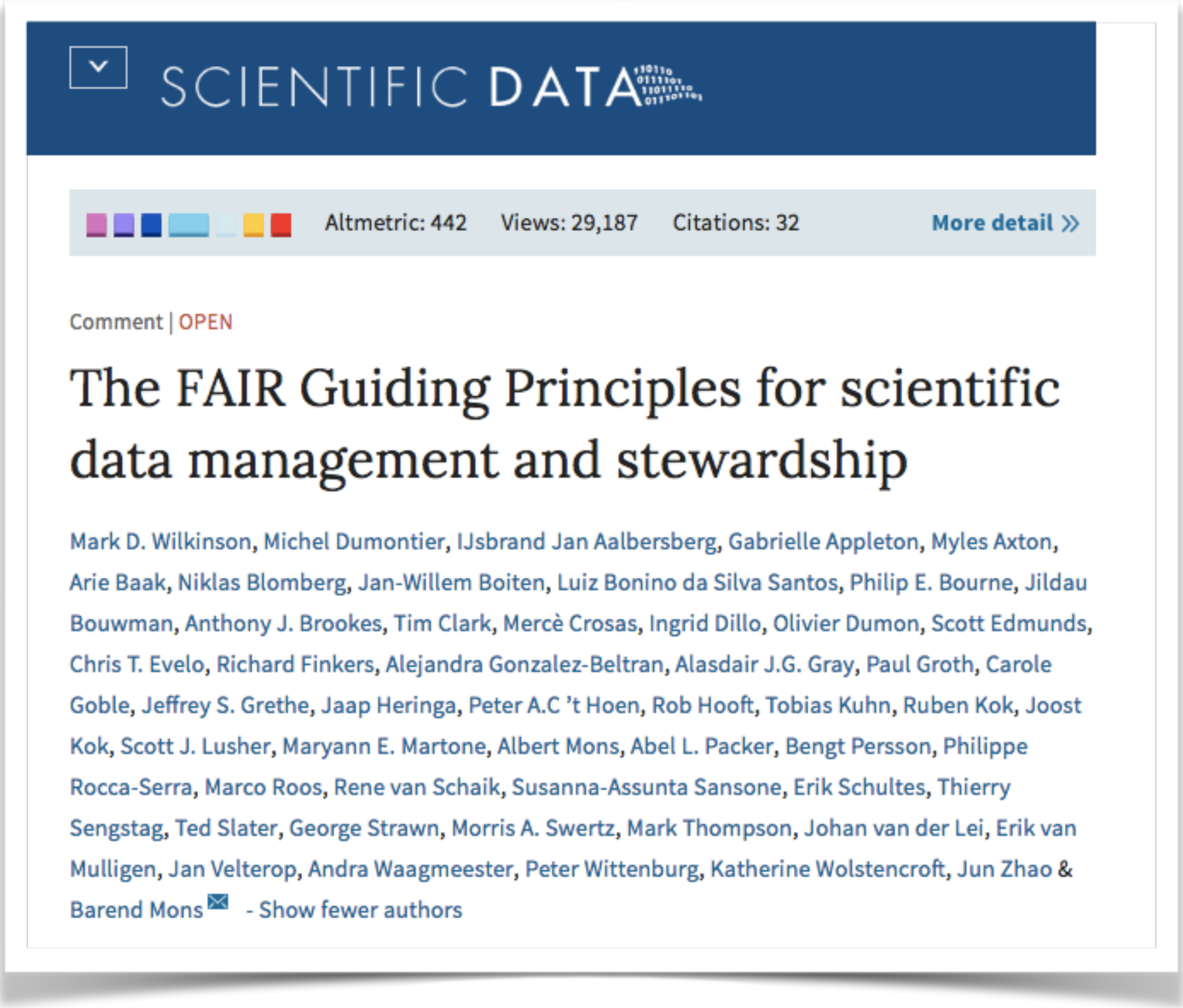
The Dataverse Project

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

Data should be Findable, Accessible, Interoperable, Reusable (FAIR) by machines



SCIENTIFIC DATA

Altmetric: 442 Views: 29,187 Citations: 32 More detail »

Comment | OPEN

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao & Barend Mons ✉ - Show fewer authors

Wilkinson et al, 'The FAIR Guiding Principles scientific data management and stewardship," Nature Scientific Data, 2016;
NIH Data Commons Principles; Joint Declaration of Data Citation Principles (Force11)

“**FAIR Principles** put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals.”

“**Good data management** is not a goal in itself, but rather is the key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community after the data publication process.”

FAIR Data Principles in Brief

- **To be Findable:**
  - (meta)data are assigned a globally unique and persistent identifier
  - data are described with rich metadata
  - metadata clearly and explicitly include the identifier of the data it describes
  - (meta)data are registered or indexed in a searchable resource
- **To be Accessible:**
  - (meta)data are retrievable by their identifier using a standardized communications protocol
  - the protocol is open, free, and universally implementable
  - the protocol allows for an authentication and authorization procedure, where necessary
  - metadata are accessible, even when the data are no longer available
- **To be Interoperable:**
  - (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
  - (meta)data use vocabularies that follow FAIR principles
  - (meta)data include qualified references to other (meta)data
- **To be Reusable:**
  - meta(data) are richly described with a plurality of accurate and relevant attributes
  - (meta)data are released with a clear and accessible data usage license
  - (meta)data are associated with detailed provenance
  - (meta)data meet domain-relevant community standards

We built Dataverse to incentivize data sharing, with “good data management” in mind

- An **open-source** platform to share and archive data
- Developed at **Harvard’s Institute for Quantitative Social Science** since 2006
- Gives **credit and control** to data authors & producers
- Builds a **community** to:
  - define new standards and best practices
  - foster new research in data sharing and reproducibility
- Has brought **data publishing** into the hands of data authors

The Dataverse Project

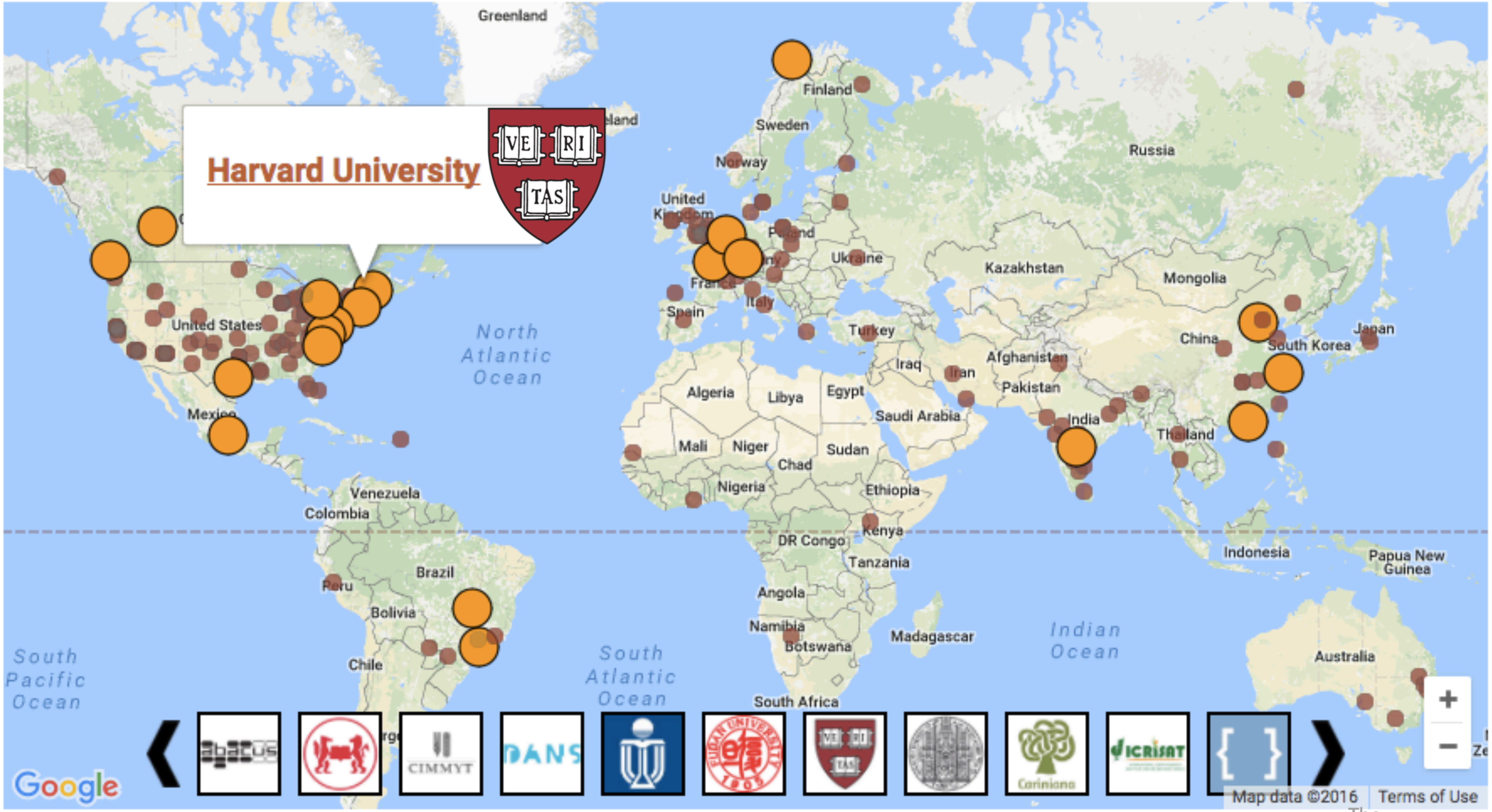
Dataverse is now a widely used repository platform

