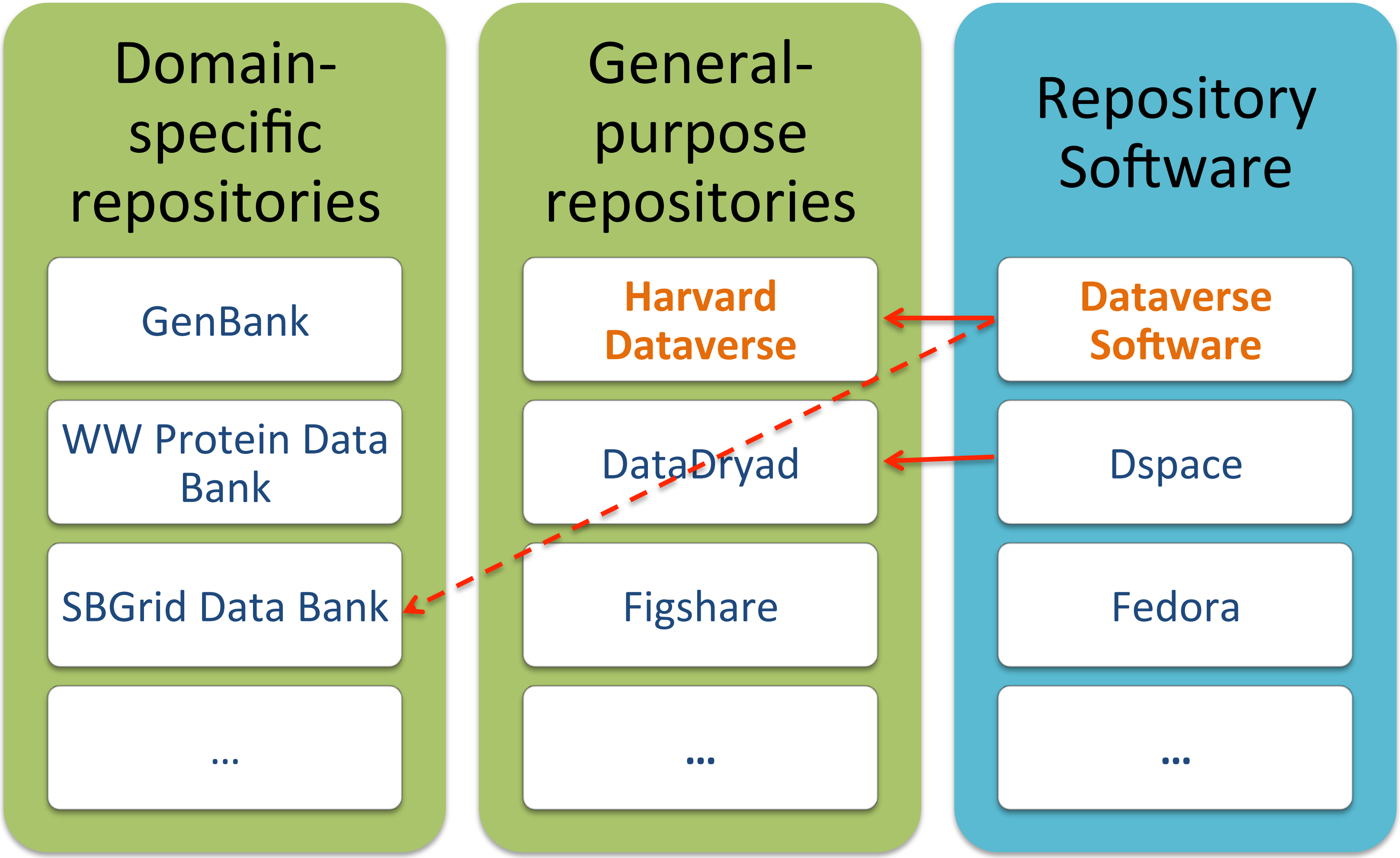
Mercè Crosas, Ph.D.
Chief Data Science and Technology Officer
Institute for Quantitative Social Science
Harvard University
@mercecrosas

DATAVERSE ON THE MOC

MOC Workshop, Boston University, November 19, 2015

Data Repositories vs Repository Software



Domain-specific repositories

- GenBank
- WW Protein Data Bank
- SBGrid Data Bank
- ...

