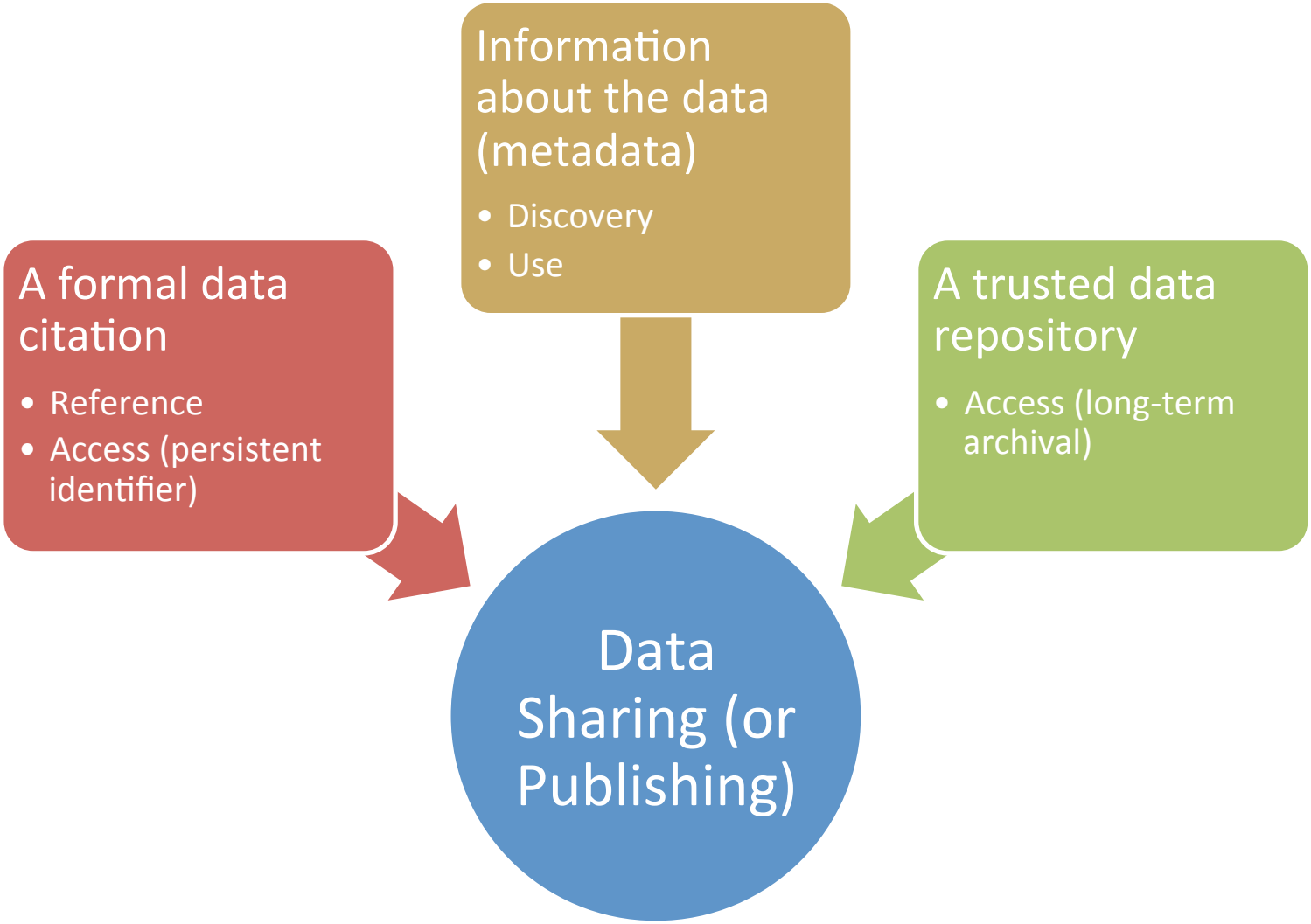
Mercè Crosas, Ph.D.
Chief Data Science and Technology Officer
Institute for Quantitative Social Science
Harvard University
@mercecrosas