21 installations around the world

Used by researchers from > 500 institutions

60,000 datasets in Harvard Dataverse repository

http://dataverse.org



The Dataverse Project

Dataverse has a growing, engaged community of developers and users

38

GitHub contributors

332

members in the community list

23

community calls

so far with 239

participants from

8 countries

Annual Community Meeting,

with 200 attendees

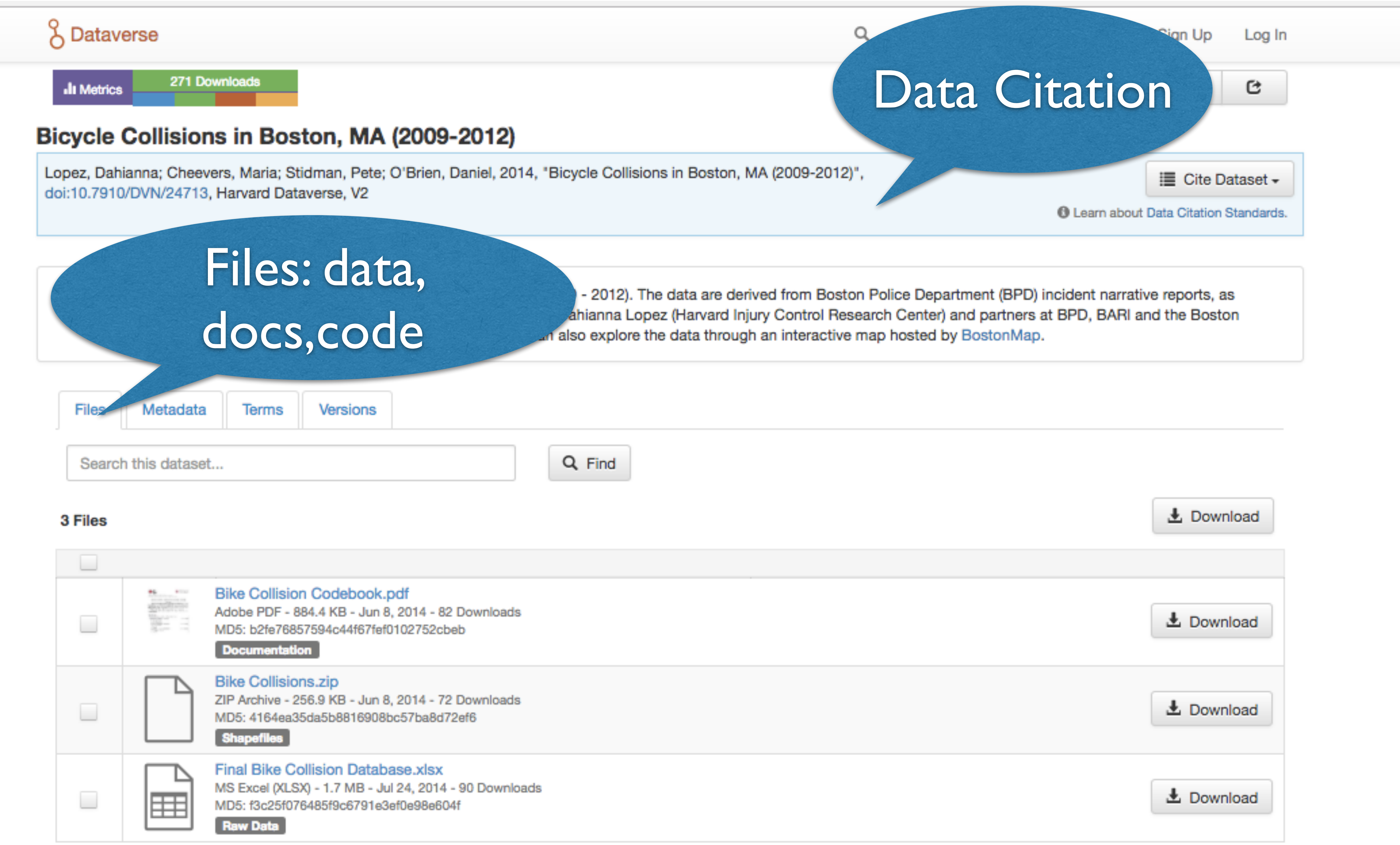
The Dataverse Project

Dataverse implements FAIR Data Principles

- **Data Citation with global persistent IDs:**
  - generate DOI automatically
  - attribution to data authors and repository
  - registration to DataCite
- **Rich Metadata:**
  - citation metadata
  - domain-specific descriptive metadata
  - variable and file metadata (extracted automatically)
- **Access and usage controls:**
  - open data as default, with CC0 waiver
  - custom terms of use and licenses, when needed
  - data can be restricted, but citation & metadata always publicly accessible
- **APIs and standards:**
  - SWORD, OAI-PMH, native API to search and get data and metadata
  - Dublin Core and DDI metadata standards
  - PROV ontology standard to capture provenance of a dataset (coming soon)

The Dataverse Project

Dataset Landing Page



Data Citation

Bicycle Collisions in Boston, MA (2009-2012)

Lopez, Dahianna; Cheevers, Maria; Stidman, Pete; O'Brien, Daniel, 2014, "Bicycle Collisions in Boston, MA (2009-2012)", doi:10.7910/DVN/24713, Harvard Dataverse, V2

Learn about Data Citation Standards.

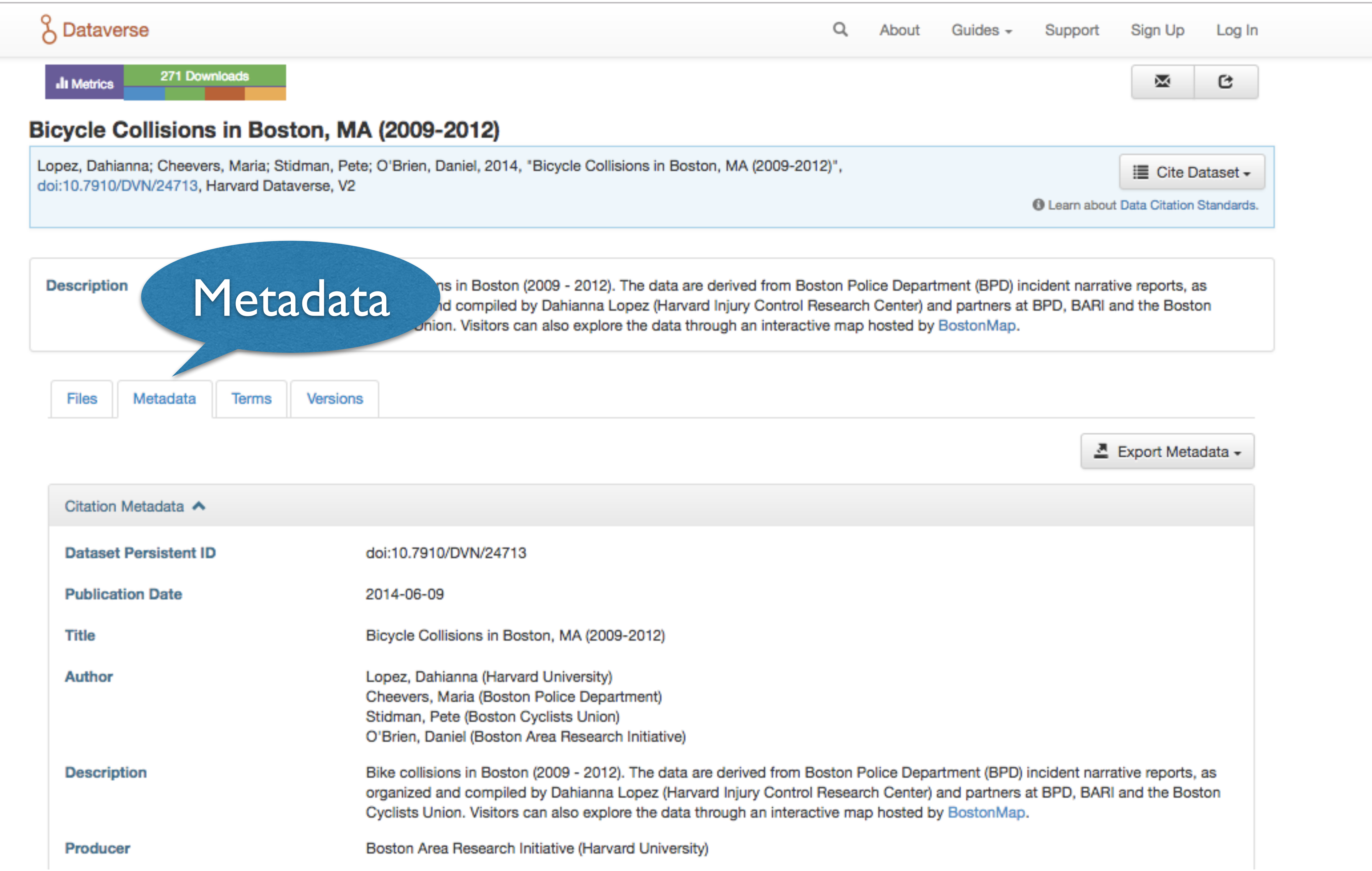
Files: data, docs,code

- 2012). The data are derived from Boston Police Department (BPD) incident narrative reports, as ahianna Lopez (Harvard Injury Control Research Center) and partners at BPD, BARI and the Boston n also explore the data through an interactive map hosted by BostonMap.

3 Files

□		
□	Bike Collision Codebook.pdf Adobe PDF - 884.4 KB - Jun 8, 2014 - 82 Downloads MD5: b2fe76857594c44f67fef0102752cbeb Documentation	Download
□	Bike Collisions.zip ZIP Archive - 256.9 KB - Jun 8, 2014 - 72 Downloads MD5: 4164ea35da5b8816908bc57ba8d72ef6 Shapefiles	Download
□	Final Bike Collision Database.xlsx MS Excel (XLSX) - 1.7 MB - Jul 24, 2014 - 90 Downloads MD5: f3c25f076485f9c6791e3ef0e98e604f Raw Data	Download

Dataset Landing Page



Metadata

Bicycle Collisions in Boston, MA (2009-2012)

Lopez, Dahianna; Cheevers, Maria; Stidman, Pete; O'Brien, Daniel, 2014, "Bicycle Collisions in Boston, MA (2009-2012)", doi:10.7910/DVN/24713, Harvard Dataverse, V2

