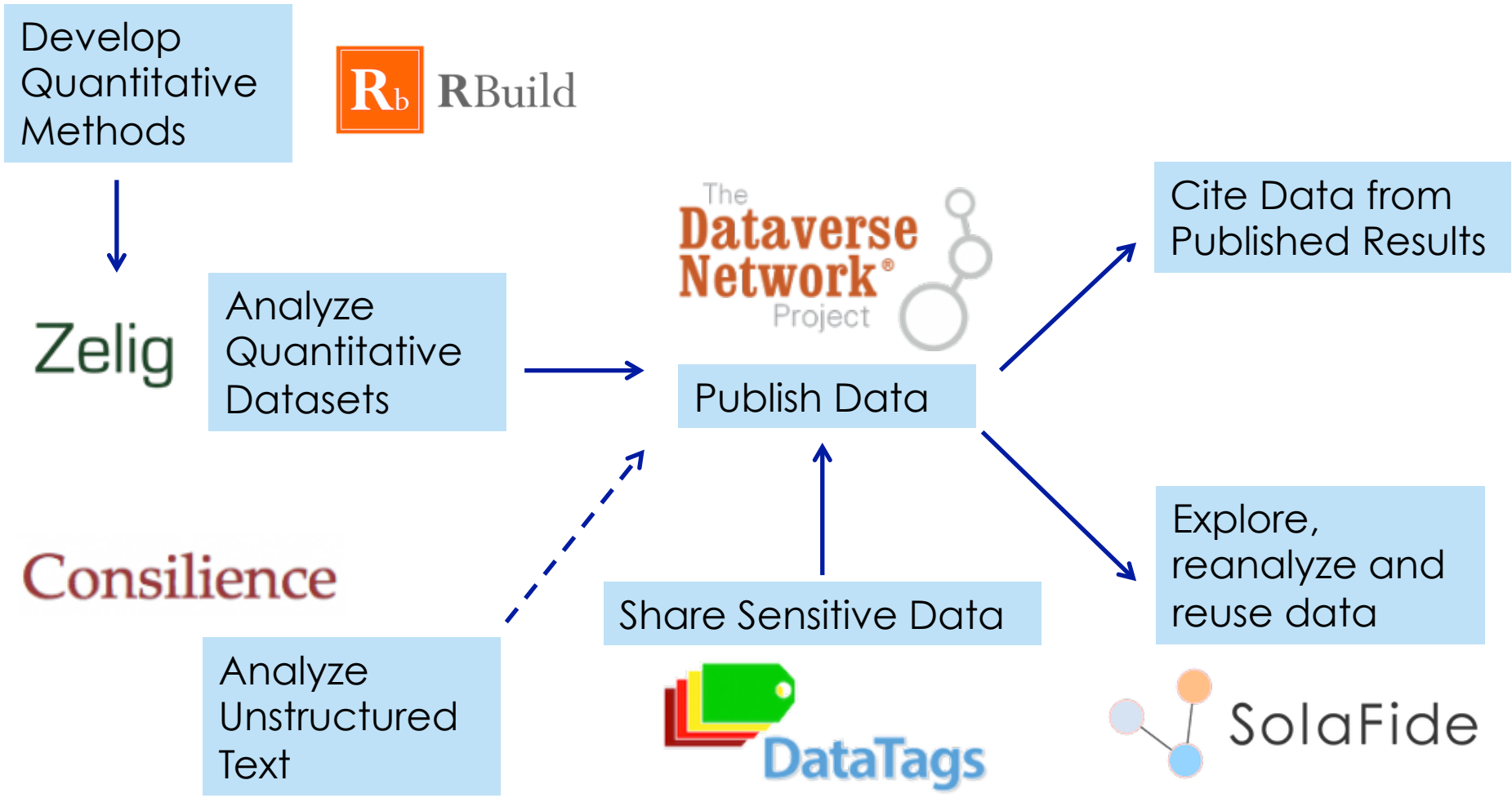
Expanding in other Areas

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

Current Research Interests and Efforts

Reproducible and Reusable Science: "encourage open data and methodological transparency, and promote and enable data citation" (with Dataverse, Zelig and SolaFide)

Computationally Assisted Exploration: "with Consilience and SolaFide, assist researchers to understand and discover new insights from their data"

Interdisciplinary Quantitative Scientific Scope: "our tools and research frameworks address broad methodological issues in quantitative science and are often employed in other domains"

When Data are Not Open: "solutions to preserve privacy, while still providing science the fundamental ability to learn, access and replicate findings, with DataTags and PrivateZelig"

Large-Scale Data Sets: "will handle large-scale data sets, as Big Data science reaches all disciplines: Consilience for millions of text documents, and Zelig and Dataverse to handle TB-PB-scale data sets."