General-purpose repositories

- **Harvard Dataverse**
- DataDryad
- Figshare
- ...

Repository Software

- **Dataverse Software**
- Dspace
- Fedora
- ...

The Dataverse Project

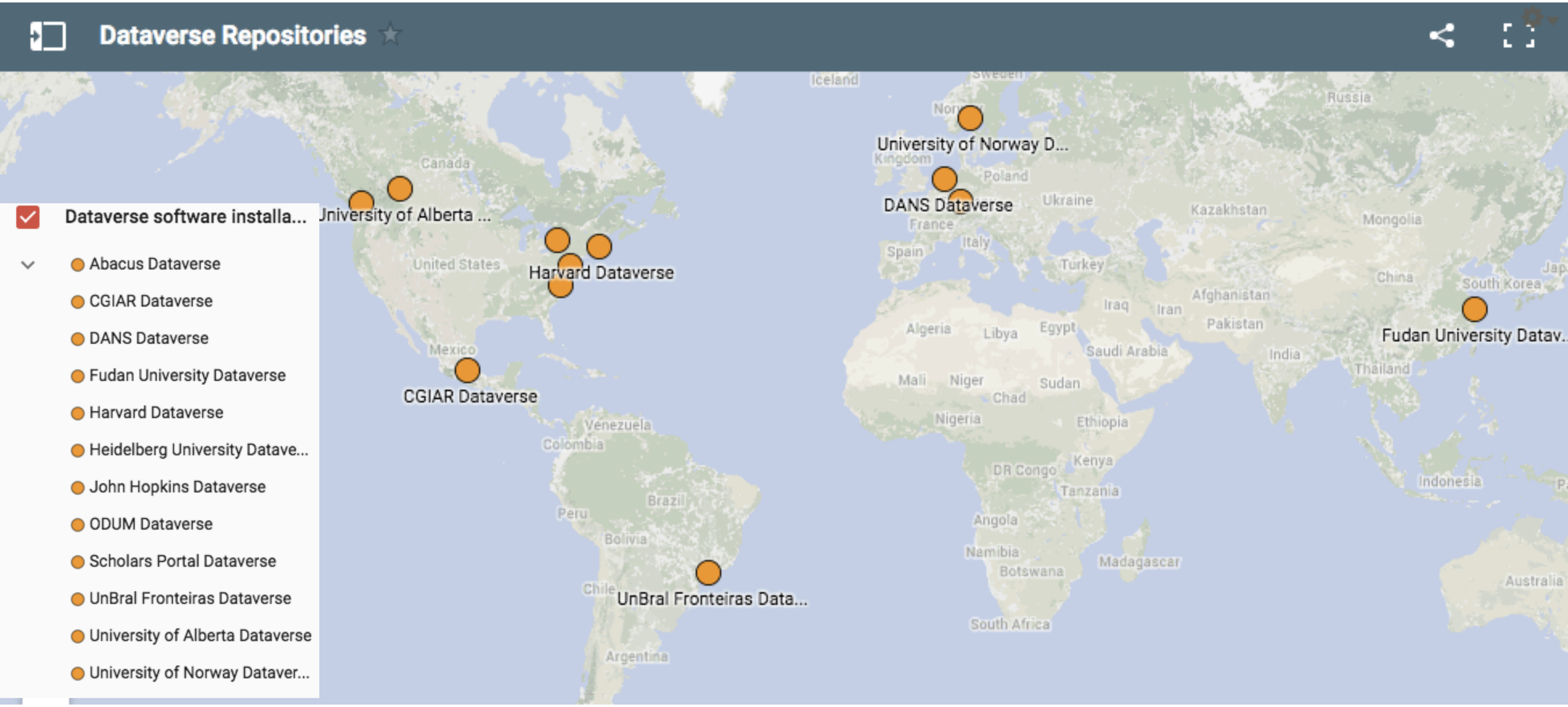
dataverse.org

Open-source software developed at Harvard’s IQSS since 2006

Used to share, publish, cite and archive research data

Installed in 12 sites world wide

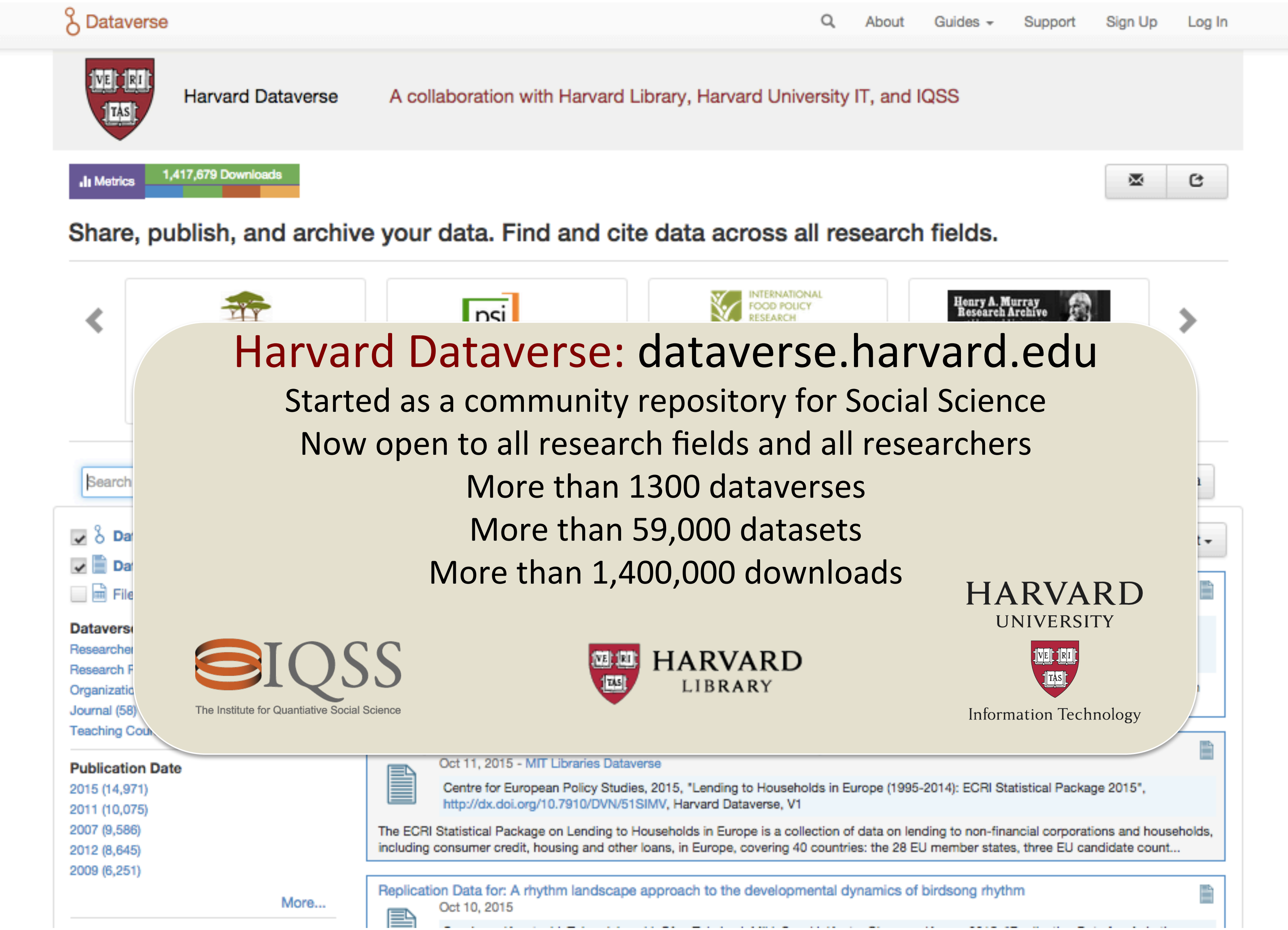
Serving 100s of universities and organizations



Dataverse Repositories ☆

☑ Dataverse software installa...

- ● Abacus Dataverse
- ● CGIAR Dataverse
- ● DANS Dataverse
- ● Fudan University Dataverse
- ● Harvard Dataverse
- ● Heidelberg University Datave...
- ● John Hopkins Dataverse
- ● ODUM Dataverse
- ● Scholars Portal Dataverse
- ● UnBral Fronteiras Dataverse
- ● University of Alberta Dataverse
- ● University of Norway Dataver...



