CLOUD DATAVERSE

Mercè Crosas, Institute for Quantitative Social Science, Harvard University

@mercecrosas

MOC WORKSHOP, OCTOBER 3, 2017, BOSTON UNIVERSITY

OUR INSTITUTE PROVIDES A TECHNOLOGY SOLUTION TO DATA SHARING

Institute for Quantitative Social Science, Harvard University

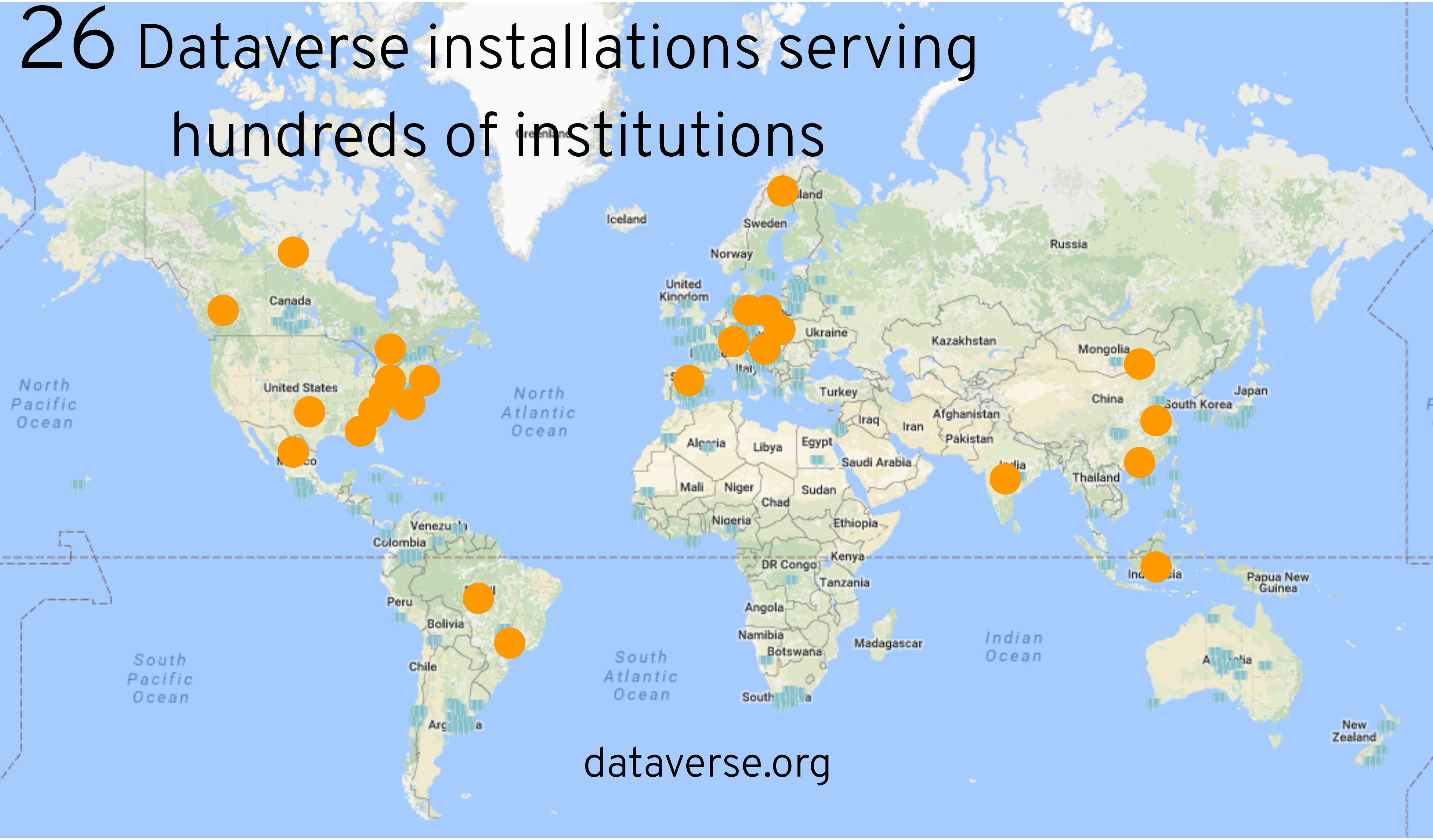
@IQSS

The Dataverse Project

An open-source software to share, cite, and find data.
Developed at Harvard's Institute for Quantitative Social Science
with the contribution of an active and growing community.

