Data Sharing in the Next 5 Years

Mercè Crosas, Ph.D.
@mercecrosas

Director of Data Science
Institute for Quantitative Social Science (IQSS)
Harvard University

Northeastern Research Computing Charrette, Spet 9, 2015

Sharing research data in a data repository enables reuse, extension and validation of previous research work

The Dataverse Project

Dedicated to sharing, archiving and citing research data.

Add Data

Find Data

Get Recognition

A widely-used, open-source data repository framework for publishing data

Research data sets are becoming larger, more sensitive, and more frequently updated

Data Sharing with Dataverse

Now

No sensitive data

Datasets up to ~ GB

Seldom Versioning

The Next 5 Years

Highly-sensitive data

Datasets > GBs, TBs, PBs

Streaming data

What are we doing towards supporting these new types of data?

+

Sharing Sensitive Data with Confidence

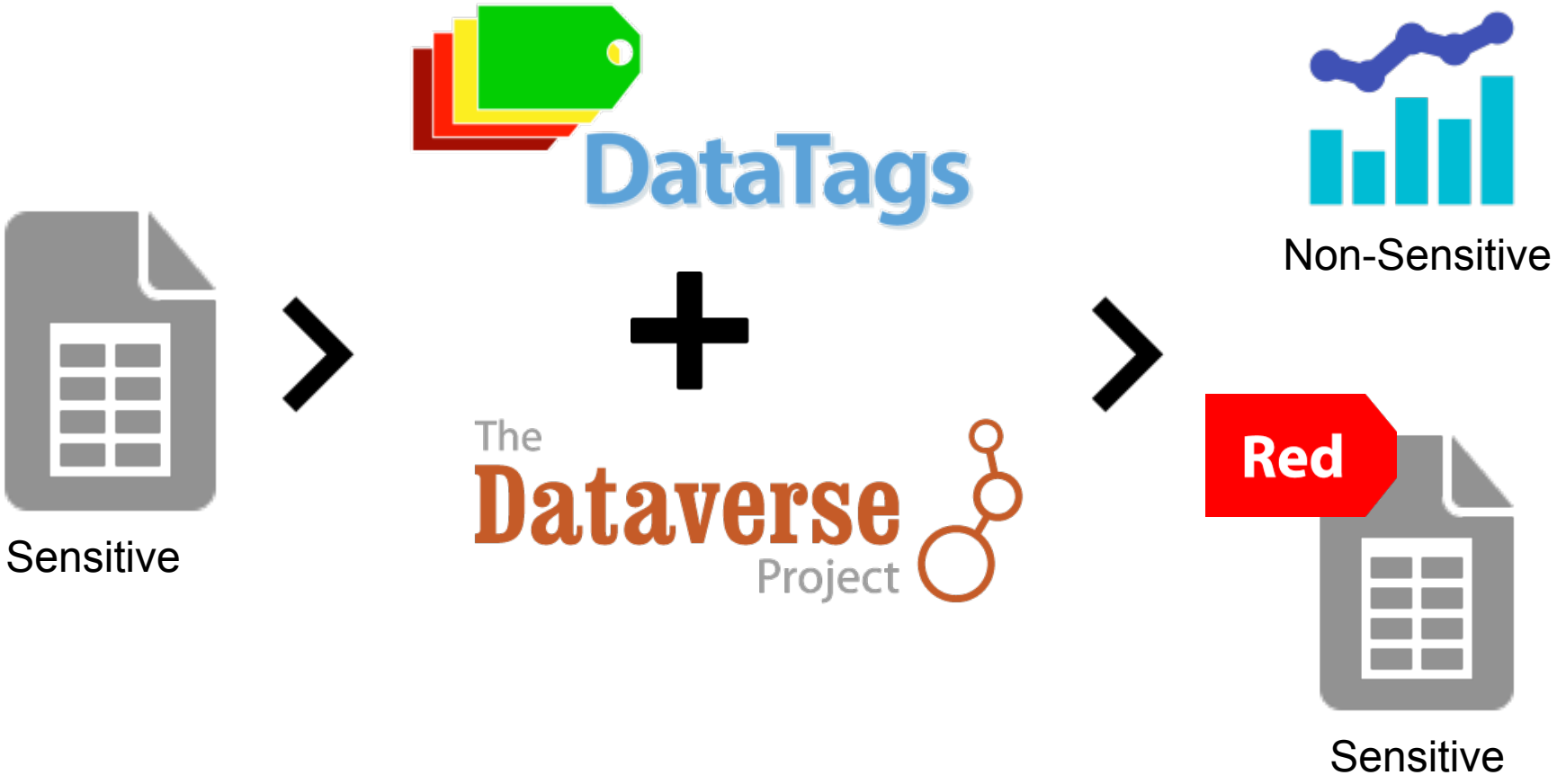
BERKMAN CENTER FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

Program on Information Science
MIT Libraries

Funded by

Standardized Levels of Data Sensitivity

Tag Type	Description	Transit	Storage	Access
Blue	Non-confidential information, stored and shared freely.	Clear	Clear	Open
Green	Not harmful personal information, shared with some access control.	Clear	Clear	Email, OAuth verified registration
Yellow	Potentially harmful personal information, shared with loosely verified and/or approved recipients.	Encrypted	Clear	Password, Registered, Approval click-through DUA
Orange	Sensitive personal information, shared with verified and/or approved recipients under agreement.	Encrypted	Encrypted	Password, Registered, Approval, signed DUA
Red	Very sensitive personal information, shared with strong verification of approved recipients under signed agreement.	Encrypted	Encrypted	Two-factor Auth, Registered, Approval, signed DUA
Crimson	Maximum sensitive, explicit permission for each transaction, strong verification of approved recipients under signed agreement.	Encrypted	Double Encrypted	Two-factor Auth, Registered, Approval, signed DUA

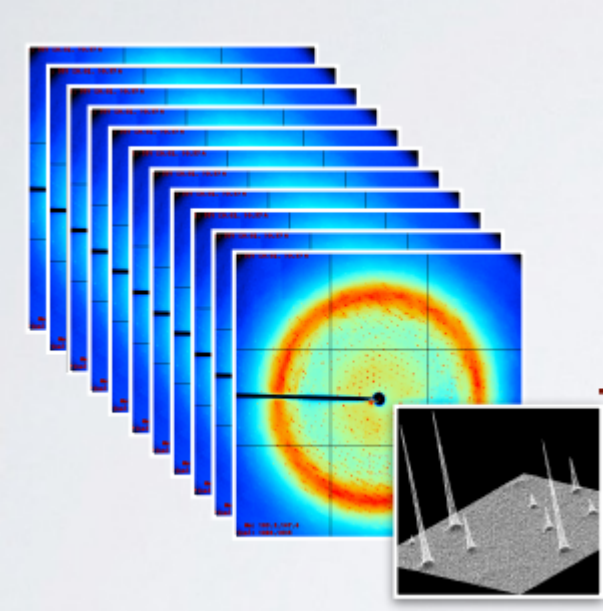

DataTags
Sensitive
The
Dataverse
Project
Non-Sensitive
Red
Sensitive

SBGrid CONSORTIUM + The Dataverse Project

Sharing and Preserving Large Structural Biology Data

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;
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

There is a need for closer integration of data repositories with research computing resources to support the new types of data

Towards an Integrated Research Ecosystem to Support Data-Intensive Research

Thanks

dataverse.org
datatags.org
sbgrid.org
scholar.harvard.edu/mercecrosas