Harvard Dataverse

A collaboration with Harvard Library, Harvard University IT, and IQSS

Share, publish, and archive your data. Find and cite data across all research fields.

psi
INTERNATIONAL FOOD POLICY RESEARCH
Henry A. Murray Research Archive
Search

Harvard Dataverse: dataverse.harvard.edu

Started as a community repository for Social Science

Now open to all research fields and all researchers

More than 1300 dataverses

More than 59,000 datasets

More than 1,400,000 downloads

IQSS
The Institute for Quantiative Social Science

HARVARD LIBRARY

HARVARD UNIVERSITY
Information Technology

Publication Date
2015 (14,971)
2011 (10,075)
2007 (9,586)
2012 (8,645)
2009 (6,251)
More...

Oct 11, 2015 - MIT Libraries Dataverse

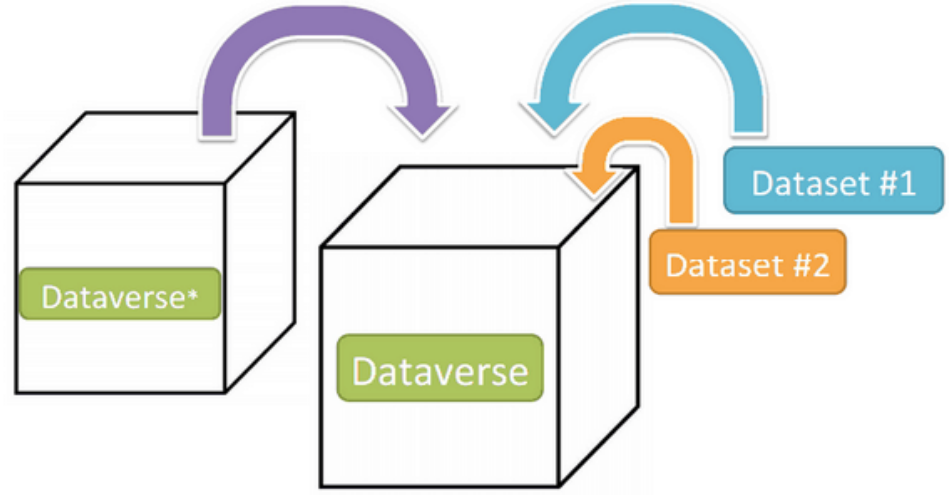
Centre for European Policy Studies, 2015, "Lending to Households in Europe (1995-2014): ECRI Statistical Package 2015", http://dx.doi.org/10.7910/DVN/51SIMV, Harvard Dataverse, V1

The ECRI Statistical Package on Lending to Households in Europe is a collection of data on lending to non-financial corporations and households, including consumer credit, housing and other loans, in Europe, covering 40 countries: the 28 EU member states, three EU candidate count...

Replication Data for: A rhythm landscape approach to the developmental dynamics of birdsong rhythm
Oct 10, 2015

Dataverses are containers for Datasets

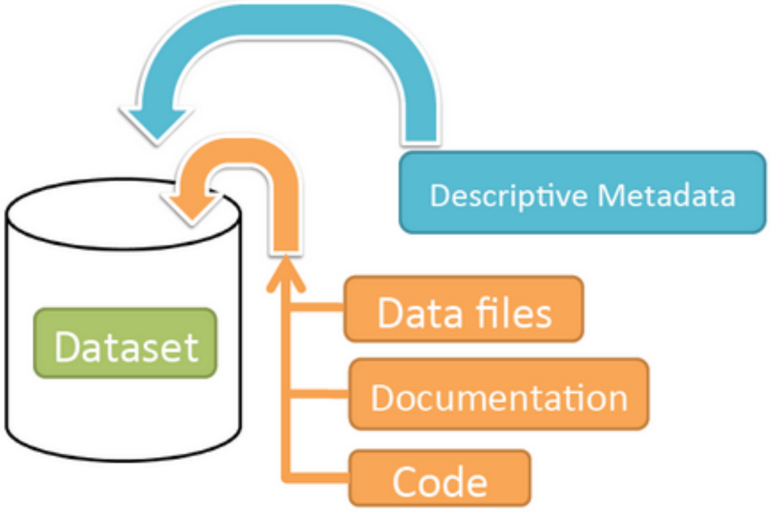
Schematic Diagram of a **Dataverse** in Dataverse 4.0



Container for your **Datasets** and/or **Dataverses***

* Dataverses can now contain other Dataverses (this replaces Collections & Subnetworks)

Schematic Diagram of a **Dataset** in Dataverse 4.0



Container for your data, documentation, and code.