2006 (we started)

2017

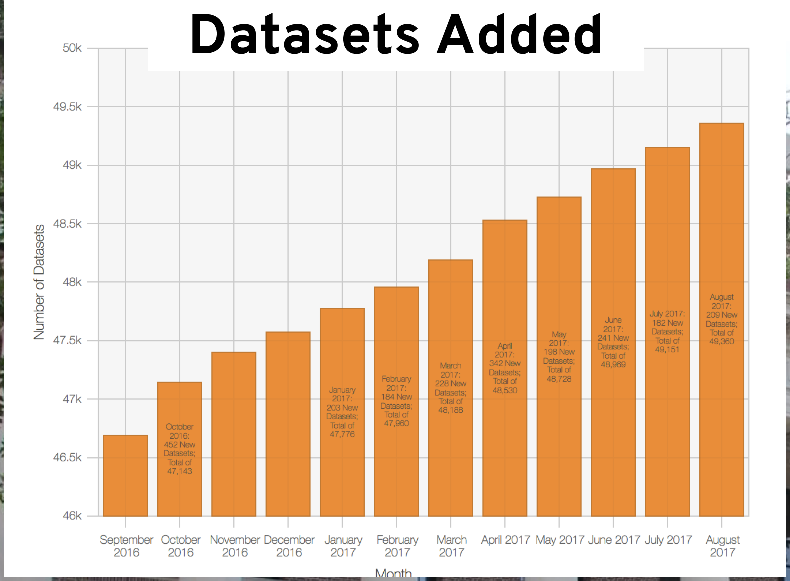


26 Dataverse installations serving hundreds of institutions

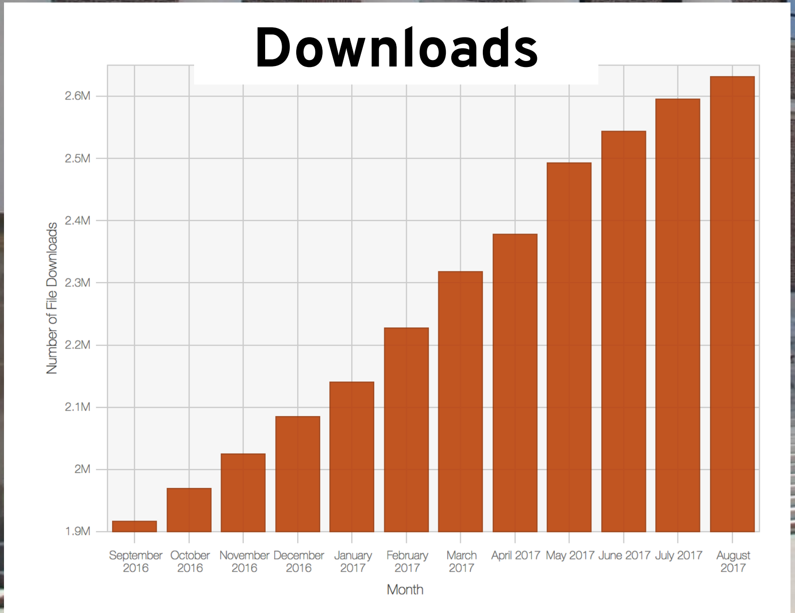
dataverse.org

HOW RESEARCHERS SHARE & USE DATA WITH DATAVERSE

Datasets Added



Downloads



Harvard Dataverse Repository

A public repository for research data

> 70,000 datasets total
> 49,000 datasets uploaded to Harvard Dataverse repository
200 datasets/month

> 340,000 files
4,000 files/month

> 2.5 M downloads
60,000 downloads/month

dataverse.harvard.edu

OUR CONTRIBUTIONS TO ENHANCE DATA SHARING

King, 1995, Replication, Replication

2014, Joint Declaration of Data Citation Principles

Wilkinson et al, 2016, The FAIR Guiding Principles for Scientific Data Management and Stewardship

Altman et al, 2001, A Digital Library for the Dissemination and Replication of Quantitative Social Science

Altman and King, 2007, A Proposed Standard for the Scholarly Citation of Quantitative Data

King, 2007, An Introduction to the Dataverse Network as an Infrastructure for Data Sharing

Crosas, 2012, The Dataverse Network: an open source application for sharing, discovering, and preserving research data

Crosas, 2013, A Data Sharing Story

Altman and Crosas, 2013, The Evolution to Data Citation: from principles to implementation

Pepe et al, 2014, How Do Astronomers Share Data?

