DataTags and OpenDP

Sharing Sensitive Data

Workshop of the OECD's Committee for Scientific and Technological Policy

Revision of Recommendation concerning access to research data from public funding

OECD Conference Center, Paris, October 15 2019

Mercè Crosas, Ph.D.

Chief Data Science and Technology Officer, IQSS

Harvard University's Research Data Officer, HUIT

@mercecrosas

DataTags

Mercè Crosas, Latanya Sweeney, Michael Bar-Sinai

Dataverse Team

A **datatag** is a set of security features and access requirements for file handling.

A **datatags repository** is one that stores and shares data files in accordance with a standardized and ordered levels of security and access requirements

A DataTags repository must:

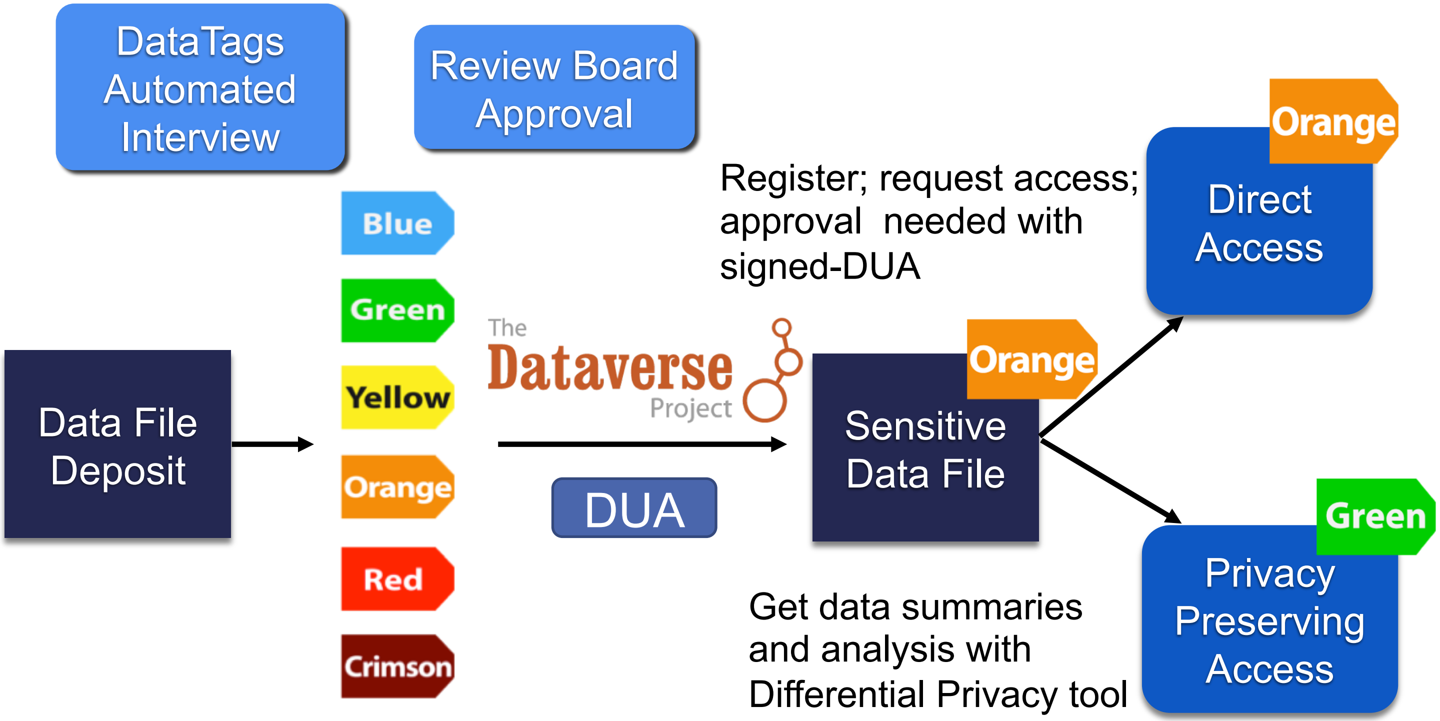
1. Support more than one **datatag**
2. Each file in the repository must have one and only one **datatag**
   a. additional requirements cannot weaken the file security
   b. and cannot required the same or more security than a more restrictive datatag
3. A recipient of a file from the **repository** must:
   a. satisfy file's access requirements,
   b. produce sufficient credentials as requested,
   c. and agree to any terms of use required to acquire the file.
4. Provide **technological guarantees** for requirements 1, 2 and 3.