Each Dataverse can be for a researcher, a research project, a department, a journal, or a larger organization.

Dataverse offers a rich feature set to publish research data

Credit and Visibility

- Standard, persistent data citation
- Branding for each dataverse
- Widgets to embed in your own website

Discovery

- Faceted search for all metadata
- Standard metadata:
  - citation
  - scientific domain
  - file-level

Access Control & Roles

- CCO waiver for public datasets
- Tiered access:
  - terms of use
  - guestbook
  - restricted data
- Publishing workflow
- Multiple roles:
  - contribute
  - curate, review
  - administrate

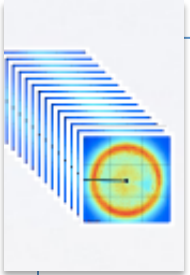
Data Features

- Versioning
- Conversion of tabular data files to standard format
- Automatic extraction of file metadata (R, STATA, SPSS, XSLX, FITS)

Interoperability through APIs

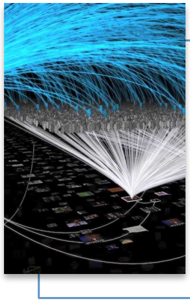
Journal Systems (Open Journal System, ScholarOne); Open Science Framework
Data Analysis (TwoRavens); Spatial Viz (WorldMap); Preservation systems (Archivematica)

Current Collaborations: Addressing the Next Challenges in Data Sharing



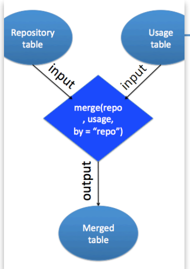
Structural Biology Grid Data Repository (Sliz, HMS, Crosas, IQSS)

THE LEONA M. AND HARRY B. HELMSLEY CHARITABLE TRUST



Social Science Big Data (King, Crosas, IQSS, CGA)

ALFRED P. SLOAN FOUNDATION 1934



Data Provenance (Seltzer, SEAS, Crosas, King, IQSS)

NSF

Privacy Tools to share sensitive data (SEAS, Berkman Center, Privacy Lab, IQSS, MIT)