Harvard Dataverse

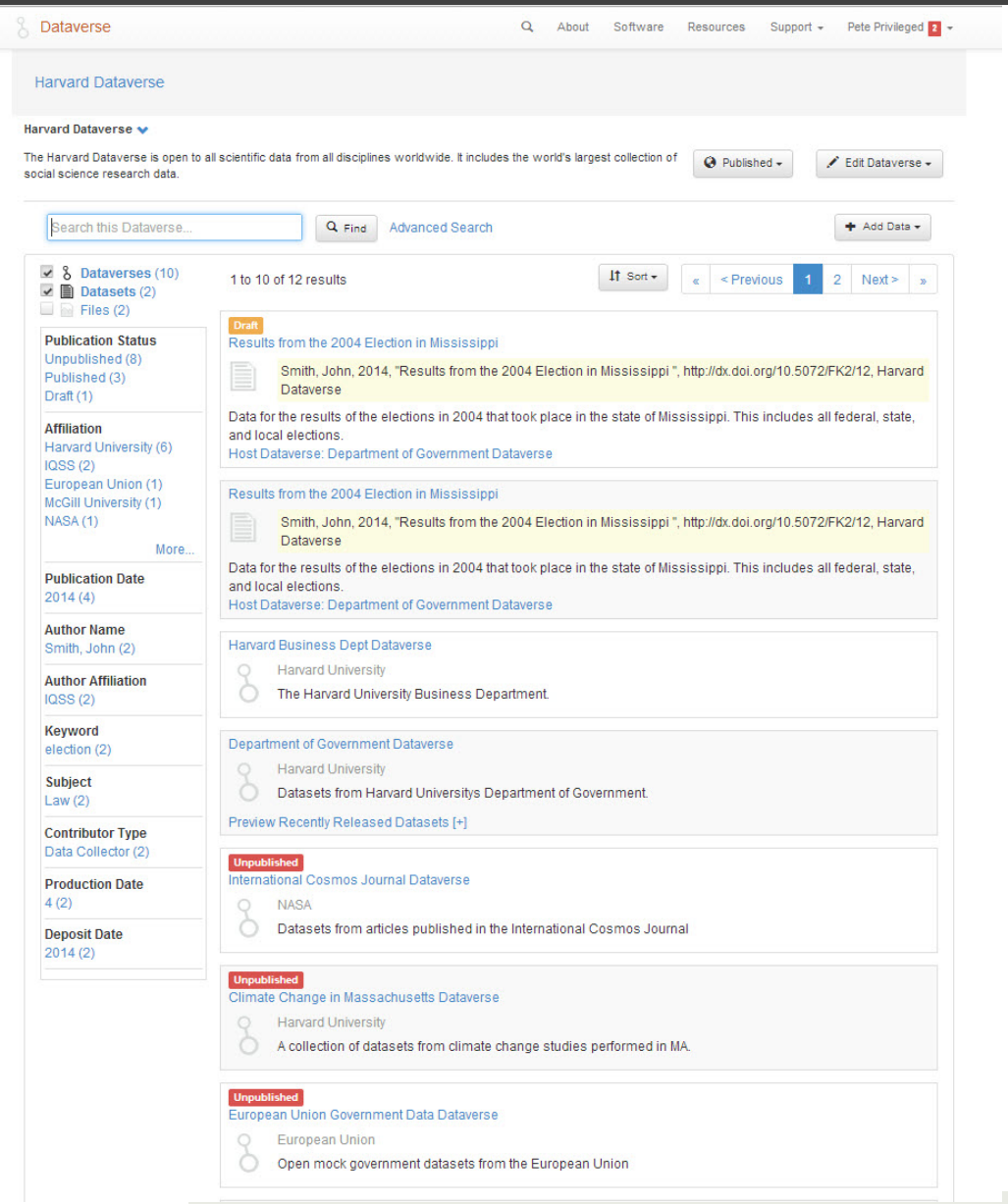
The Harvard Dataverse Repository

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

Collaborations with MIT

- Membership through the Harvard-MIT Data Center (e.g., statistics training, access to ICPSR collection)
- The MIT Libraries Dataverse disseminates data purchased by the MIT Libraries (with Kate McNeill):
  - http://thedata.harvard.edu/dvn/dv/mit
- MIT faculty and research groups are already disseminating their data through the Harvard Dataverse
- Research collaborations (with Micah Altman):
  - Integration of Publications with Data (Funded by Sloan): http://projects.iq.harvard.edu/ojs-dvn
  - Privacy Tools for Sharing Research Data (Funded by NSF): http://privacytools.seas.harvard.edu/