Goodman et al, 2014, Ten Simple Rules for the Care and Feeding of Scientific Data

Crosas, Honaker, King, Sweeney, 2015, Automating Open Science for Big Data

Castro et al, 2015, Achieving Human and Machine Accessibility of Cited Data

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

Meyer et al. 2016, Data Publication with the Structural Biology Data Grid Supports Live Analysis

Bierer, Crosas, Pierce, 2017, Data Authorship as an Incentive to Data Sharing

2017

Data should be ...

FINDABLE

ACCESSIBLE

INTERPOPERABLE

REUSABLE

Wilkinson et al. , 2016, "The FAIR Guiding Principles for Scientific Data Management and Stewardship"

Nature Scientific Data

FAIR DATA IN DATAVERSE

Data Citation with Persistent Identifier

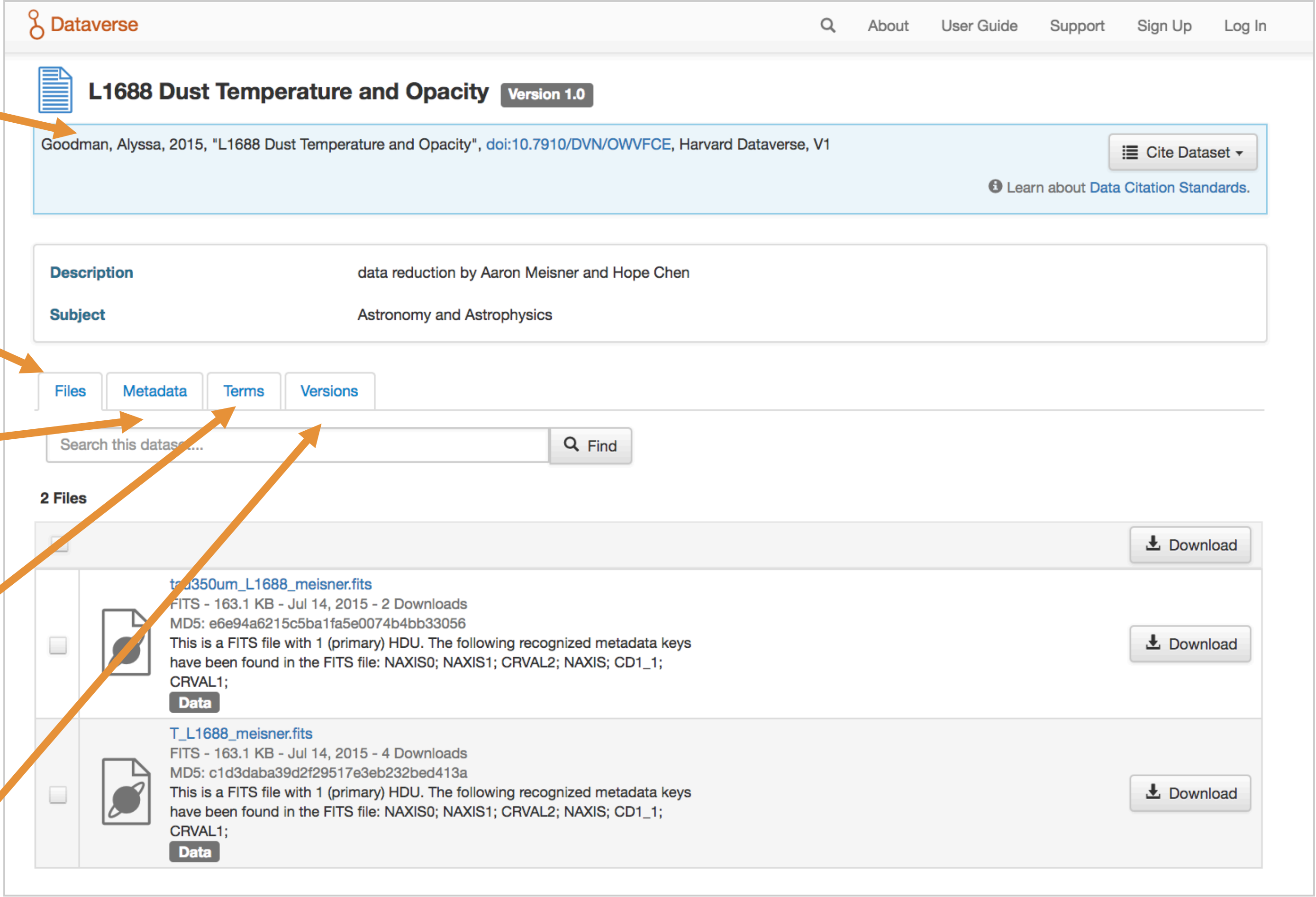
Data Files

Metadata

Data Licenses, User Agreements, Restrictions

Versions

APIs