NSF

ALFRED P. SLOAN FOUNDATION 1934

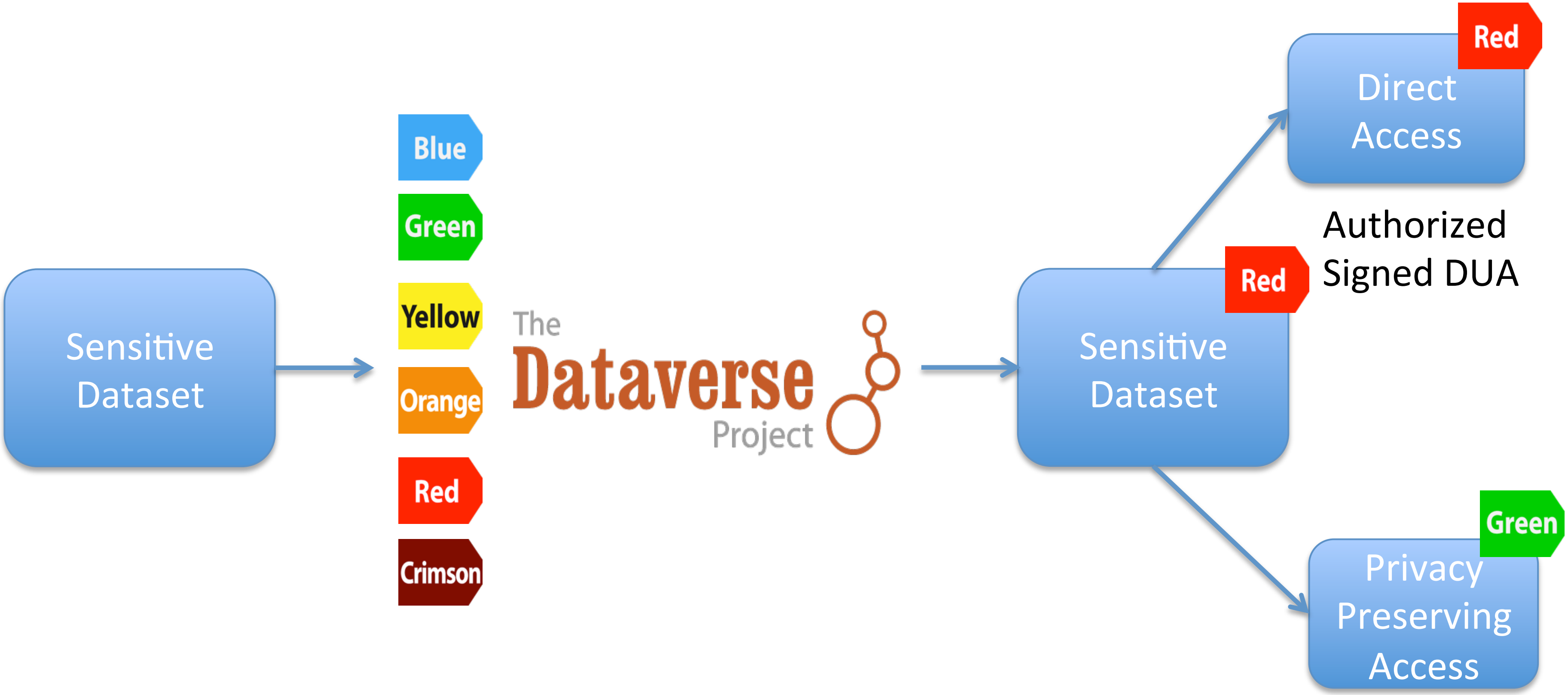
Sharing Sensitive Data with Confidence: DataTags System

Tag Type	Description	Security Features	Access Credentials
Blue	Public	Clear storage, Clear transmit	Open
Green	Controlled public	Clear storage, Clear transmit	Email- or OAuth Verified Registration
Yellow	Accountable	Clear storage, Encrypted transmit	Password, Registered, Approval, Click-through DUA
Orange	More accountable	Encrypted storage, Encrypted transmit	Password, Registered, Approval, Signed DUA
Red	Fully accountable	Encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA
Crimson	Maximally restricted	Multi-encrypted storage, Encrypted transmit	Two-factor authentication, Approval, Signed DUA