ADDRESSING THE NEXT CHALLENGES IN DATA SHARING: LARGE-SCALE DATA AND SENSITIVE DATA

Data sharing:
good for you and good for the world

Researchers	→	Get credit for their data
Publishers and Journals	→	Verify published work
Federal funding agencies	→	Make public assets accessible
Science	→	Validate, reuse and extend previous work

Data Sharing needs to support data discovery, referencing, access, and reuse



The Dataverse Project

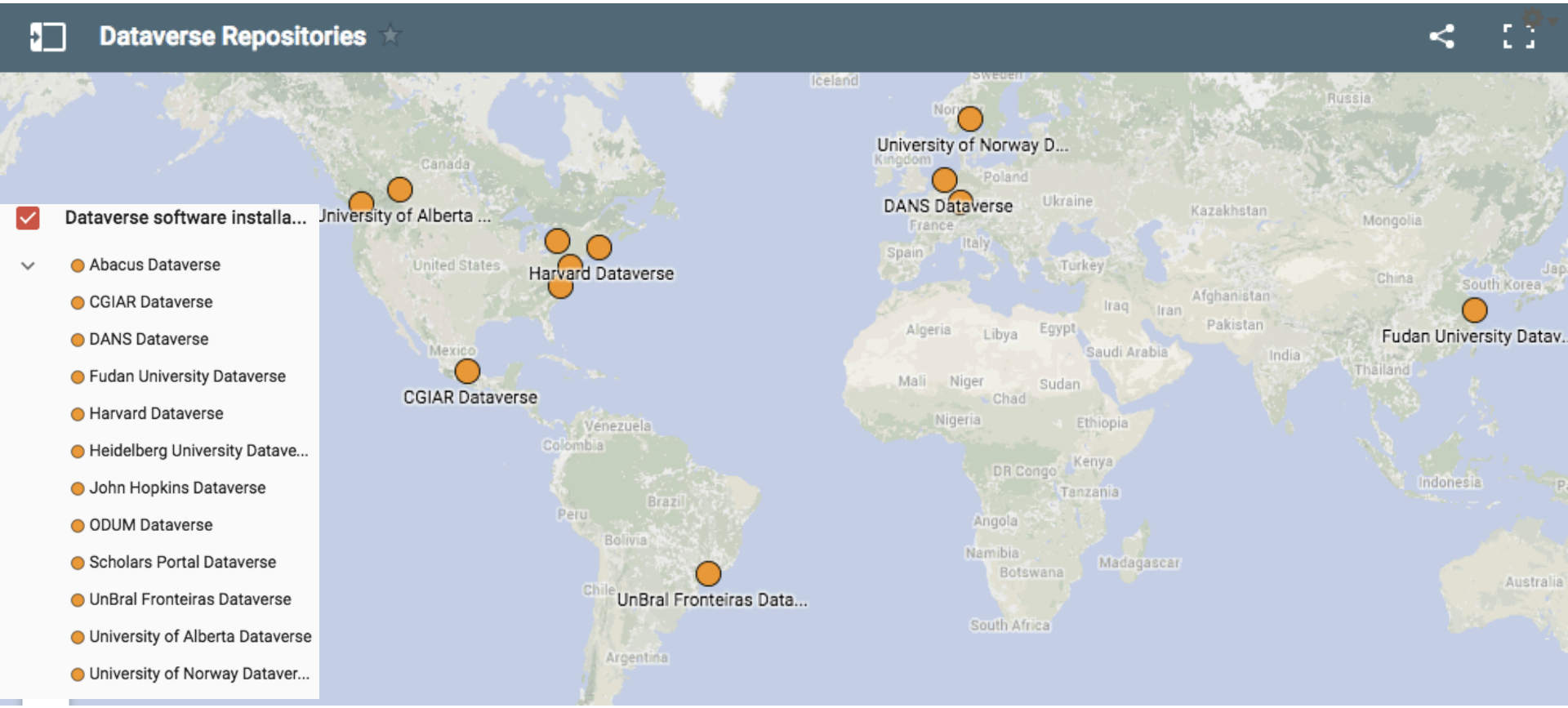
dataverse.org

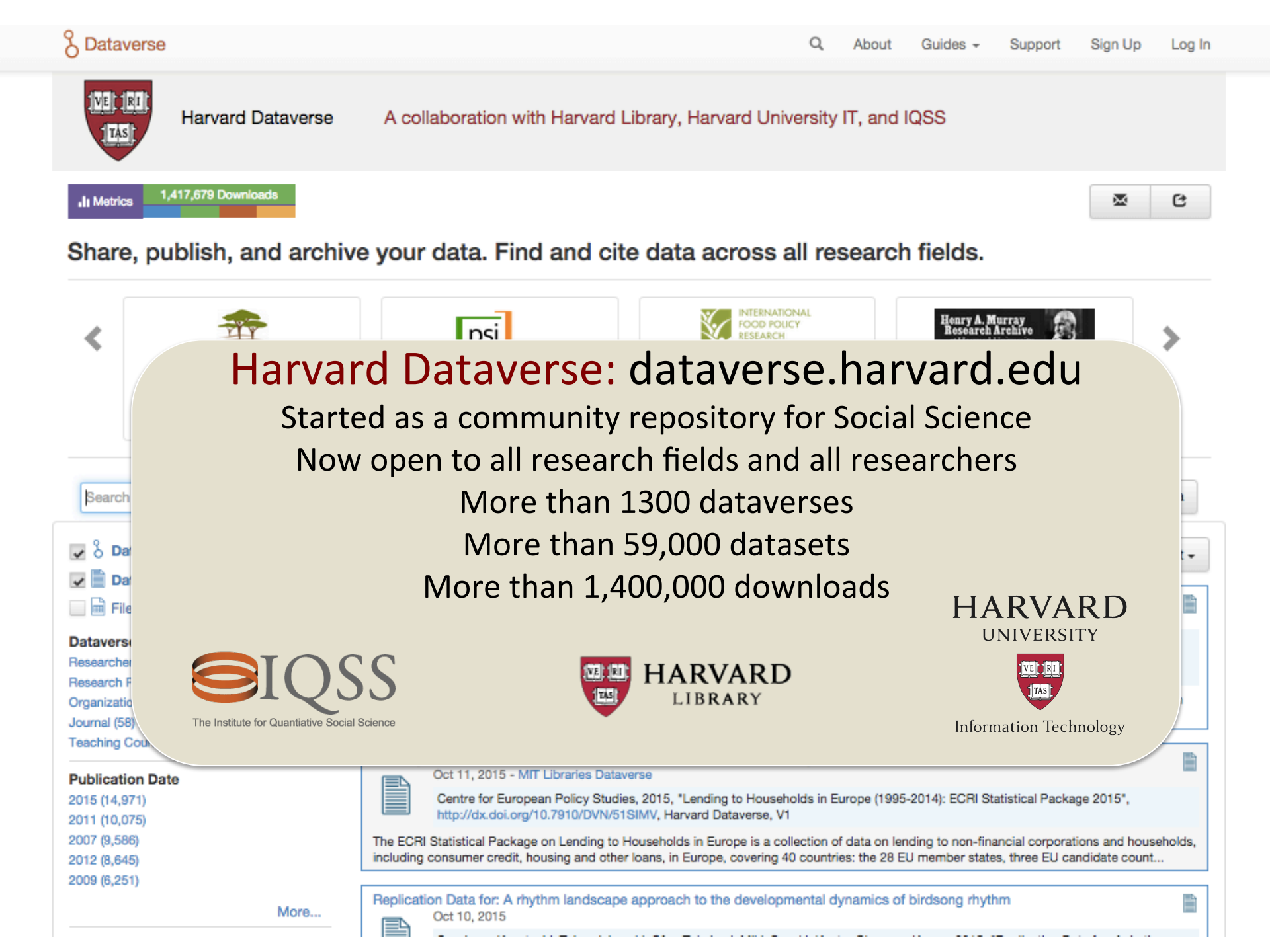
Open-source software developed at Harvard's IQSS since 2006