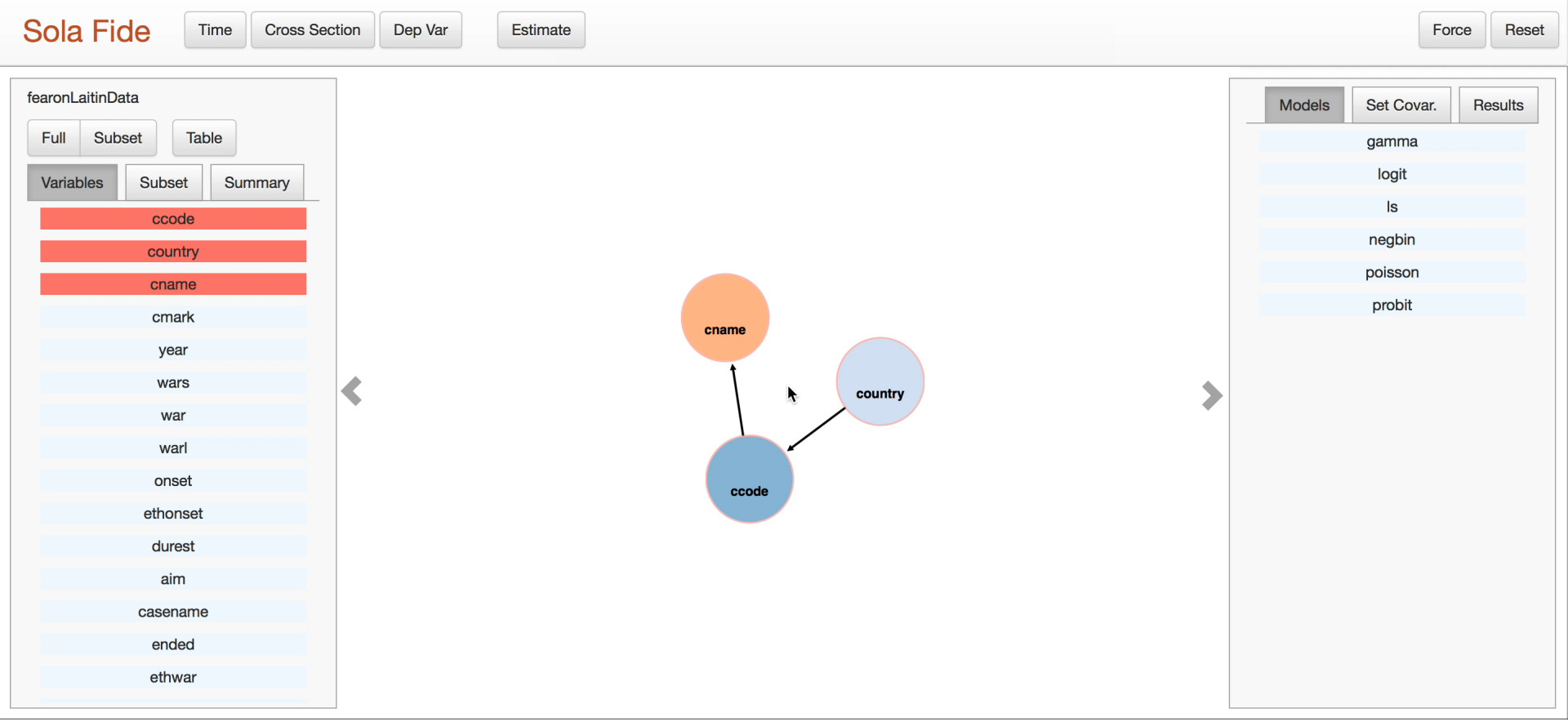
Dataverse 4.0



Target release date: June 23

- 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 **SolaFide**.

SolaFide Demo



Data Publishing Workflows