L1688 Dust Temperature and Opacity

Goodman, Alyssa, 2015, "L1688 Dust Temperature and Opacity", doi:10.7910/DVN/OWVFCE, Harvard Dataverse, V1

Learn about Data Citation Standards.

Description	data reduction by Aaron Meisner and Hope Chen
Subject	Astronomy and Astrophysics

2 Files

tau350um_L1688_meisner.fits
FITS - 163.1 KB - Jul 14, 2015 - 2 Downloads
MD5: e6e94a6215c5ba1fa5e0074b4bb33056
This is a FITS file with 1 (primary) HDU. The following recognized metadata keys have been found in the FITS file: NAXIS0; NAXIS1; CRVAL2; NAXIS; CD1_1; CRVAL1;

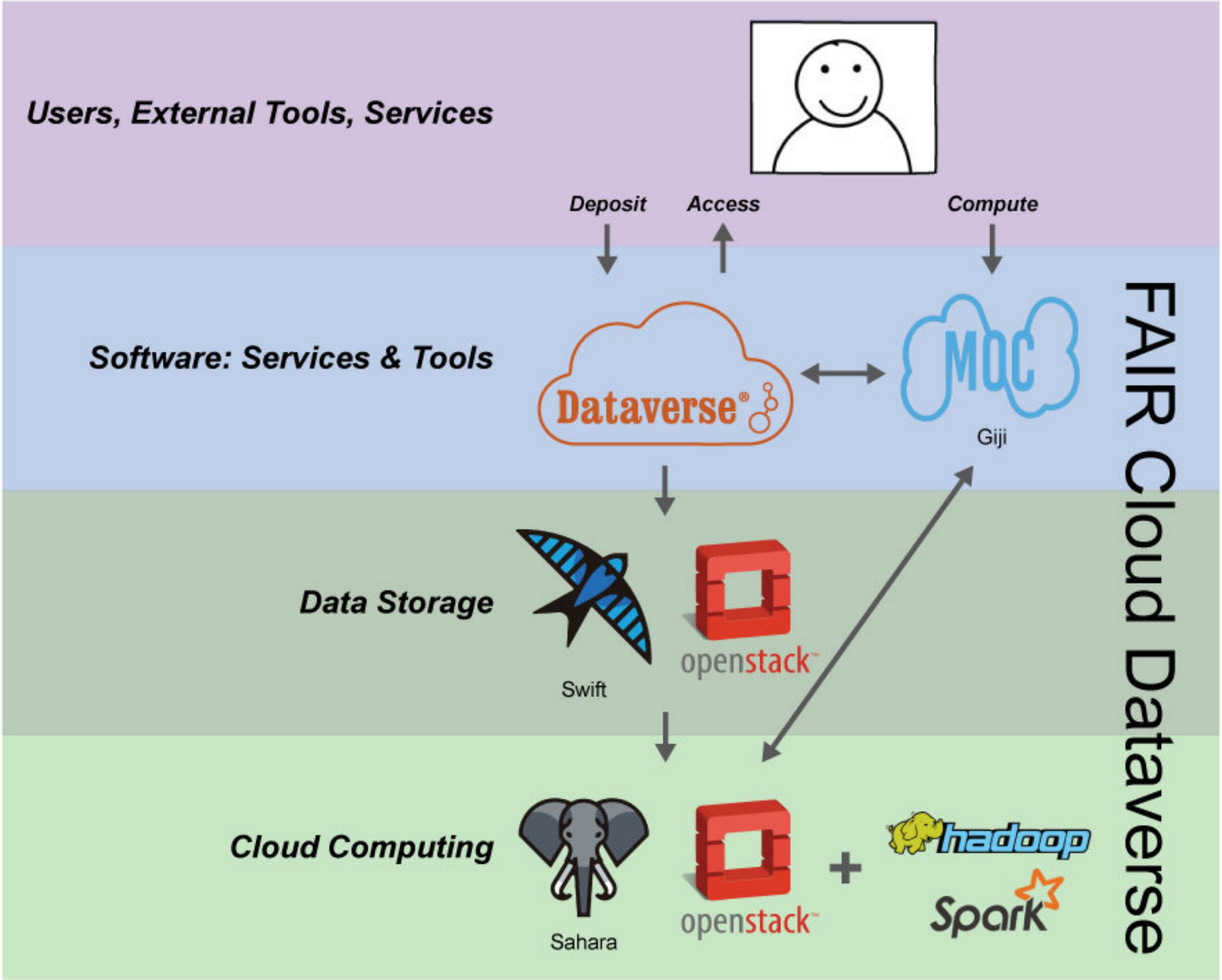
T_L1688_meisner.fits
FITS - 163.1 KB - Jul 14, 2015 - 4 Downloads
MD5: c1d3daba39d2f29517e3eb232bed413a
This is a FITS file with 1 (primary) HDU. The following recognized metadata keys have been found in the FITS file: NAXIS0; NAXIS1; CRVAL2; NAXIS; CD1_1; CRVAL1;