Used to share, publish, cite and archive research data

Installed in 12 sites world wide

Serving 100s of universities and organizations




Dataverse
About
Guides
Support
Sign Up
Log In
Harvard Dataverse
A collaboration with Harvard Library, Harvard University IT, and IQSS
Metrics
1,417,679 Downloads
Share, publish, and archive your data. Find and cite data across all research fields.
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

Data Sharing with Dataverse

Now

- No sensitive data
- Seldom versioning
- Datasets up to ~GB

The Next 5 Years

- Highly-sensitive data
- Streaming or frequently updated data
- Datasets > GBs, TBs, PBs
  - Thousands of files per dataset
  - Large dataset in a Big Data, NoSQL storage (MongoDB, Cassandra, Lucene)

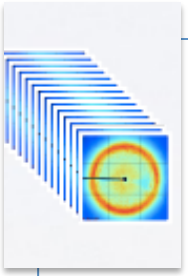
Large-scale data sharing needs to continue supporting discovery, referencing, access and reuse.

Adhering to the same high standards for large-scale data

- Metadata for discovery:
  - citation metadata
  - domain-specific descriptive metadata
  - file-level or variable metadata
- Data citation for reference and access:
  - for entire dataset and for subsets of the dataset (based on time of retrieval or variables selected)
- Fast queries, data exploration and visualizations for reuse:
  - might not be able to download entire dataset

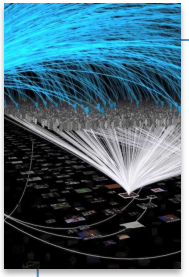
Data retrieval, explorations and visualizations of large-scale datasets require data repositories be closer to computing resources.

Current collaborations to address the next challenges in data sharing



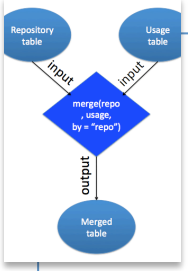
SB Grid Data Repository (HMS, IQSS)

THE LEONA M. AND HARRY B. HELMSLEY CHARITABLE TRUST



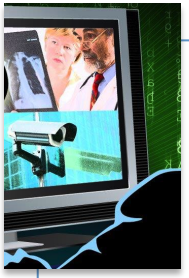
Social Science Big Data (IQSS)

ALFRED P. SLOAN FOUNDATION 1934



Data Provenance (SEAS, IQSS)

NSF



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

NSF

ALFRED P. SLOAN FOUNDATION 1934

Structural Biology Data Grid

+

The Dataverse Project

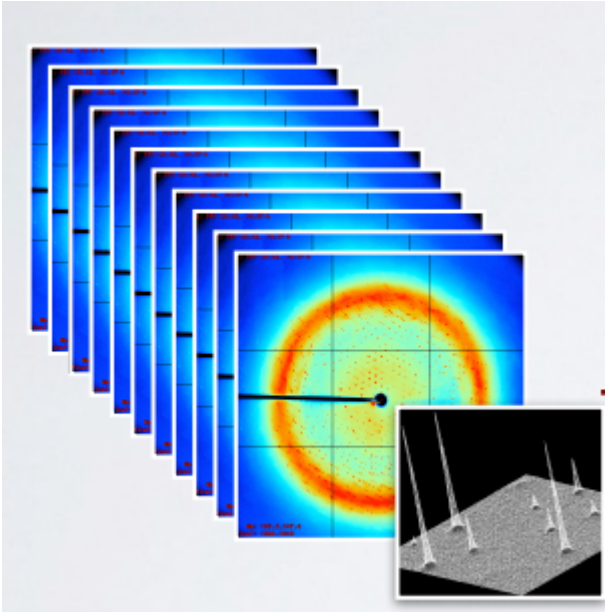
Sharing and Preserving Large Structural Biology Data

