Recommendations for Implementing Open Science

Keynote, Swiss Open Science Action Plan
Lausanne, October 17, 2019

Mercè Crosas, Ph.D.
Chief Data Science and Technology Officer, IQSS
Harvard University's Research Data Officer, HUIT
@mercecrosas

An Open Science solution should be researcher-centric

National Academies of Sciences New Report: Open Science by Design (2018)

Researcher at the center:

- **Researcher contributes to open science**
- **Researcher takes advantage of the open science practices**
- **From data generation to validation, dissemination, and preservation**

OpenAire New White Paper: Achieving Open Science in the European Science Cloud (2019)

Researcher at the center:

- **Researcher publishes all kinds of scientific products (data, software, workflows)**
- **Open as the default**
- **Services provide technology and training for researchers**

Data Sharing is key to Open Science

Dataverse: an open-source platform for sharing and archiving research data

- Launched at Harvard in 2006
- Used in 6 continents
- 48 Dataverse sites
- 5500 dataverses (branded datasets collections)
- 120K datasets
- 10M data files downloads
- Vibrant open-source community

A National Dataverse Site: DataverseNO

DataverseNO
Dataverse Network Norway

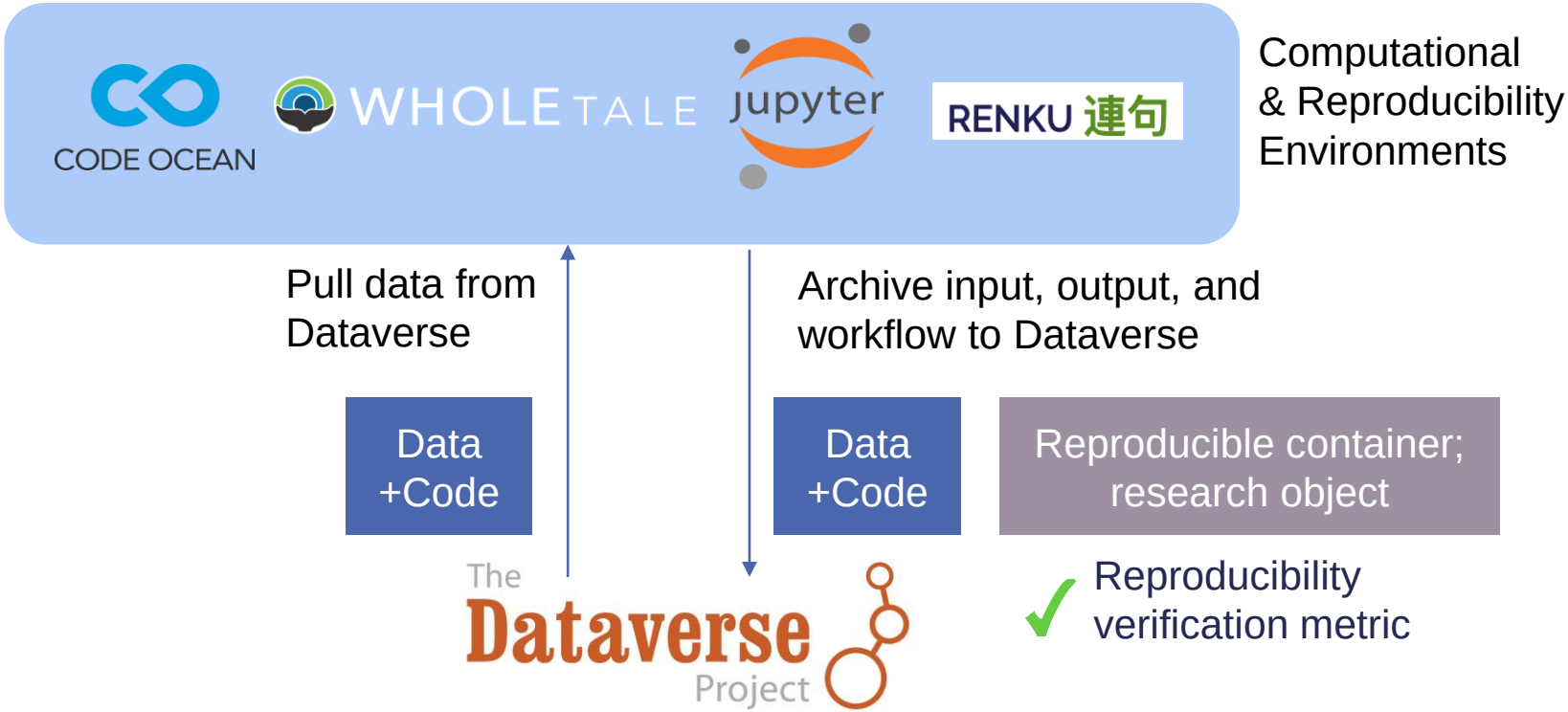
- **8 universities in Norway** as members**;** the other 3 to join soon
- **Policies and guidelines common** to all DataverseNO members
- Global and local support
- Applied for Core Trust Seal certificate
- Dataverse Europe Workshop January 2020
- https://site.uit.no/dataverseno/