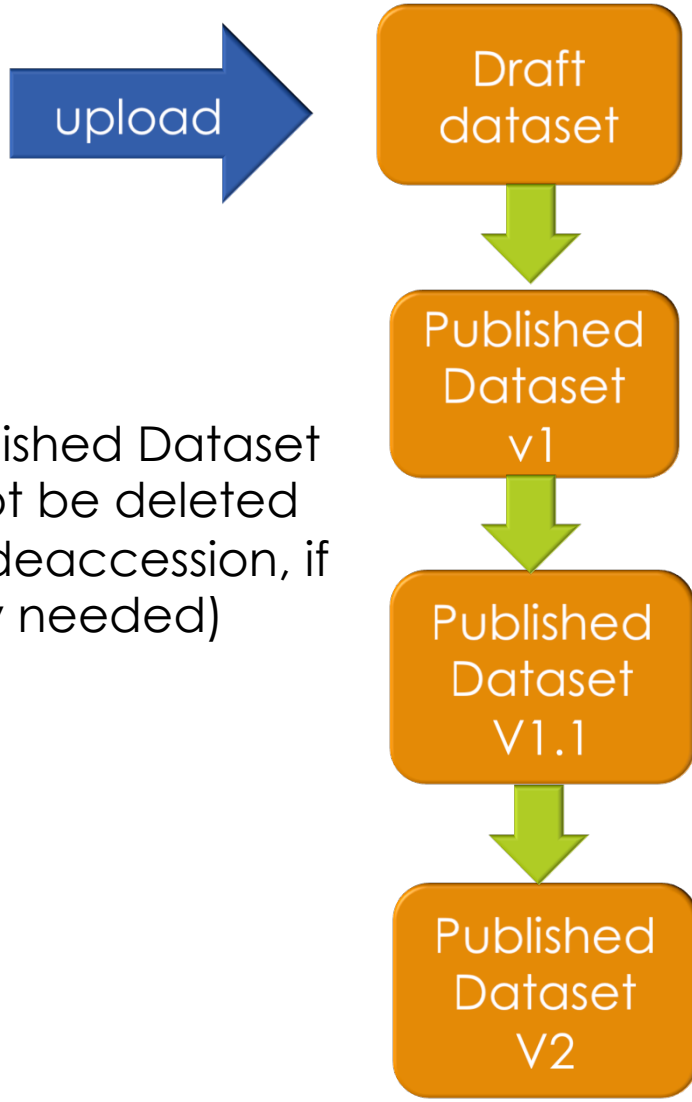
Data Publishing Guidelines

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 Publishing Workflow



A Published Dataset cannot be deleted (only deaccession, if legally needed)