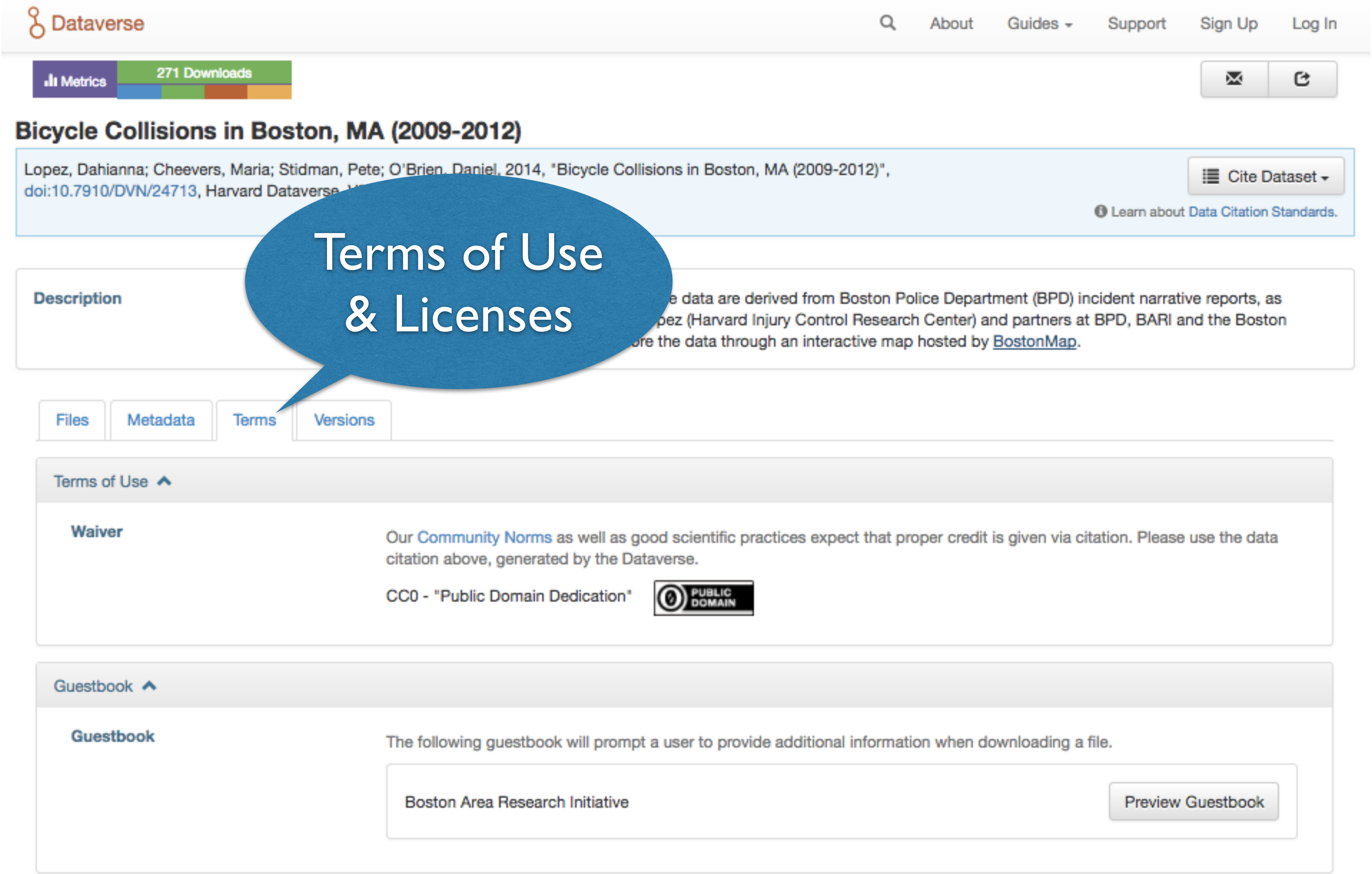
Learn about Data Citation Standards.

Description ns in Boston (2009 - 2012). The data are derived from Boston Police Department (BPD) incident narrative reports, as nd compiled by Dahianna Lopez (Harvard Injury Control Research Center) and partners at BPD, BARI and the Boston nion. Visitors can also explore the data through an interactive map hosted by BostonMap.

Citation Metadata

Dataset Persistent ID	doi:10.7910/DVN/24713
Publication Date	2014-06-09
Title	Bicycle Collisions in Boston, MA (2009-2012)
Author	Lopez, Dahianna (Harvard University) Cheevers, Maria (Boston Police Department) Stidman, Pete (Boston Cyclists Union) O'Brien, Daniel (Boston Area Research Initiative)
Description	Bike collisions in Boston (2009 - 2012). The data are derived from Boston Police Department (BPD) incident narrative reports, as organized and compiled by Dahianna Lopez (Harvard Injury Control Research Center) and partners at BPD, BARI and the Boston Cyclists Union. Visitors can also explore the data through an interactive map hosted by BostonMap.
Producer	Boston Area Research Initiative (Harvard University)

Dataset Landing Page



Bicycle Collisions in Boston, MA (2009-2012)

Lopez, Dahianna; Cheevers, Maria; Stidman, Pete; O'Brien, Daniel, 2014, "Bicycle Collisions in Boston, MA (2009-2012)", doi:10.7910/DVN/24713, Harvard Dataverse

Description

...e data are derived from Boston Police Department (BPD) incident narrative reports, as ...pez (Harvard Injury Control Research Center) and partners at BPD, BARI and the Boston ...re the data through an interactive map hosted by BostonMap.

Terms of Use
& Licenses

Terms of Use

Waiver

Our Community Norms as well as good scientific practices expect that proper credit is given via citation. Please use the data citation above, generated by the Dataverse.

CC0 - "Public Domain Dedication"

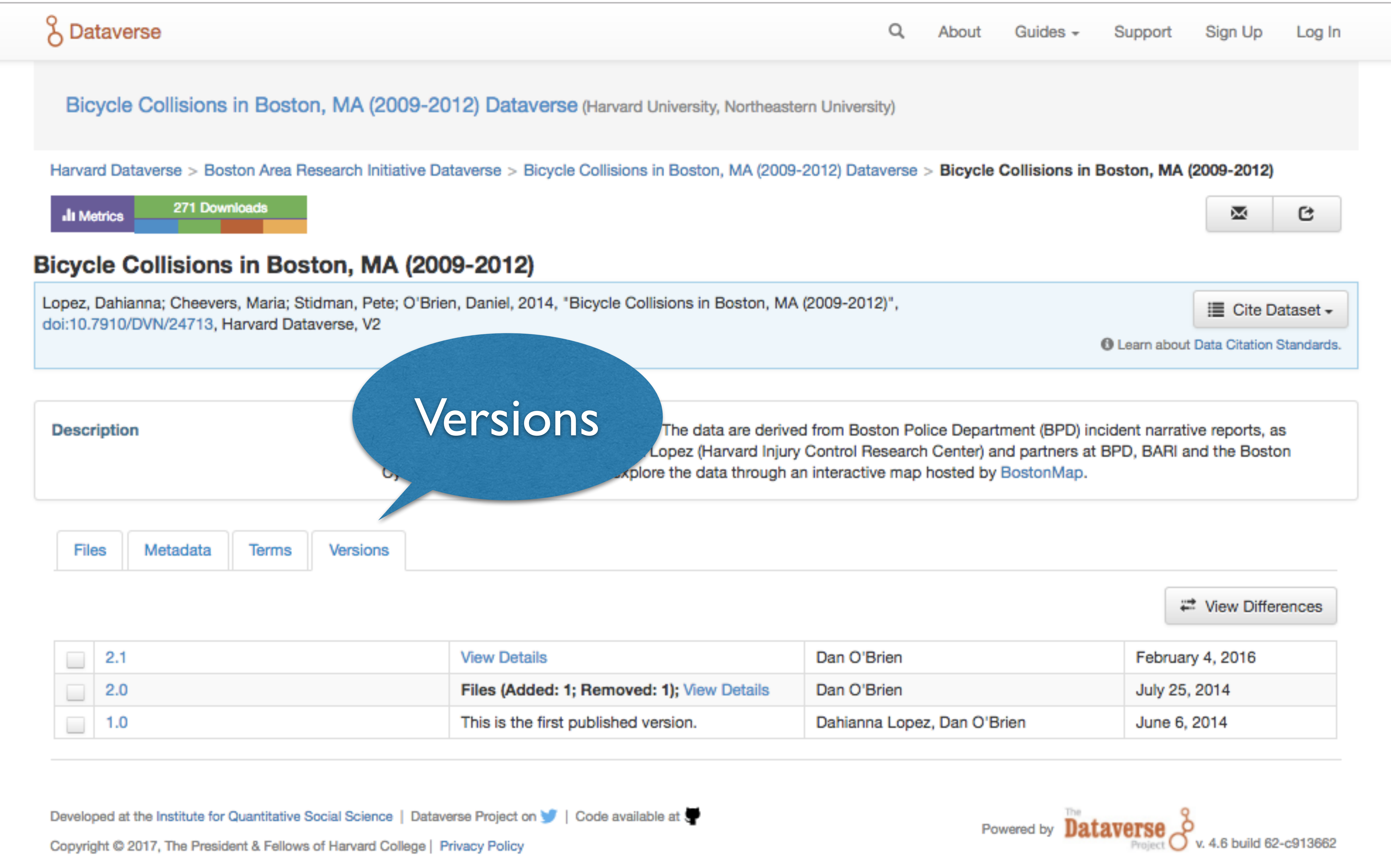
Guestbook

Guestbook

The following guestbook will prompt a user to provide additional information when downloading a file.

Boston Area Research Initiative

Dataset Landing Page



Bicycle Collisions in Boston, MA (2009-2012) Dataverse (Harvard University, Northeastern University)

Harvard Dataverse > Boston Area Research Initiative Dataverse > Bicycle Collisions in Boston, MA (2009-2012) Dataverse > Bicycle Collisions in Boston, MA (2009-2012)

Bicycle Collisions in Boston, MA (2009-2012)

Lopez, Dahianna; Cheevers, Maria; Stidman, Pete; O'Brien, Daniel, 2014, "Bicycle Collisions in Boston, MA (2009-2012)", doi:10.7910/DVN/24713, Harvard Dataverse, V2

Learn about Data Citation Standards.

Versions

Description The data are derived from Boston Police Department (BPD) incident narrative reports, as Lopez (Harvard Injury Control Research Center) and partners at BPD, BARI and the Boston Cy xplore the data through an interactive map hosted by BostonMap.

