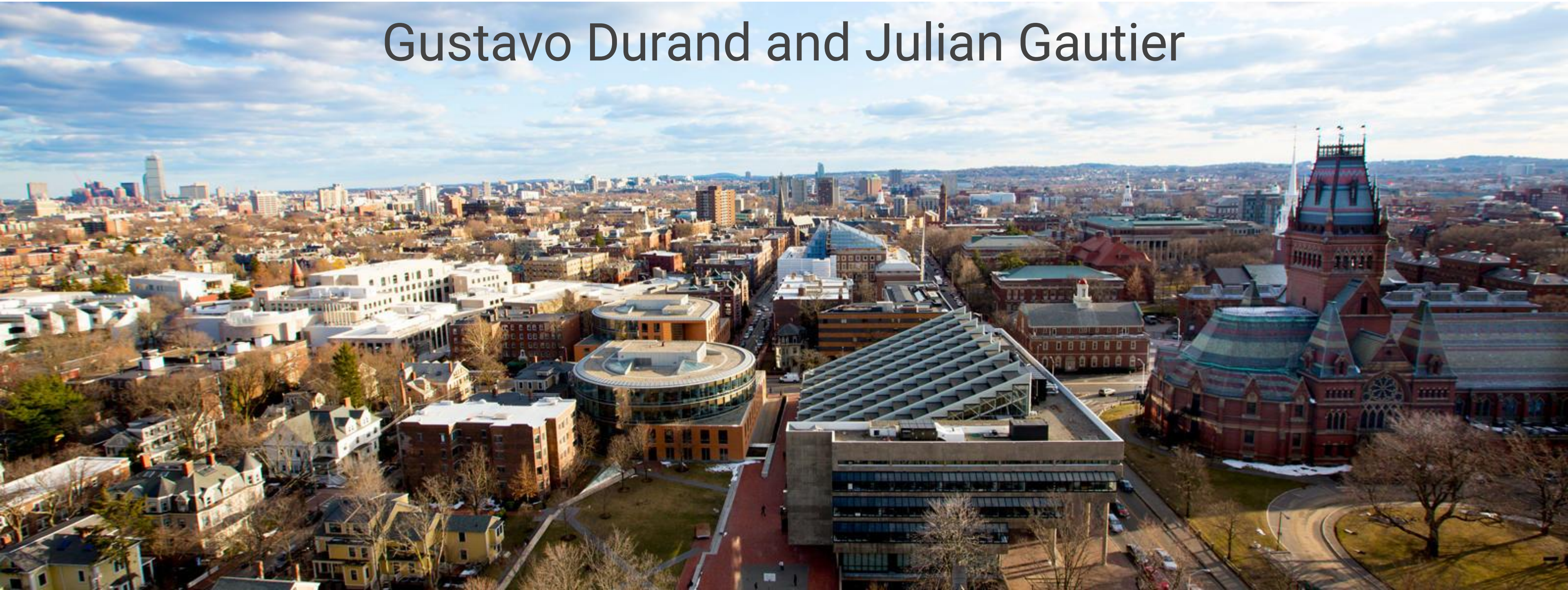
Managing, Exploring, and Sharing Data with Dataverse

The Dataverse® Project

IQSS

The Institute for Quantitative Social Science



Gustavo Durand and Julian Gautier

Introduction to Dataverse

Overview

- **An open-source platform to publish, cite, and archive research data**
- Built to support multiple types of data, users, and workflows
- Developed at Harvard’s Institute for Quantitative Social Science (IQSS) since 2006
- Development funded by IQSS and with grants, in collaboration with institutions around the world
- 12 on the core team - developers, designers, UI/UX, metadata specialists, curation team, leadership team

Dataverse Technology

Glassfish Server 4.1

Java SE8

Java EE7

- Presentation: JSF (PrimeFaces), RESTful API
- Business: EJB, Transactions, Asynchronous, Timers
- Storage: JPA (Entities), Bean Validation

Storage: Postgres, Solr, File System / Swift / S3

Dataverse Features - Data

- Persistent IDs / URLs
  - DataCite
  - Handle
- Automatically Generated Citations with attribution
- Compliant with FAIR and data citation principles
- Domain-specific Metadata
- Versioning
- File Storage
  - Local
  - Swift (OpenStack)
  - S3 (Amazon)

Dataverse Features - Users

- Multiple Sign In options
    - Native
    - Shibboleth
    - OAuth (ORCID)
- Dataverses within Dataverses
- Branding
- Widgets