MOC + The Dataverse Project

Cloud Dataverse combines the power of cloud computing and storage with access to thousands of datasets from a feature-rich data repository platform

WHY CLOUD DATAVERSE?

- Big Data should also be **FAIR Data**
- Datasets are replicated to the Cloud for efficient access and reuse
- Computing on a dataset is enabled directly from any repository


Users, External Tools, Services
Deposit
Access
Compute
Software: Services & Tools
Dataverse®
MOC
Giji
Data Storage
Swift
openstack
Cloud Computing
Sahara
openstack
+
hadoop
Spark
FAIR Cloud Dataverse

WHAT WE HAVE BUILT

- Dataverse integration with Swift storage
- Compute access to MOC from a dataset page in Dataverse
- Temporary url to access restricted files in MOC

IN PROGRESS

- Implement Swift Access Control List (ACL) for file restriction
- Support InCommon for MOC to use same credentials as in Dataverse

NEXT

- Replicate data from any Dataverse to Cloud Dataverse
- Upload data directly in Swift; publish dataset from Swift to Dataverse

INTEGRATION WITH OTHER PROJECTS

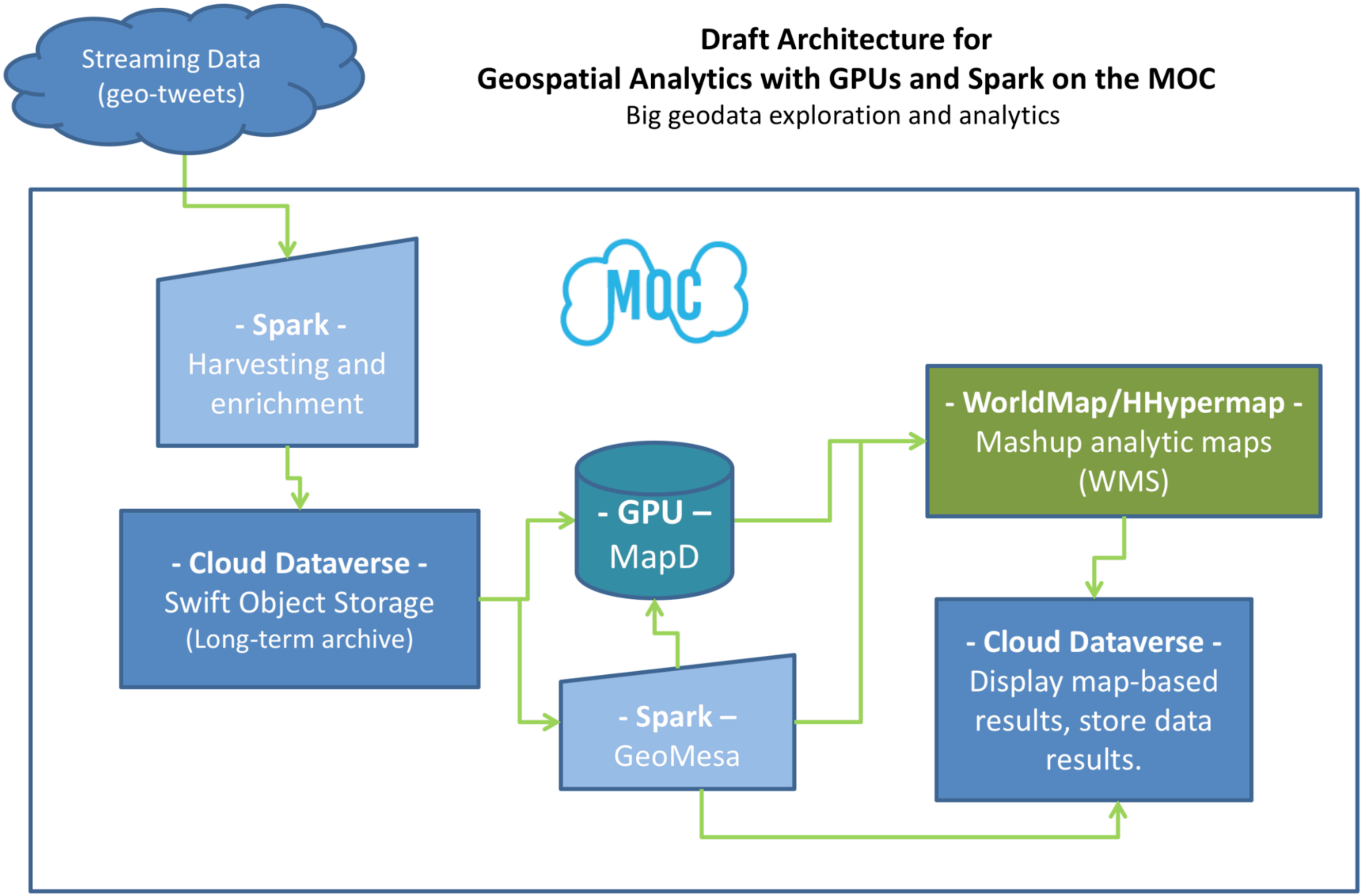
BILLION OBJECT PLATFORM

BIG GEODATA EXPLORATION AND ANALYTICS

ALFRED P. SLOAN FOUNDATION
1934

Draft Architecture for
Geospatial Analytics with GPUs and Spark on the MOC

Big geodata exploration and analytics



Streaming Data (geo-tweets)

MOC

- Spark -
Harvesting and enrichment

- Cloud Dataverse -
Swift Object Storage
(Long-term archive)

- GPU –
MapD

- Spark –
GeoMesa

- WorldMap/HHypermap -
Mashup analytic maps
(WMS)

- Cloud Dataverse -
Display map-based results, store data results.