□	2.1	View Details	Dan O'Brien	February 4, 2016
□	2.0	**Files (Added: 1; Removed: 1);** View Details	Dan O'Brien	July 25, 2014
□	1.0	This is the first published version.	Dahianna Lopez, Dan O'Brien	June 6, 2014

Developed at the Institute for Quantitative Social Science | Dataverse Project on | Code available at

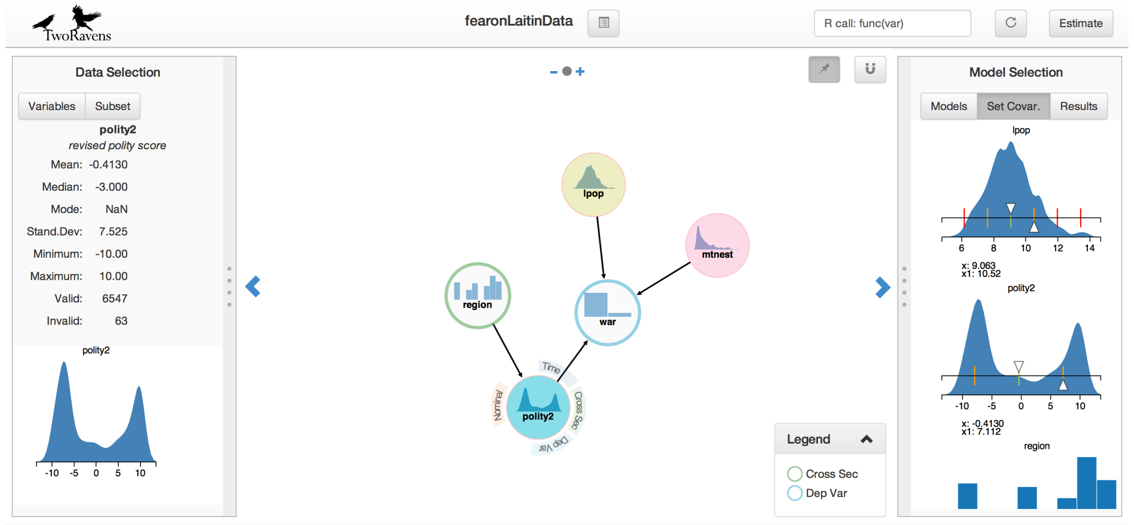
Copyright © 2017, The President & Fellows of Harvard College | Privacy Policy

Powered by The Dataverse Project v. 4.6 build 62-c913662

Standard file formats and automatic metadata extraction allow data exploration

Var1	Var2	Var3	Var4

TwoRavens: summary stats & analysis



Data Selection

polity2

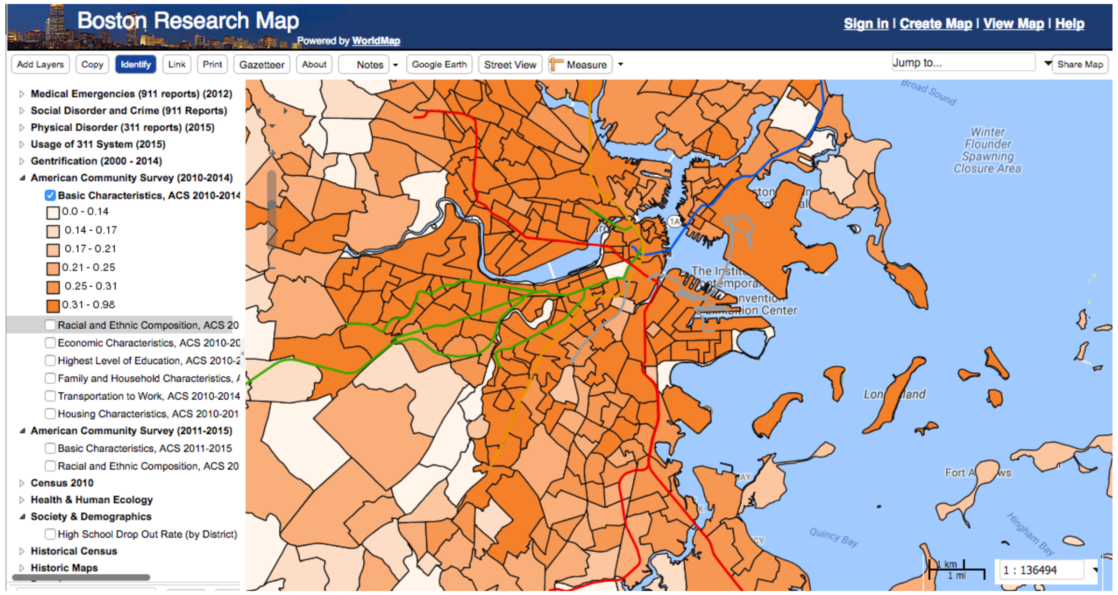
revised polity score

Mean:	-0.4130
Median:	-3.000
Mode:	NaN
Stand.Dev:	7.525
Minimum:	-10.00
Maximum:	10.00
Valid:	6547
Invalid:	63

geospatial variable

Var1	Var2	Var3	Var4

WorldMap: geospatial exploration



Boston Research Map

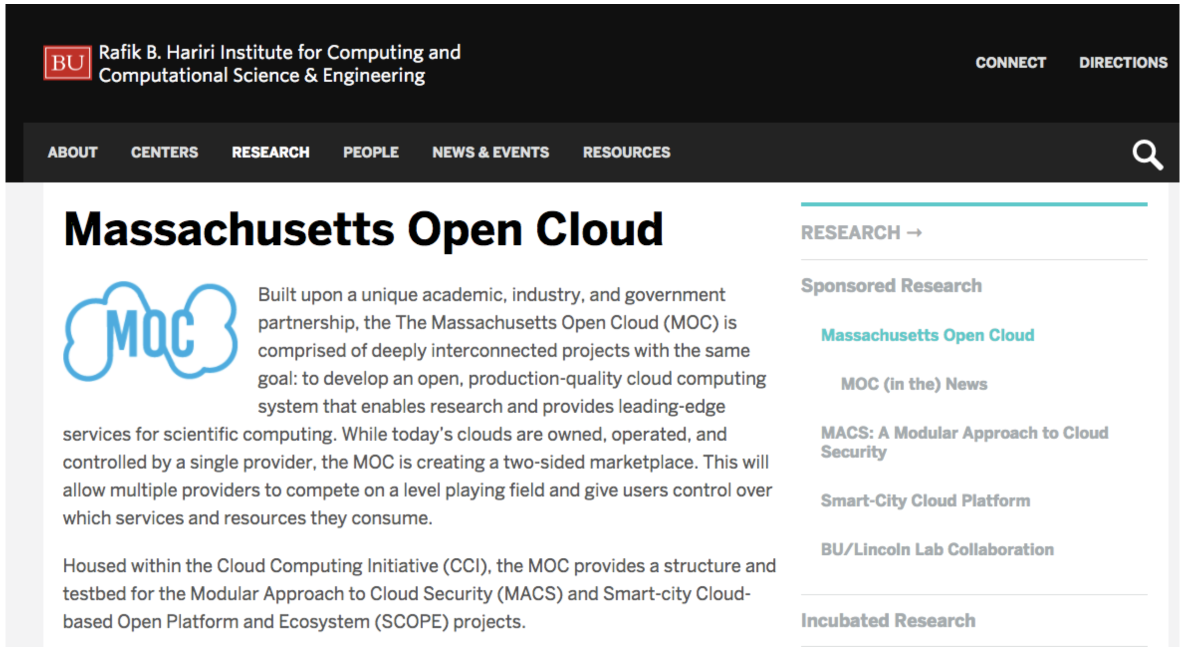
- Medical Emergencies (911 reports) (2012)
- Social Disorder and Crime (911 Reports)
- Physical Disorder (311 reports) (2015)
- Usage of 311 System (2015)
- Gentrification (2000 - 2014)
- American Community Survey (2010-2014)
  - ☑ Basic Characteristics, ACS 2010-2014
    - □ 0.0 - 0.14
    - □ 0.14 - 0.17
    - □ 0.17 - 0.21
    - □ 0.21 - 0.25
    - □ 0.25 - 0.31
    - □ 0.31 - 0.98
  - □ Racial and Ethnic Composition, ACS 20
  - □ Economic Characteristics, ACS 2010-20
  - □ Highest Level of Education, ACS 2010-2
  - □ Family and Household Characteristics, /
  - □ Transportation to Work, ACS 2010-2014
  - □ Housing Characteristics, ACS 2010-201
- American Community Survey (2011-2015)
  - □ Basic Characteristics, ACS 2011-2015
  - □ Racial and Ethnic Composition, ACS 20
- Census 2010
- Health & Human Ecology
- Society & Demographics
  - □ High School Drop Out Rate (by District)
- Historical Census
- Historic Maps

The Dataverse Project

Cloud Dataverse:

- is a collaboration between **Massachusetts Open Cloud (MOC)** and **Dataverse**
- will allow replication of data in multiple storage locations, and access to cloud computing

Led by Boston University, the MOC is a collaborative effort among BU, Harvard, UMass Amherst, MIT, and Northeastern University, as well as the Massachusetts Green High-Performance Computing Center (MGHPCC) and Oak Ridge National Laboratory (ORNL)


BU Rafik B. Hariri Institute for Computing and Computational Science & Engineering
CONNECT
DIRECTIONS
ABOUT
CENTERS
RESEARCH
PEOPLE
NEWS & EVENTS
RESOURCES
MOC

Massachusetts Open Cloud

Built upon a unique academic, industry, and government partnership, the The Massachusetts Open Cloud (MOC) is comprised of deeply interconnected projects with the same goal: to develop an open, production-quality cloud computing system that enables research and provides leading-edge services for scientific computing. While today's clouds are owned, operated, and controlled by a single provider, the MOC is creating a two-sided marketplace. This will allow multiple providers to compete on a level playing field and give users control over which services and resources they consume.

Housed within the Cloud Computing Initiative (CCI), the MOC provides a structure and testbed for the Modular Approach to Cloud Security (MACS) and Smart-city Cloud-based Open Platform and Ecosystem (SCOPE) projects.