Dataverse Features - Workflows

- Permissions
- Access Controls and Terms of Use
- Publishing Workflows
- Private URLs
- Upload / Download Workflows
    - Browser
    - Dropbox
    - Rsync (for big data “packages”)

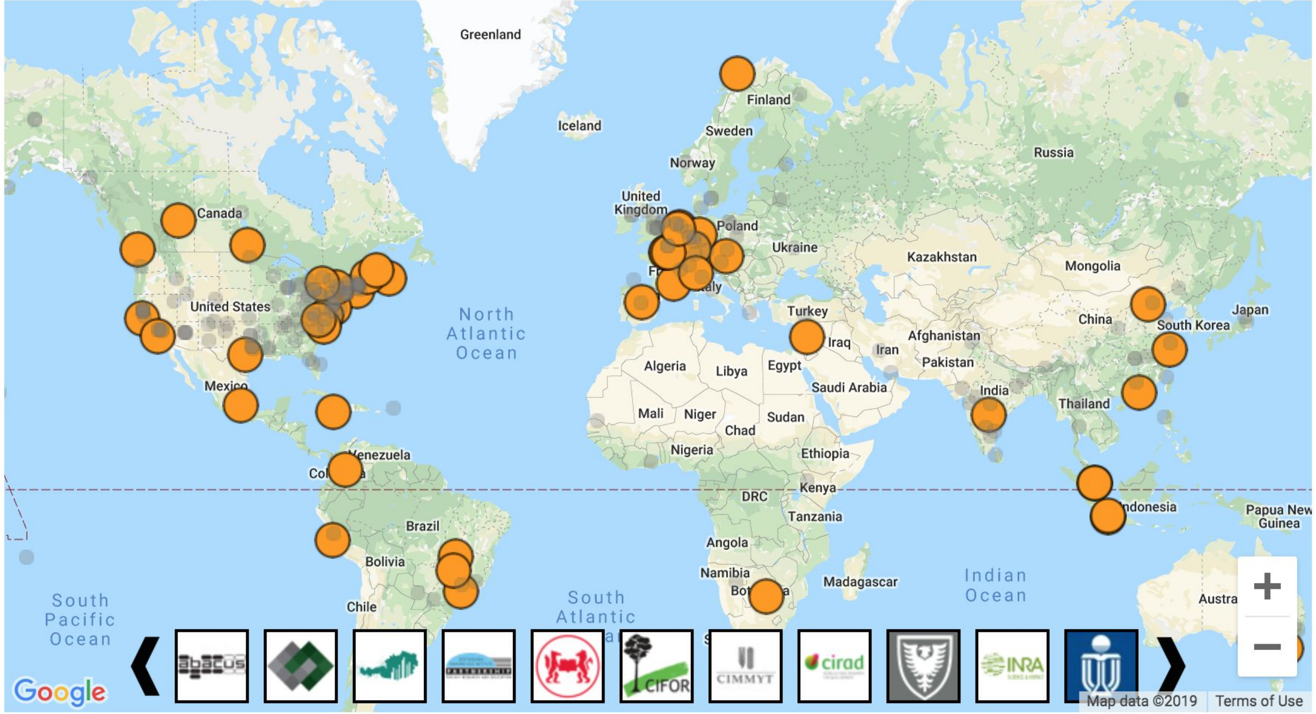
Dataverse Features - Interoperability

- APIs
    - SWORD
    - Native
- Harvesting (OAI-PMH)
    - Client
    - Server
- Modular External Tools
    - Explore
    - Configure