But data sharing only is not sufficient for Open Science

Open Science should include sharing of:

- Code associated with data to reproduce research results
- Workflows to describe research steps and data transformations
- Research software, algorithms, and tools for reuse

Data repositories should be integrated with computational environments



Sharing data and computing together helps create a Data Commons to collaborate

"…*a data commons brings together (or co-locates) data with cloud computing infrastructure and commonly used software services, tools & applications for managing, analyzing and sharing data to create an interoperable resource for a research community*"

Robert Grossman

Data scientist at the University of Chicago; Director of the Open Commons Consortium

https://medium.com/@rgrossman1/a-proposed-end-to-end-principle-for-data-commons-5872f2fa8a47

Open Science does not always mean fully open data

Metadata should be open for discoverability;
But data restricted when needed.

DataTags: Standardized Data Policies

DataTag	Data Access	Authorization	Data Use Agreement	Encryption
Blue	Public			
Green	Public	+ Register		
Yellow	Restricted	+ Approval Needed	+ Click-thru DUA	+ Encrypted transit
Orange	Restricted	+ Approval Needed	+ Signed DUA	+ Encrypted transit + Encrypted storage
Red	Restricted	+ Approval Needed	+ Signed DUA + Two-factor Auth	+ Encrypted transit + Encrypted storage
Crimson	Restricted	+ Approval Needed	+ Signed DUA + Two-factor Auth	+ Encrypted transit + Multi-encrypted storage

Sweeney, Crosas, Bar-Sinai, 2015. Sharing Sensitive Data with Confidence: The DataTags System, Technology Science