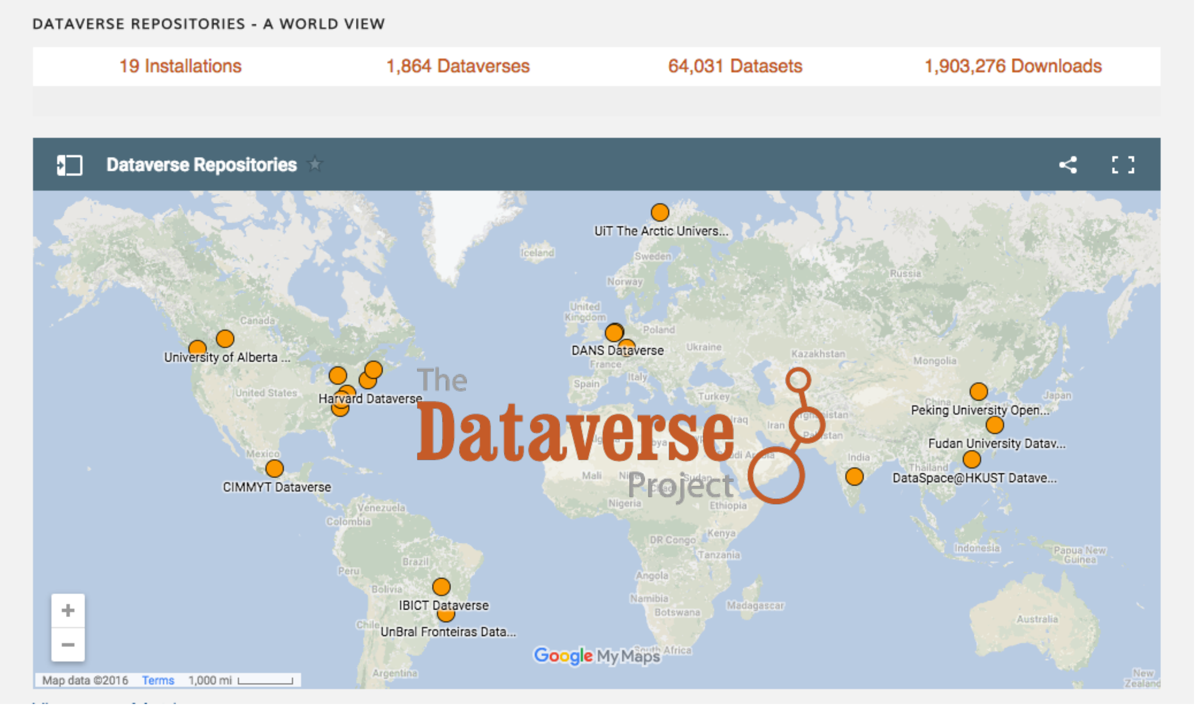
RESEARCH →

Sponsored Research

- Massachusetts Open Cloud
  - MOC (in the) News
- MACS: A Modular Approach to Cloud Security
- Smart-City Cloud Platform
- BU/Lincoln Lab Collaboration

Incubated Research

Dataverse led by Harvard's IQSS

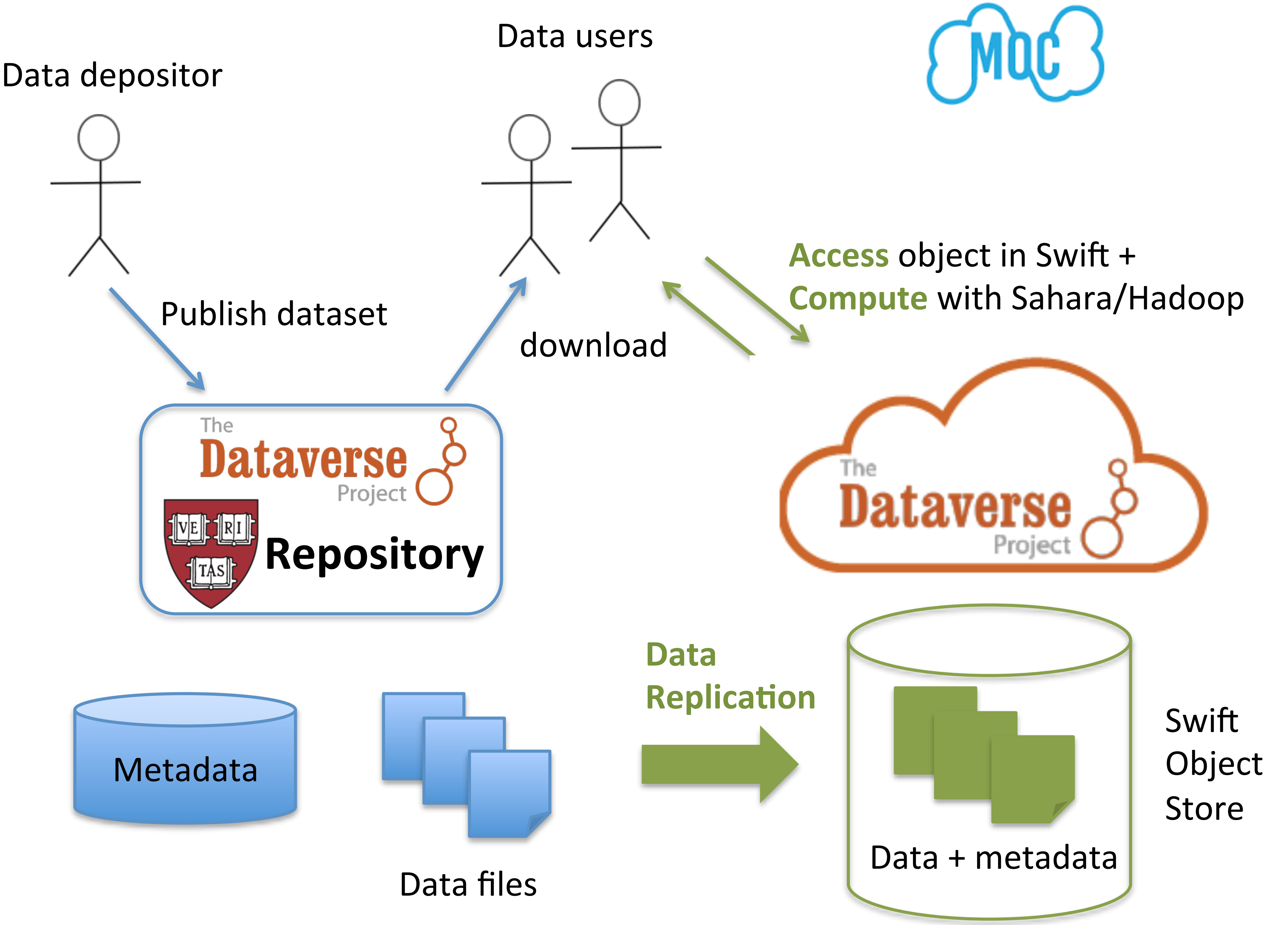


DATAVERSE REPOSITORIES - A WORLD VIEW

19 Installations	1,864 Dataverses	64,031 Datasets	1,903,276 Downloads

Dataverse Repositories
UiT The Arctic Univers...
University of Alberta ...
Harvard Dataverse
DANS Dataverse
CIMMYT Dataverse
Peking University Open...
Fudan University Datav...
DataSpace@HKUST Datave...
IBICT Dataverse
UnBral Fronteiras Data...
The Dataverse Project
Google My Maps
Map data ©2016 Terms 1,000 mi

Dataverse Now with Cloud Dataverse



Data depositor

Data users

Publish dataset

download

Access object in Swift + Compute with Sahara/Hadoop

The Dataverse Project Repository

The Dataverse Project

Metadata

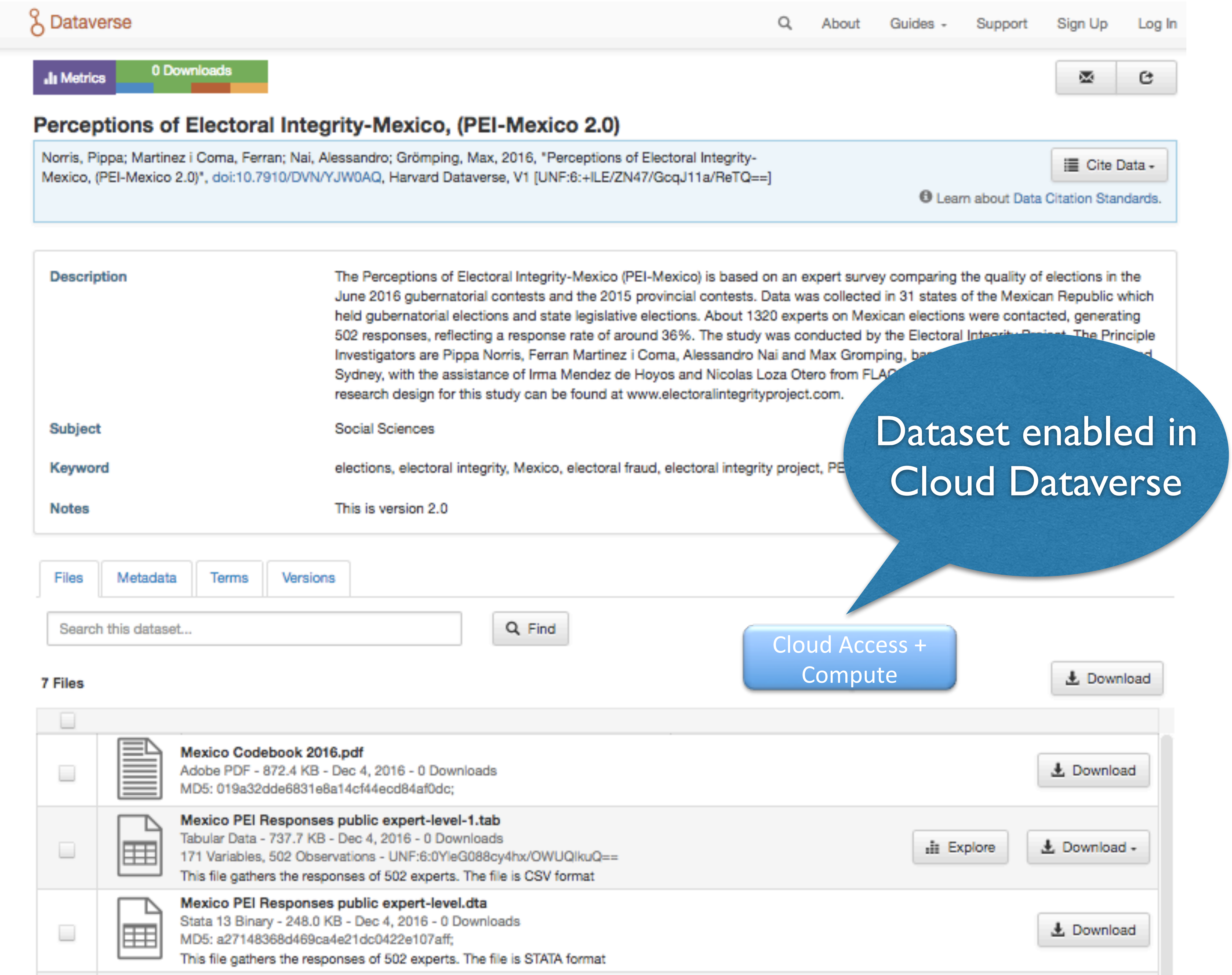
Data files

Data Replication

Data + metadata

Swift Object Store

Future Dataset Landing Page: Access to Compute



Dataverse

Perceptions of Electoral Integrity-Mexico, (PEI-Mexico 2.0)

Norris, Pippa; Martinez i Coma, Ferran; Nai, Alessandro; Grömping, Max, 2016, "Perceptions of Electoral Integrity-Mexico, (PEI-Mexico 2.0)", doi:10.7910/DVN/YJW0AQ, Harvard Dataverse, V1 [UNF:6:+ILE/ZN47/GcqJ11a/ReTQ==]

Learn about Data Citation Standards.