The Dataverse Project

geomesa

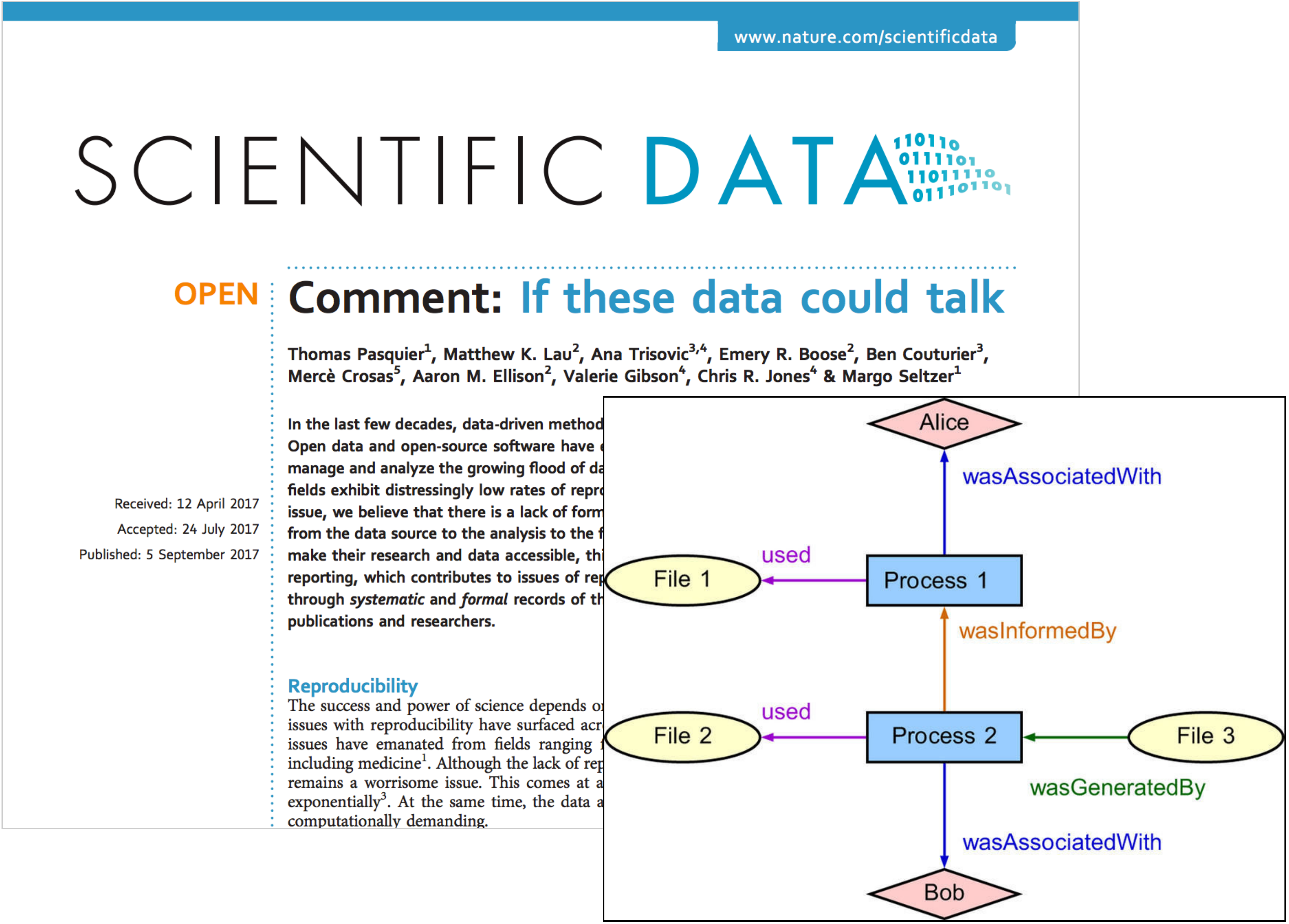
MAPD

WorldMap
HARVARD UNIVERSITY

STC
Spatiotemporal Thinking, Computing and Applications

DATA PROVENANCE

TRACK THE ORIGINAL SOURCE OF A DATASET



www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

Comment: If these data could talk

Thomas Pasquier[1], Matthew K. Lau[2], Ana Trisovic[3,4], Emery R. Boose[2], Ben Couturier[3], Mercè Crosas[5], Aaron M. Ellison[2], Valerie Gibson[4], Chris R. Jones[4] & Margo Seltzer[1]

Received: 12 April 2017
Accepted: 24 July 2017
Published: 5 September 2017

**In the last few decades, data-driven method
Open data and open-source software have
manage and analyze the growing flood of da
fields exhibit distressingly low rates of repr
issue, we believe that there is a lack of form
from the data source to the analysis to the f
make their research and data accessible, thi
reporting, which contributes to issues of rep
through *systematic* and *formal* records of th
publications and researchers.**

Reproducibility

The success and power of science depends o
issues with reproducibility have surfaced acr
issues have emanated from fields ranging
including medicine[1]. Although the lack of rep
remains a worrisome issue. This comes at a
exponentially[3]. At the same time, the data a
computationally demanding.