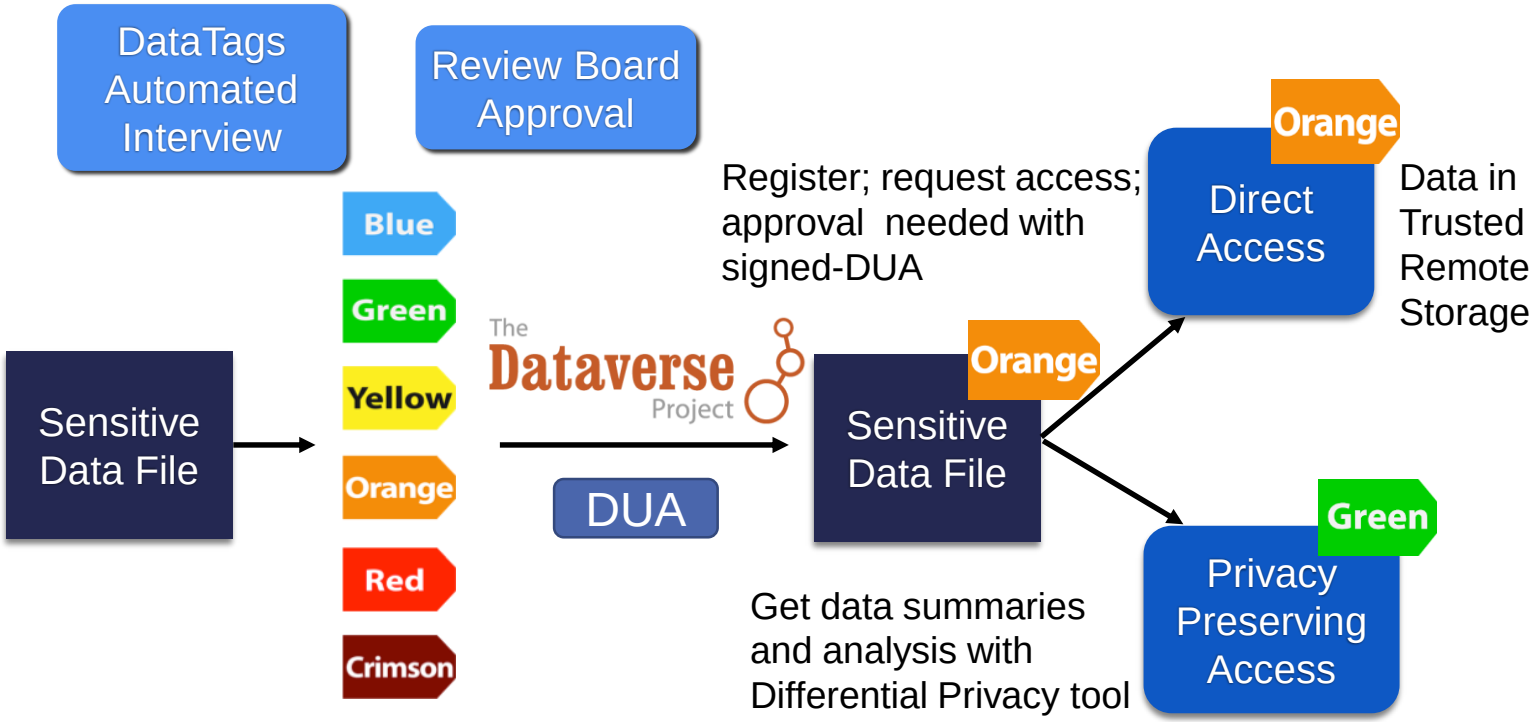
OpenDP: A New Project for Sensitive Data

A **community effort** to build a **trustworthy** and **open-source** suite of **differential privacy** tools that can be **easily adopted** by custodians of sensitive data to make it available for statistical research.

- To be launched in **2020** with Sloan Foundation funding
- Initially led by **Harvard Privacy Tools** project (PIs: Vadhan, Honaker, King, Crosas)

*A tool (algorithm) is **differentially private** if its output cannot reveal whether any individual's data was included in the original dataset or not*.

Dataverse + DataTags + OpenDP: Sharing and analyzing sensitive data



In conclusion, an Open Science implementation should include:

- Incentives to share data, software/code, and other research outputs
- Metadata and format standards for discovery and reuse
- Machine-readable data for management and usage by computers
- Sufficient information to reuse the data, software, workflows (all research outputs)
- Support for software licenses and data use agreements
- **Public metadata (at a minimum for citation) even when data are restricted**
- **Integration of archival repositories with computational environments**
- **Solutions for collaborations that access sensitive, private data**

Thanks

The Dataverse® Project

IQSS
The Institute for Quantitative Social Science

HARVARD UNIVERSITY

GDCC
The Global Dataverse Community Consortium
Supporting Dataverse repositories around the world.

Harvard University
Privacy Tools Project

dataverse.org | dataversecommunity.global/ | scholar.harvard.edu/mercecrosas