Description The Perceptions of Electoral Integrity-Mexico (PEI-Mexico) is based on an expert survey comparing the quality of elections in the June 2016 gubernatorial contests and the 2015 provincial contests. Data was collected in 31 states of the Mexican Republic which held gubernatorial elections and state legislative elections. About 1320 experts on Mexican elections were contacted, generating 502 responses, reflecting a response rate of around 36%. The study was conducted by the Electoral Integ The Principle Investigators are Pippa Norris, Ferran Martinez i Coma, Alessandro Nai and Max Gromping, ba Sydney, with the assistance of Irma Mendez de Hoyos and Nicolas Loza Otero from FLAC research design for this study can be found at www.electoralintegrityproject.com.

Subject Social Sciences

Keyword elections, electoral integrity, Mexico, electoral fraud, electoral integrity project, PE

Notes This is version 2.0

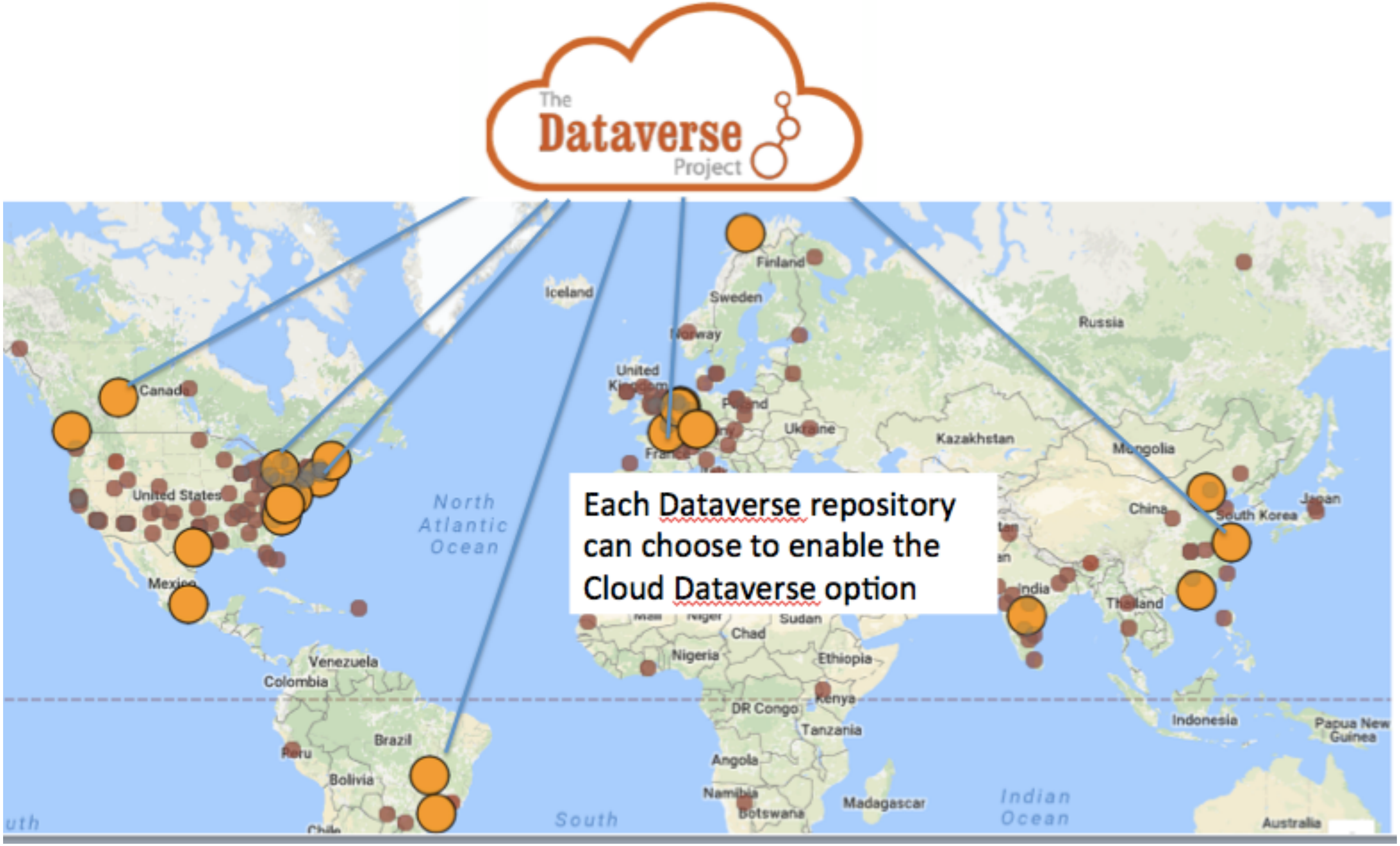
Dataset enabled in Cloud Dataverse

7 Files

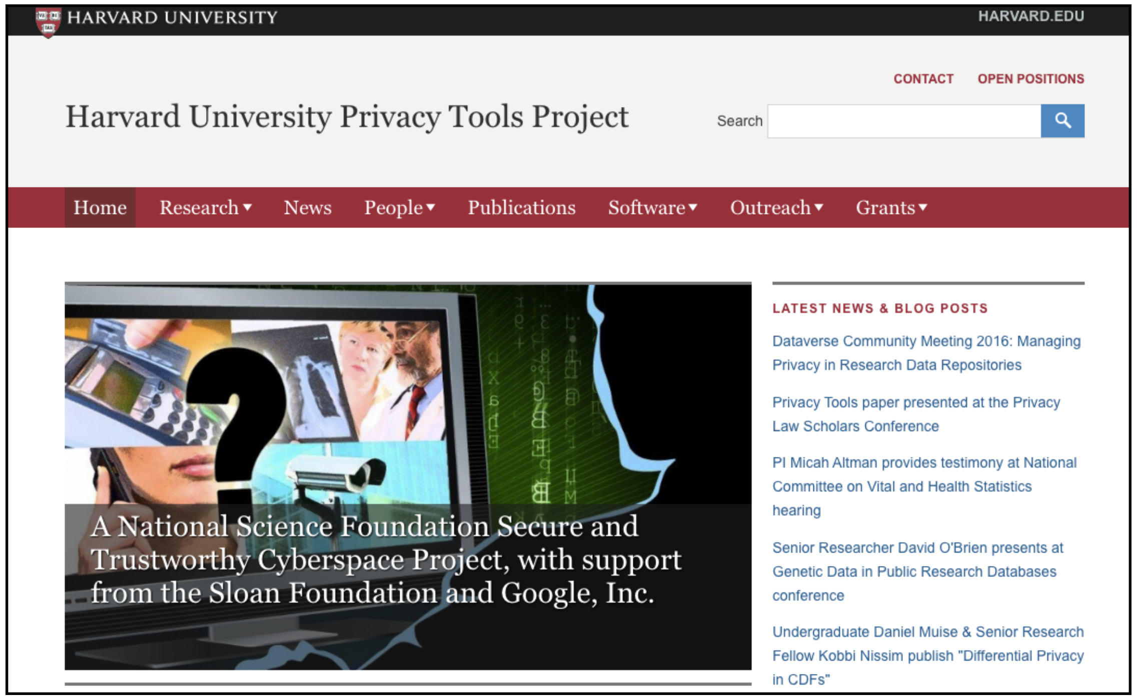
Mexico Codebook 2016.pdf
Adobe PDF - 872.4 KB - Dec 4, 2016 - 0 Downloads
MD5: 019a32dde6831e8a14cf44ecd84af0dc;

Mexico PEI Responses public expert-level-1.tab
Tabular Data - 737.7 KB - Dec 4, 2016 - 0 Downloads
171 Variables, 502 Observations - UNF:6:0YleG088cy4hx/OWUQlkuQ==
This file gathers the responses of 502 experts. The file is CSV format

Mexico PEI Responses public expert-level.dta
Stata 13 Binary - 248.0 KB - Dec 4, 2016 - 0 Downloads
MD5: a27148368d469ca4e21dc0422e107aff;
This file gathers the responses of 502 experts. The file is STATA format



Each Dataverse repository can choose to enable the Cloud Dataverse option http://privacytools.seas.harvard.edu



HARVARD UNIVERSITY HARVARD.EDU

CONTACT OPEN POSITIONS

Harvard University Privacy Tools Project

Search

Home Research ▾ News People ▾ Publications Software ▾ Outreach ▾ Grants ▾

A National Science Foundation Secure and Trustworthy Cyberspace Project, with support from the Sloan Foundation and Google, Inc.