Pasquier, Lau, Trisovic, Boose, Coutierer, Crosas, Ellison, GIbson, Jones, Seltzer, 2017, *If These Data Could Talk*, Nature Scientific Data

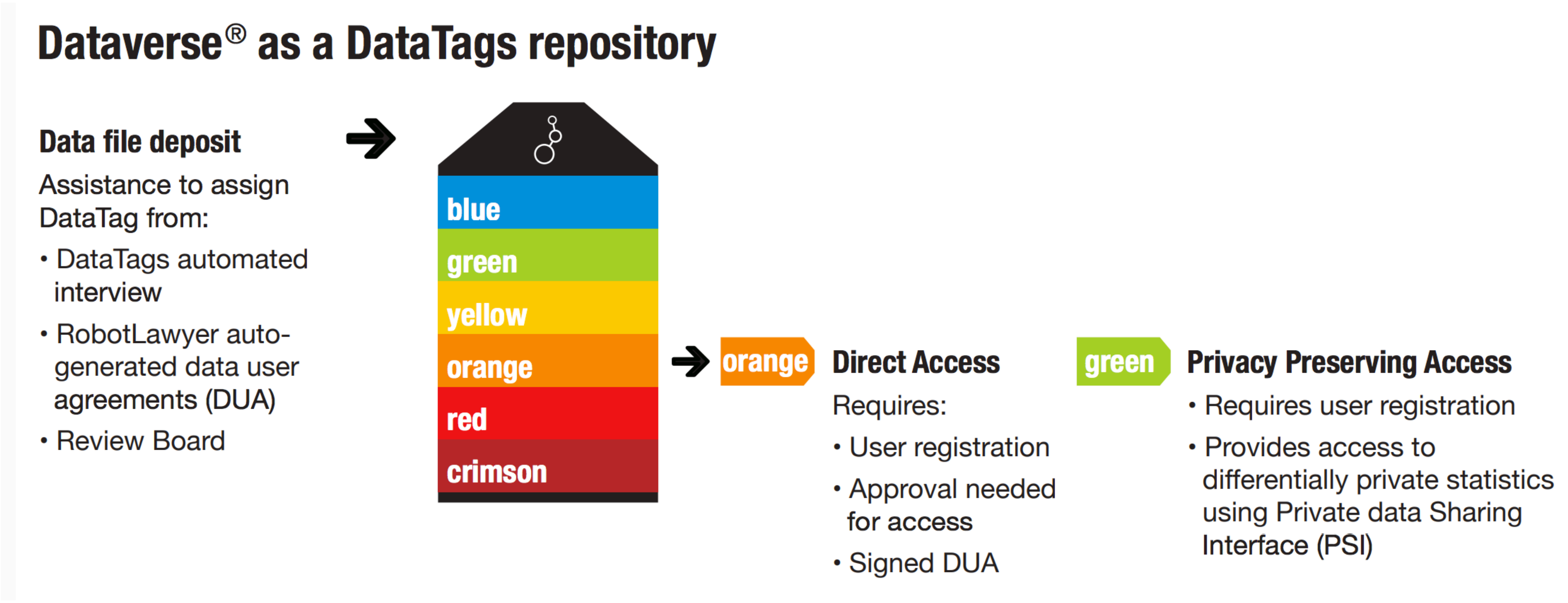
(Data Provenance examples from CERN and Harvard Forest)

DATA PRIVACY

CLASSIFY AND HANDLE DATASETS BASED ON THEIR PRIVACY LEVEL

NSF

ALFRED P. SLOAN FOUNDATION 1934



Dataverse® as a DataTags repository

Data file deposit

Assistance to assign DataTag from:

- DataTags automated interview
- RobotLawyer auto-generated data user agreements (DUA)
- Review Board

Direct Access

Requires:

- User registration
- Approval needed for access
- Signed DUA

Privacy Preserving Access

- Requires user registration
- Provides access to differentially private statistics using Private data Sharing Interface (PSI)

Harvard Data Privacy Tools Project: privacytools.seas.harvard.edu

DataTags Project: datatags.org

THANKS

@mercecrosas

@iqss

scholar.harvard.edu/mercecrosas

dataverse.org

Text