Workflows that Integrate with Journals

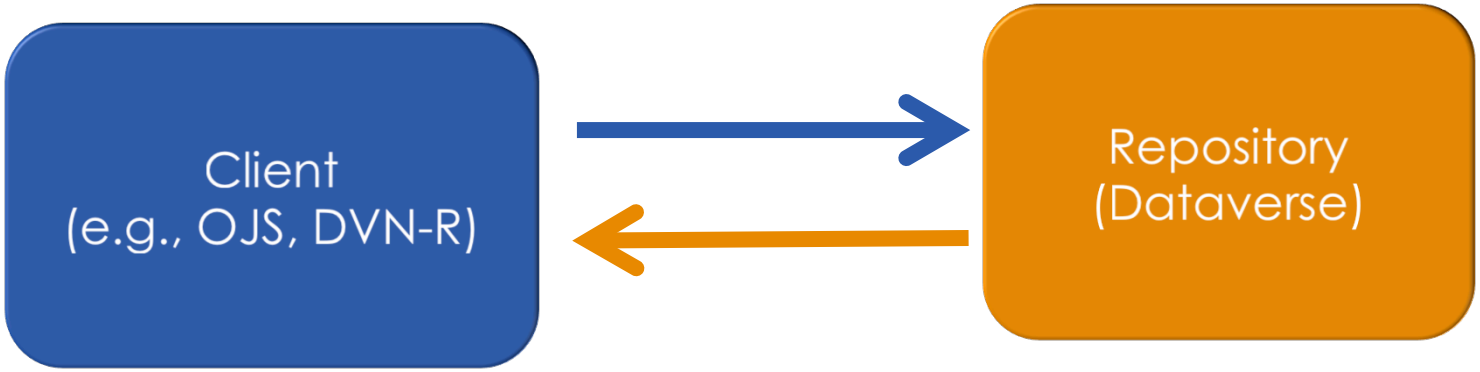
1. Publish a dataset to your Dataverse, then provide the Data Citation to the journal.
2. Contribute to a journal Dataverse:
    1. Add dataset to Journal Dataverse as a draft.
    2. Journal Editor reviews it, and approves it for release.
    3. Dataset is published with Data Citation and link from journal article to the data.
3. Seamless Integration between journal system and Dataverse.

OJS and Dataverse Integration

- 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

Detailed System Integration

- ✓ 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.
- ✓ Support HTTP verbs: GET, PUT, POST, and DELETE.



- ✓ XML file: "Deposit Receipt"
- ✓ HTTP status code: 200, 201, 204, 404, 405, 406, 412, 415

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

Deposit API based on SWORD

- Follows SWORD2 specifications
- SWORD is known and supported within academic publishing; a "profile" of the AtomPub standard.
- 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:
  - Use own Data API
  - Plan to add this support in the future
- Add XML attribute to pass article citation from client:
  - Allow DCterms:isReferencedby to contain attributes such as HoldingsURI to link back to article from Dataverse
  - This is now part of the SWORD PHP client library
- Use “In-Progress: false” to indicate that dataset is ready to be published (In SWORD spec means deposit complete)

Support for Metadata Standards

- A core or **citation metadata** that applies to all datasets – Supported currently by Data Deposit API
- Extensible metadata blocks for specific domains:
  - **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.

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 | End | +

YYYY-MM-DD | YYYY-MM-DD

Date of Collection

Start | End | +

YYYY-MM-DD | YYYY-MM-DD

Country/Nation

Geographic Coverage

Geographic Unit

Geographic Bounding Box

West Longitude | East Longitude

North Latitude | South Latitude

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	End	
YYYY-MM-DD	YYYY-MM-DD	+