https://data.sbgrid.org/

Funded by

THE LEONA M. AND HARRY B. HELMSLEY CHARITABLE TRUST

Structural Biology
Primary Data

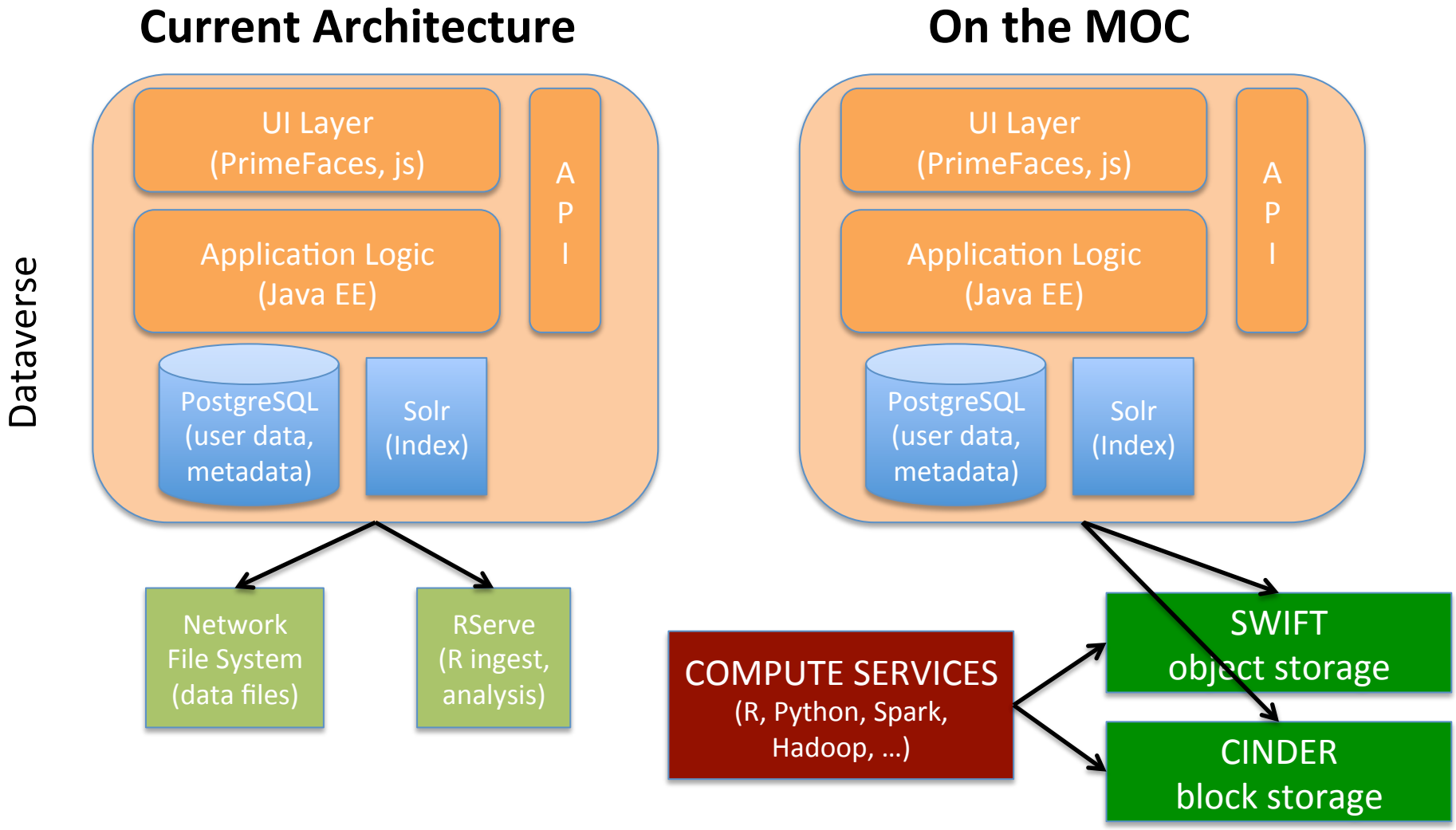


1 Dataset is 180-360 images of X-ray diffraction data, 3.5-7 GB; ~ 1TB per dataset, with a total up to 100 PBs

Integration with Dataverse:

- Long-term access
- Formal Data Citation
- Standard Metadata
- Data Exploration (OME)
- Preservation, with copies in multiple sites (following dataPASS approach)