LATEST NEWS & BLOG POSTS

Dataverse Community Meeting 2016: Managing Privacy in Research Data Repositories

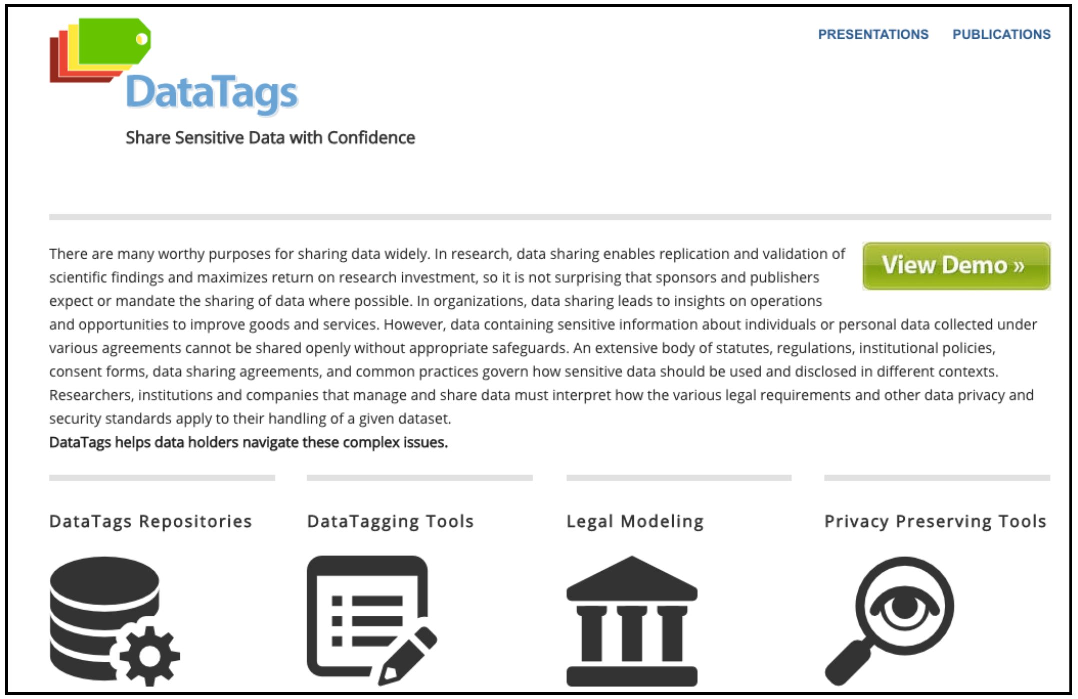
Privacy Tools paper presented at the Privacy Law Scholars Conference

PI Micah Altman provides testimony at National Committee on Vital and Health Statistics hearing

Senior Researcher David O'Brien presents at Genetic Data in Public Research Databases conference

Undergraduate Daniel Muise & Senior Research Fellow Kobbi Nissim publish "Differential Privacy in CDFs"

http://datatags.org



PRESENTATIONS PUBLICATIONS

DataTags

Share Sensitive Data with Confidence

There are many worthy purposes for sharing data widely. In research, data sharing enables replication and validation of scientific findings and maximizes return on research investment, so it is not surprising that sponsors and publishers expect or mandate the sharing of data where possible. In organizations, data sharing leads to insights on operations and opportunities to improve goods and services. However, data containing sensitive information about individuals or personal data collected under various agreements cannot be shared openly without appropriate safeguards. An extensive body of statutes, regulations, institutional policies, consent forms, data sharing agreements, and common practices govern how sensitive data should be used and disclosed in different contexts. Researchers, institutions and companies that manage and share data must interpret how the various legal requirements and other data privacy and security standards apply to their handling of a given dataset.

DataTags helps data holders navigate these complex issues.

View Demo »

DataTags Repositories DataTagging Tools Legal Modeling Privacy Preserving Tools

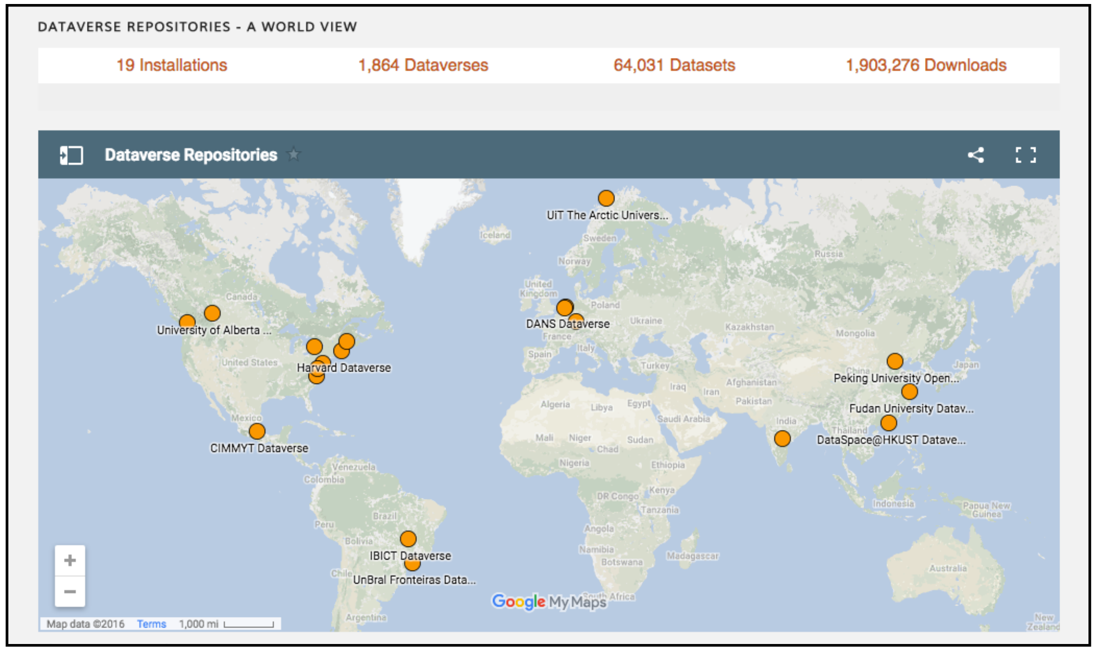
DataTags:

- is part of the Harvard University Privacy Tools Project
- will be integrated with Dataverse

http://dataverse.org

The Dataverse Project

Open source research data repository software



DATAVERSE REPOSITORIES - A WORLD VIEW

19 Installations 1,864 Dataverses 64,031 Datasets 1,903,276 Downloads

Dataverse Repositories

UiT The Arctic Univers...
University of Alberta ...
Harvard Dataverse
DANS Dataverse
Peking University Open...
Fudan University Datav...
DataSpace@HKUST Datave...
CIMMYT Dataverse
IBICT Dataverse
UnBral Fronteiras Data...
Google My Maps
Map data ©2016 Terms 1,000 mi

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

DataTags Levels

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

DataTags and their respective policies

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

Requirements for a DataTags Repository

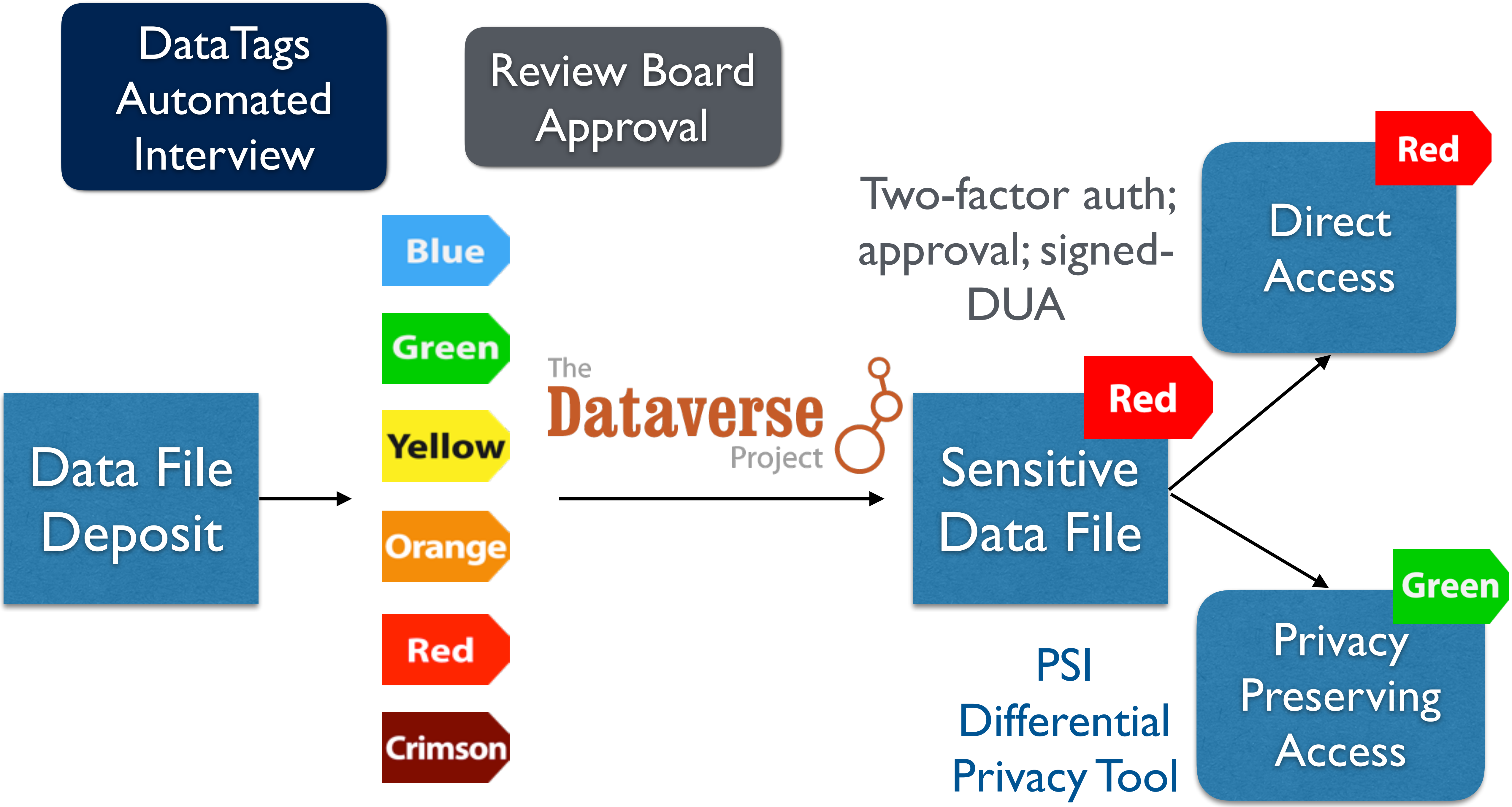
1. Supports more than one datatag
2. Each file in the repository must have one and only one datatag
   - additional requirements cannot weaken the file security
   - and cannot required the same or more security than a more restrictive datatag
3. A recipient of a file from the repository must
   - satisfy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

Repositories today do not have an easy, standard way to support sensitive data

The Dataverse Project Terms of Use

“User Uploads must be void of all identifiable information, such that re-identification of any subjects from the amalgamation of the information available from all of the materials (across datasets and dataverses) uploaded under any one author and/or user should not be possible.”