DataTag: A set of security features and access requirements for file handling

Sweeney, Crosas, Bar-Sinai, 2015, Technology Science

Data Sharing Workflow for Sensitive Data



Sensitive Dataset

Blue

Green

Yellow

Orange

Red

Crimson

The Dataverse Project

Red

Sensitive Dataset

Red

Direct Access

Authorized Signed DUA

Green

Privacy Preserving Access

DATA PRIVACY LAB

IQSS The Institute for Quantitative Social Science

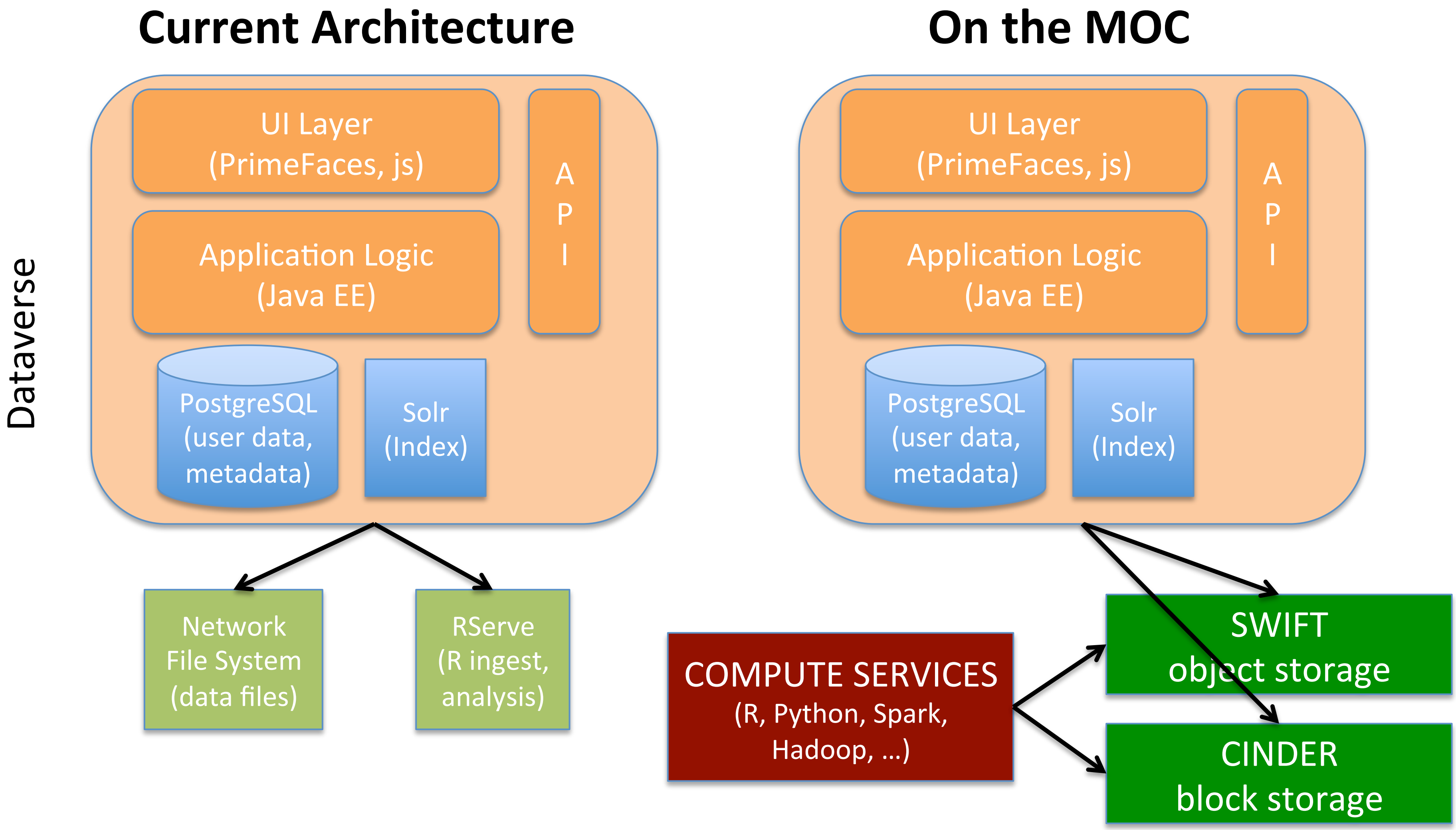
BERKMAN CENTER FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

CRCS Center for Research on Computation and Society at Harvard School of Engineering and Applied Sciences

Program on Information Science MIT Libraries

http://datatags.org
http://privacytools.seas.harvard.edu

Dataverse on the MOC



Current Architecture
On the MOC
Dataverse
UI Layer (PrimeFaces, js)
Application Logic (Java EE)
A P I
PostgreSQL (user data, metadata)
Solr (Index)
Network File System (data files)
RServe (R ingest, analysis)
UI Layer (PrimeFaces, js)
Application Logic (Java EE)
A P I
PostgreSQL (user data, metadata)
Solr (Index)
COMPUTE SERVICES (R, Python, Spark, Hadoop, …)
SWIFT object storage
CINDER block storage

Dataverse-MOC Projects

We propose to pilot a framework for sharing and publishing large and streaming datasets, and enabling collaborative computing on them.

Boston Data from City Hall and Boston Area Research Initiative (BARI):

- Dan O'Brien (Northeastern University)
- Storage: 911 and 311 calls streaming data:
  - 311 calls: 800,000 reports, at 500 reports/day
  - 911 calls: 2,500,000 reports, at 1,500 reports/day
- Computing: merge data + geospatial, temporal exploration

Partners Healthcare Clinical Trial Data:

- Shawn Murphy (Partners Healthcare)
- Scalable Collaborative Infrastructure for a Learning Healthcare System (SCILHS)
  - Data from 14 health systems sites
- Sensitive data sets, categorized using the DataTags levels

@mercecrosas

mcrosas@iq.harvard.edu

http://scholar.harvard.edu/mercecrosas

THANKS