Standardized Access and Security Levels

DataTag	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

DataTags, Dataverse, Privacy-Preserving Tools



DataTags Automated Interview

Review Board Approval

Data File Deposit

Blue

Green

Yellow

Orange

Red

Crimson

The Dataverse Project

DUA

Orange

Sensitive Data File

Register; request access; approval needed with signed-DUA

Orange

Direct Access

Get data summaries and analysis with Differential Privacy tool

Green

Privacy Preserving Access

OpenDP

Salil Vadhan, James Honaker, Gary King, Mercè Crosas

Harvard Privacy Tools team

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** by Sloan Foundation funding
- Initially led by **Harvard Privacy Tools** project

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

OpenDP: Use Cases Motivation

- **Archival data repositories** (e.g., Dataverse) enabling secondary reuse and reproducibility
- **Government agencies** making data available to the public, both for official statistics and open data mandates
- **Companies** sharing data for academic research or internal research.
- **Focus on "centralized model" for Differential Privacy:**
  - A central aggregator has accessed to the raw data. The aggregator transforms the data with a differentially private mechanism.

OpenDP: Principles

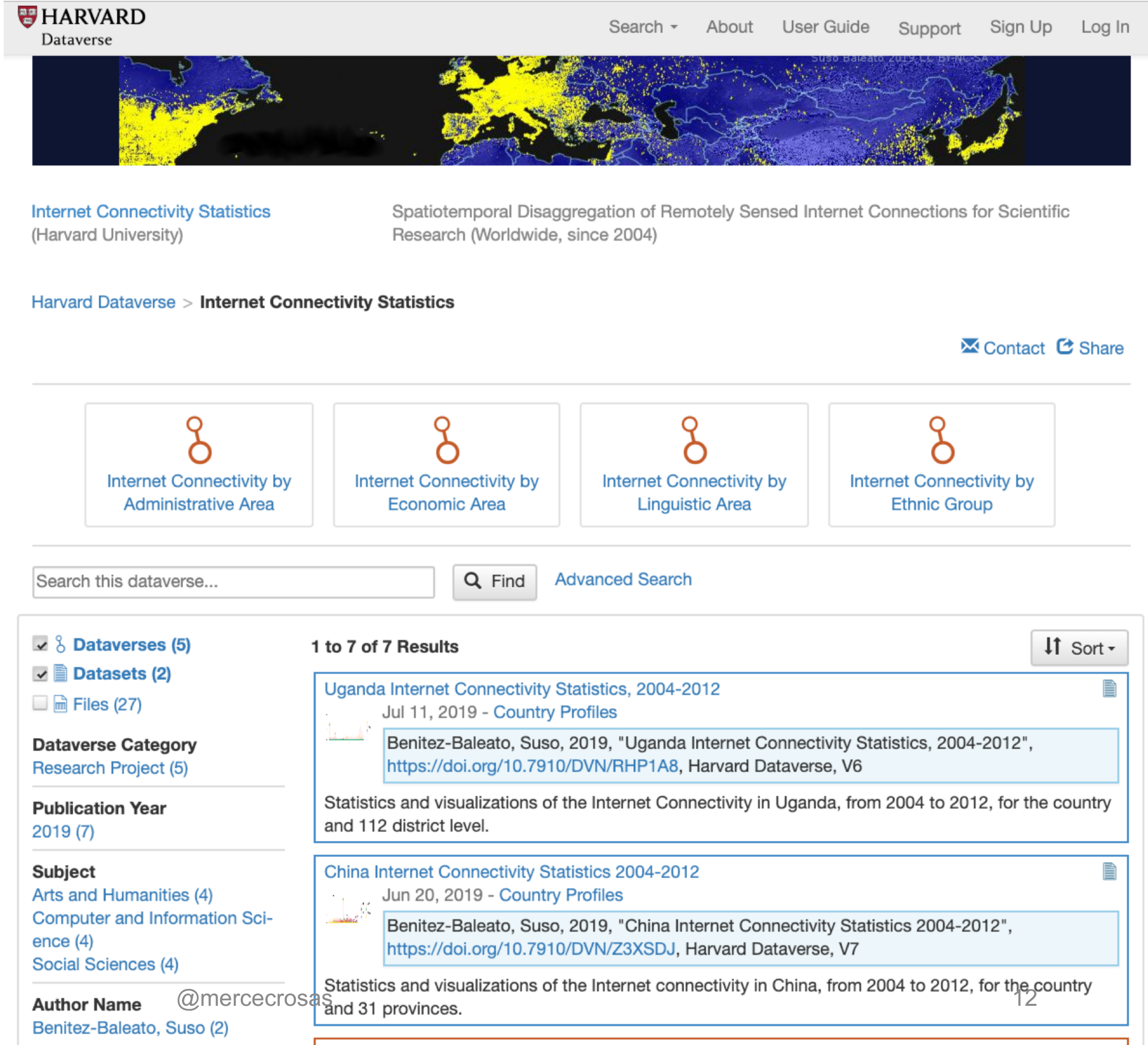
- **Open Source**
  - worldwide open-source community
  - processes and recognition for contribution
- **Security & Privacy**
  - careful vetting of any security-critical or privacy-critical code
  - can ship code to the sensitive data
- **Scalability**
  - handle petabyte-scale data
- **Extensibility**
  - can grow from the continuing research developments in the field

Example: Enabling Reproducibility with Privacy Preserving Data Sharing

Suso Baleato, James Honaker, Mercè Crosas

Internet Connectivity Statistics

- Data from measurements of the used **IPv4** address space across countries
- Close **correlation with internet penetration statistics** from International Telecommunications Union (ITU) and **OECD**



Internet Connectivity Statistics
(Harvard University)

Spatiotemporal Disaggregation of Remotely Sensed Internet Connections for Scientific Research (Worldwide, since 2004)

Harvard Dataverse > **Internet Connectivity Statistics**

Internet Connectivity by Administrative Area

Internet Connectivity by Economic Area

Internet Connectivity by Linguistic Area

Internet Connectivity by Ethnic Group

Search this dataverse...