We are making Dataverse a DataTags Repository to share sensitive data



DataTags Automated Interview

Review Board Approval

Blue

Green

Yellow

Orange

Red

Crimson

Data File Deposit

The Dataverse Project

Sensitive Data File

Red

Two-factor auth; approval; signed-DUA

Direct Access

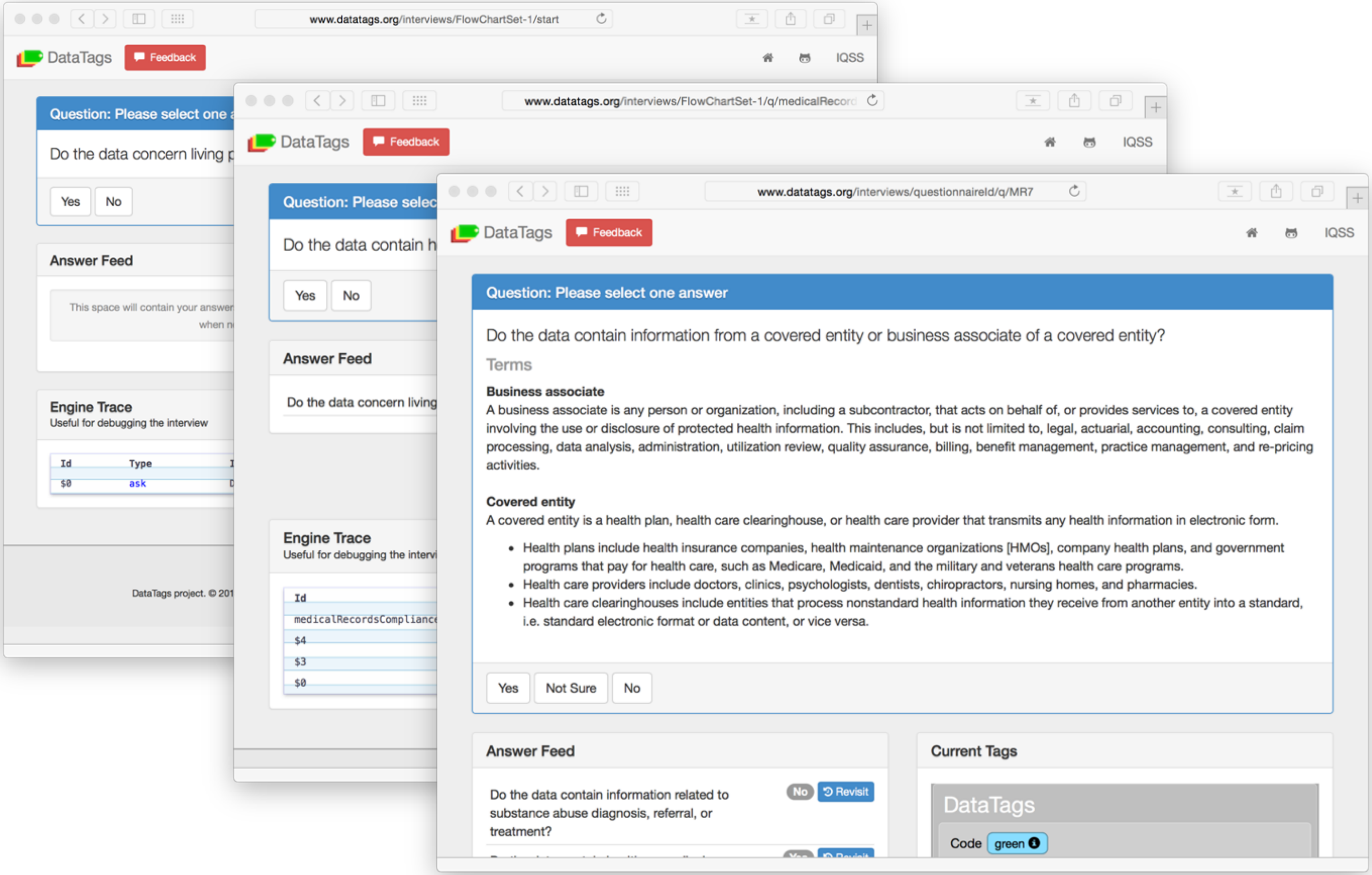
Red

PSI Differential Privacy Tool

Privacy Preserving Access

Green

The DataTags automated interview …



www.datatags.org/interviews/FlowChartSet-1/start

Question: Please select one a

Do the data concern living p

Answer Feed

This space will contain your answer when n

Engine Trace

Useful for debugging the interview

```
Id      Type
$0      ask
```

DataTags project. © 201

www.datatags.org/interviews/FlowChartSet-1/q/medicalRecord

Question: Please selec

Do the data contain h

Answer Feed

Do the data concern living

Engine Trace

Useful for debugging the intervi

```
Id
medicalRecordsCompliance
$4
$3
$0
```

www.datatags.org/interviews/questionnaireId/q/MR7

Question: Please select one answer

Do the data contain information from a covered entity or business associate of a covered entity?

Terms

Business associate

A business associate is any person or organization, including a subcontractor, that acts on behalf of, or provides services to, a covered entity involving the use or disclosure of protected health information. This includes, but is not limited to, legal, actuarial, accounting, consulting, claim processing, data analysis, administration, utilization review, quality assurance, billing, benefit management, practice management, and re-pricing activities.

Covered entity

A covered entity is a health plan, health care clearinghouse, or health care provider that transmits any health information in electronic form.

- Health plans include health insurance companies, health maintenance organizations [HMOs], company health plans, and government programs that pay for health care, such as Medicare, Medicaid, and the military and veterans health care programs.
- Health care providers include doctors, clinics, psychologists, dentists, chiropractors, nursing homes, and pharmacies.
- Health care clearinghouses include entities that process nonstandard health information they receive from another entity into a standard, i.e. standard electronic format or data content, or vice versa.

Answer Feed

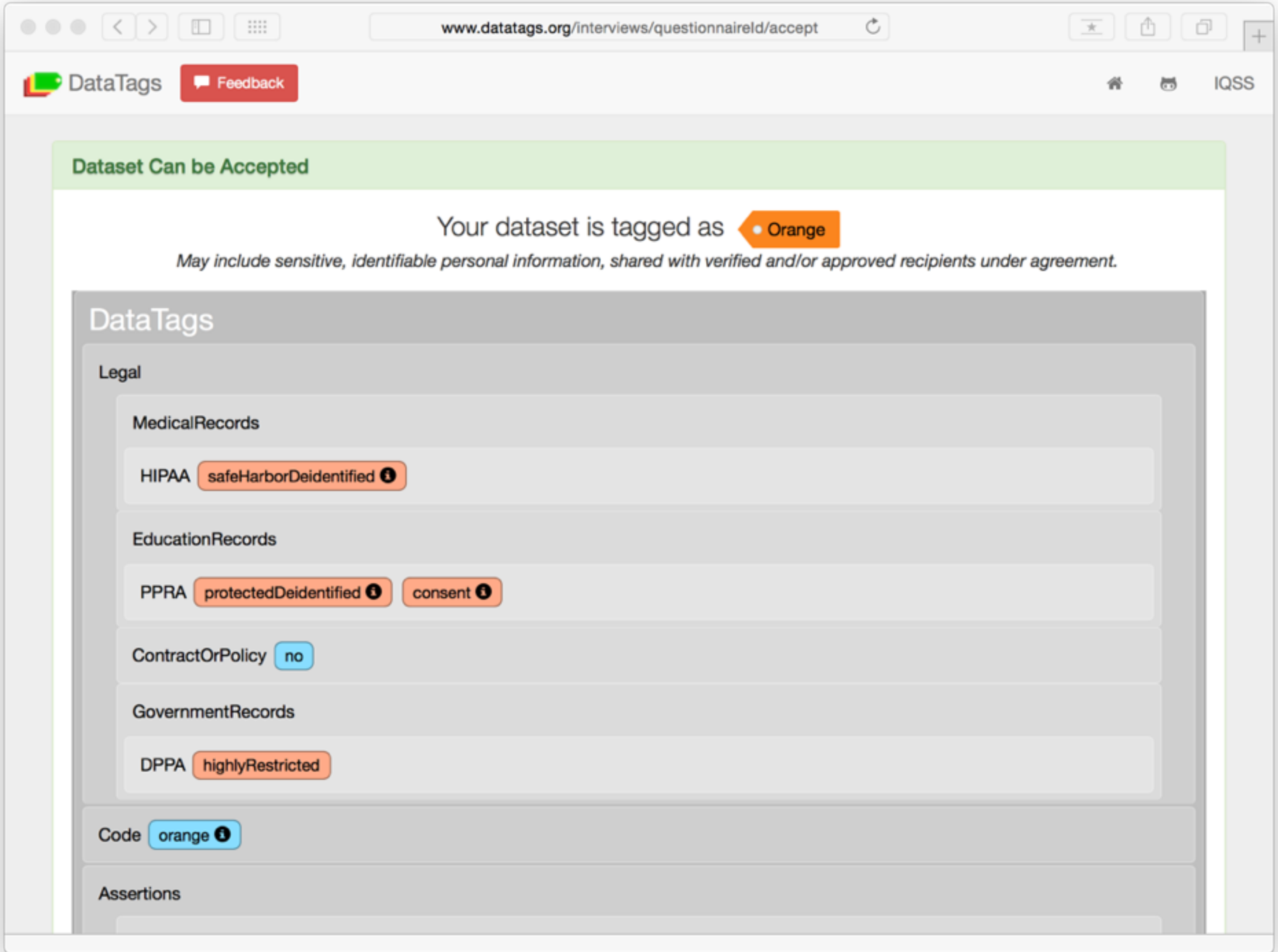
Do the data contain information related to substance abuse diagnosis, referral, or treatment? No

Current Tags

DataTags

Code: green

… helps you generate a machine-readable datatag for your data



Dataset Can be Accepted

Your dataset is tagged as Orange

May include sensitive, identifiable personal information, shared with verified and/or approved recipients under agreement.

DataTags

Legal

MedicalRecords

HIPAA safeHarborDeidentified

EducationRecords

PPRA protectedDeidentified consent

ContractOrPolicy no

GovernmentRecords

DPPA highlyRestricted

Code orange

Assertions

Thanks!

Learn more at http//dataverse.org

@mercecrosas   mercecrosas.com