Dataverse on the Massachusetts Open Cloud (MOC): Computing closer to data storage



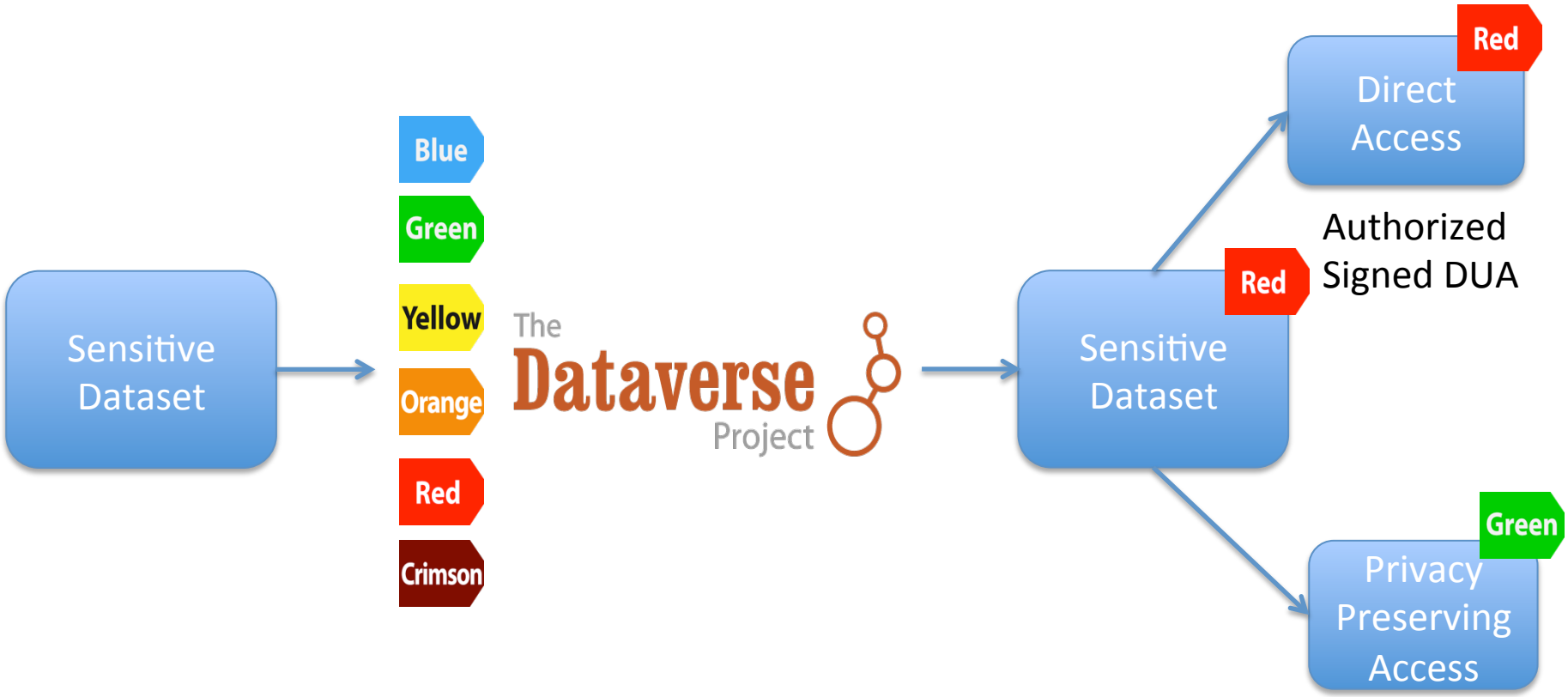
Sharing Sensitive Data with Confidence: DataTags System

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, "Sharing Sensitive Data with Confidence: The DataTags System" Technology Science

Data Sharing Workflow for Sensitive Data



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

Piotrek Sliz (SBGrid, HMS), Latanya Sweeney (Data Privacy Lab, Harvard), Dataverse team (IQSS, Harvard)
@mercecrosas

THANKS