☑ **Dataverses (5)**
☑ **Datasets (2)**
□ Files (27)

Dataverse Category
Research Project (5)

Publication Year
2019 (7)

Subject
Arts and Humanities (4)
Computer and Information Science (4)
Social Sciences (4)

Author Name
Benitez-Baleato, Suso (2)

1 to 7 of 7 Results

Uganda Internet Connectivity Statistics, 2004-2012
Jul 11, 2019 - Country Profiles

Benitez-Baleato, Suso, 2019, "Uganda Internet Connectivity Statistics, 2004-2012", https://doi.org/10.7910/DVN/RHP1A8, Harvard Dataverse, V6

Statistics and visualizations of the Internet Connectivity in Uganda, from 2004 to 2012, for the country and 112 district level.

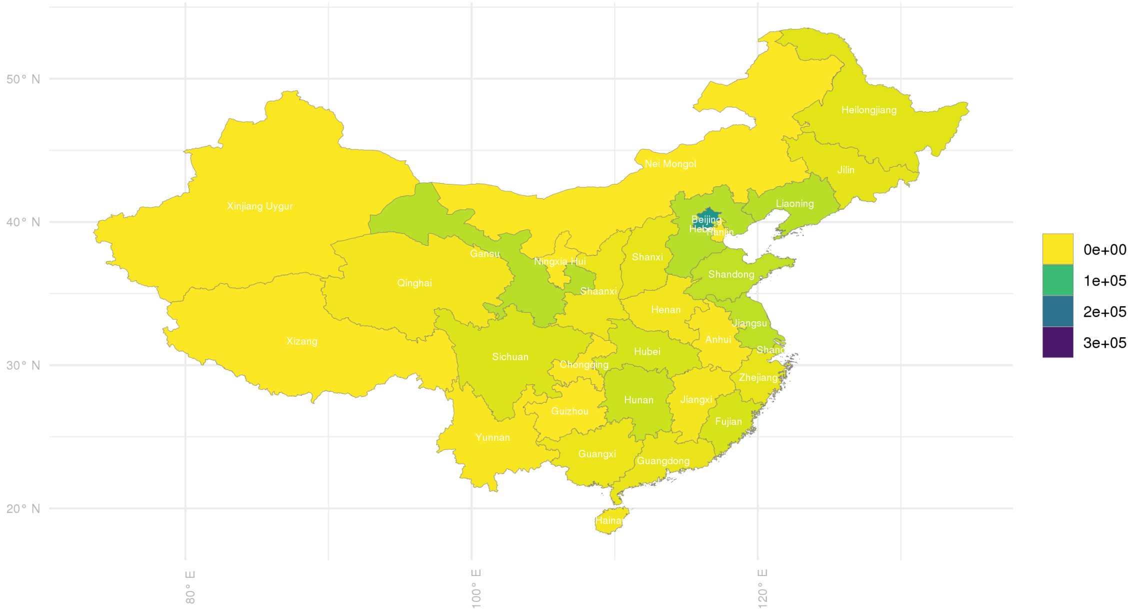
China Internet Connectivity Statistics 2004-2012
Jun 20, 2019 - Country Profiles

Benitez-Baleato, Suso, 2019, "China Internet Connectivity Statistics 2004-2012", https://doi.org/10.7910/DVN/Z3XSDJ, Harvard Dataverse, V7

Statistics and visualizations of the Internet connectivity in China, from 2004 to 2012, for the country and 31 provinces.

Differentially private data released in Dataverse

- Original data might contain **sensitive** information:
  - **DataTag = Orange**
- Statistics published in Dataverse are **differentially private**:
  - **DataTag = Green**
- With DP epsilon (noise) = 0.01, error added only 10^{-4}



Suso Baleato CC 3.0 BY-NC-SA <susobaleato@iq.harvard.edu>
Source & cite: https://doi.org/10.7910/DVN/Z3XSDJ Run: 20190322-11:01:50 (0.1-kaweco)

Take Away

- A solution for reproducibility and publishing of **sensitive data**:
  - **Dataverse + DataTags + Differential Privacy**
  - Acceptable precision loss; e.g. Error = 10^{-5} with Epsilon = 0.2 or Error = 10^{-4} with Epsilon = 0.1 (smaller epsilon, more privacy protection)
- Internet Connectivity Statistics as a use case:
  - https://dataverse.harvard.edu/dataverse/ics
- OpenDP will be launched by Harvard Privacy Tools project:
  - https://privacytools.seas.harvard.edu

Thanks

scholar.harvard.edu/mercecrosas