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

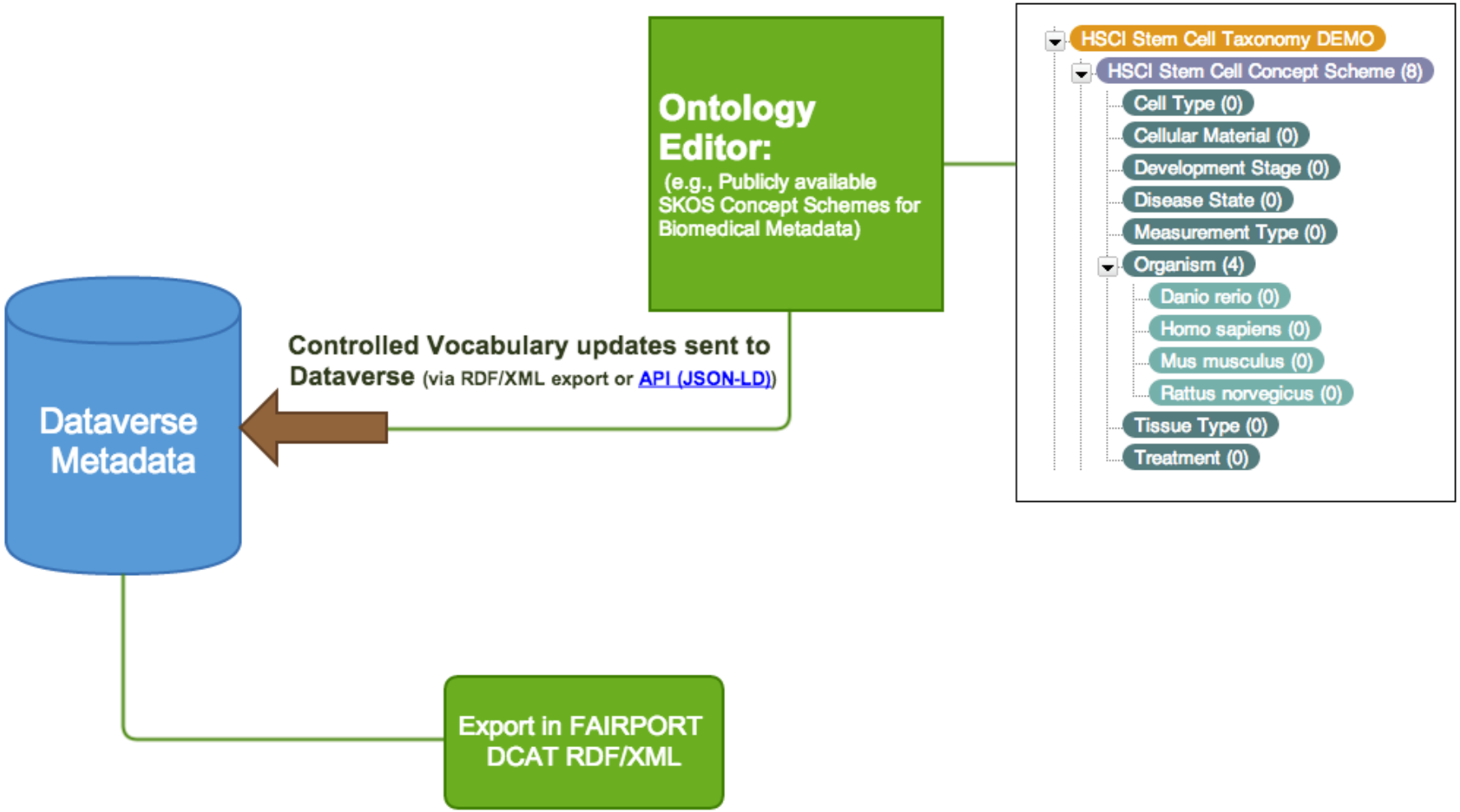
- [ ] DNA Methylation Profiling (Bisulfite-Seq)
- [ ] DNA Methylation Profiling (MeDIP-Seq)
- [ ] Histone Modification (ChIP-Seq)
- [ ] Protein-RNA Binding (RIP-Seq)
- [ ] Transcription Factor Binding (ChIP-Seq)

Organism

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

Cell Type

+


Ontology Editor:
(e.g., Publicly available SKOS Concept Schemes for Biomedical Metadata)
HSCI Stem Cell Taxonomy DEMO
HSCI Stem Cell Concept Scheme (8)
Cell Type (0)
Cellular Material (0)
Development Stage (0)
Disease State (0)
Measurement Type (0)
Organism (4)
Danio rerio (0)
Homo sapiens (0)
Mus musculus (0)
Rattus norvegicus (0)
Tissue Type (0)
Treatment (0)
Controlled Vocabulary updates sent to Dataverse (via RDF/XML export or API (JSON-LD))
Dataverse Metadata
Export in FAIRPORT DCAT RDF/XML

Upcoming

Expanding to Larger and More Types of Data

- Sharing sensitive data with DataTags and Secure Dataverse
- Integration with other systems:
  - OSF
  - DataUp
  - WorldMap
  - DataBridge
  - ORCID
  - DASH (at Harvard)
- Expand to Larger data sets

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

Tagging Complete!

Direct Data Access

CriminalPenalties

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

THANKS

mcrosas@iq.harvard.edu Twitter: mercecrosas

http://datascience.iq.harvard.edu (Beta)