Dataverse Community

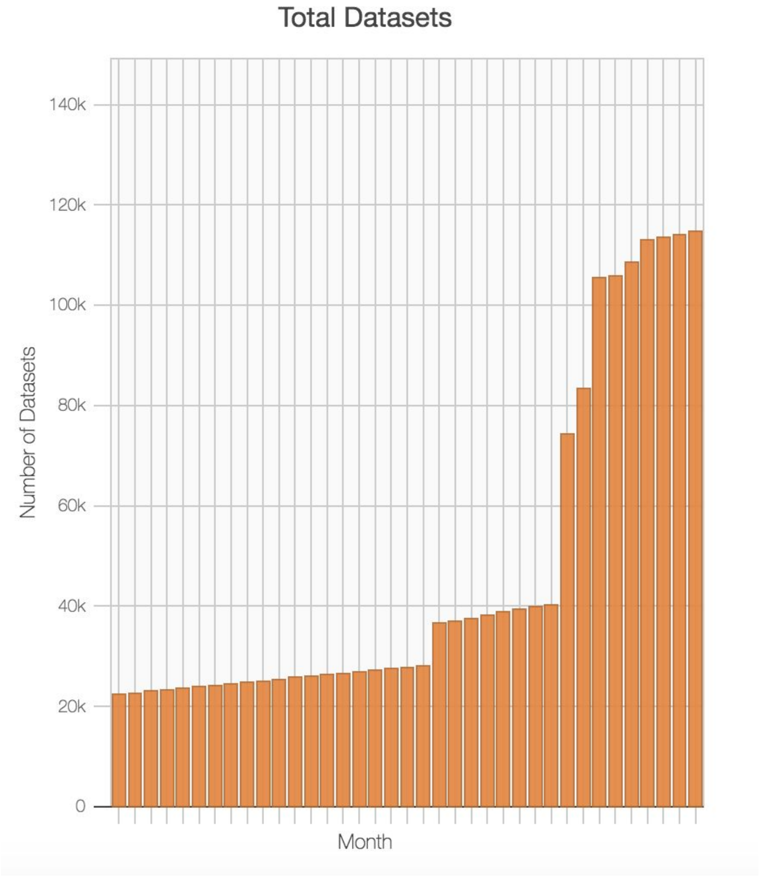
Dataverse Community

- 46 installations around the world



The Data (dataverse.org/metrics)

- 46 installations
- 4,300 Dataverses
- 115,000 Datasets
- 441,000 Files
- 7,710,000 File Downloads



Dataverse Community

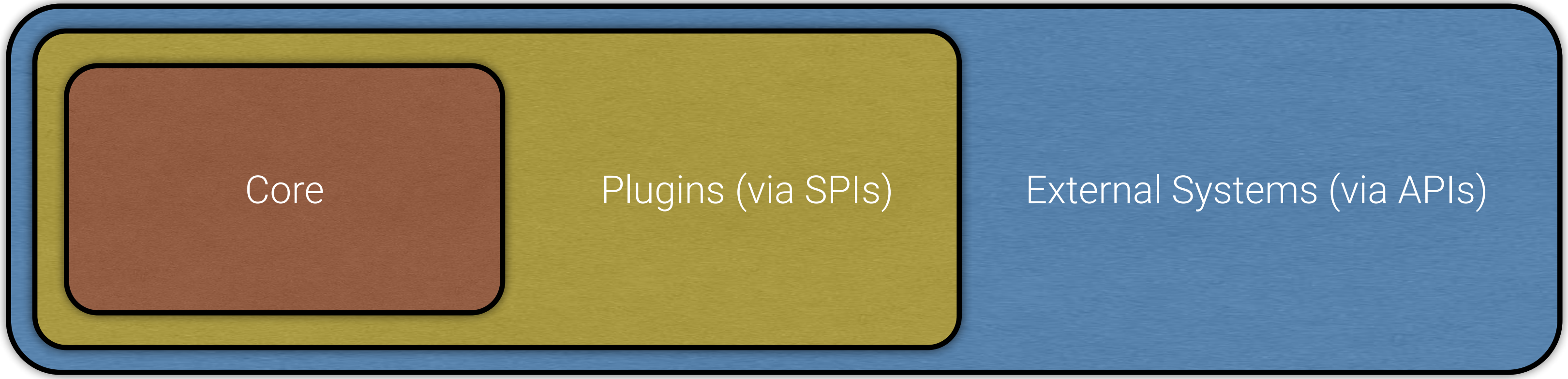
- 90+ Code Contributors
- Hundreds of members of the Dataverse Community - developers, researchers, librarians, data scientists
    - Dataverse Google Group
    - Dataverse Community Calls
    - Dataverse Community Meeting
    - Global Dataverse Community Consortium

https://dataverse.org/events

The Dataverse Cup 🏆

Community Development

Dataverse Ecosystem



Core - Contributing Code to the Dataverse Repo

- Let’s talk early and often!
  - Preview vs Review
- We like small batches, but we’ll follow your lead
- References
  - Developer’s Guide
  - Style Guide
  - API Guide

SPIs / APIs - Why Modularity Matters

- Dataverse is a big application that serves many disciplines with various different needs
  - Almost no-one uses the full functionality
- Modular design allows:
  - Easier code contributions
  - Tailoring installations to institution needs
  - Smaller, more efficient, core
- SPIs - Dataverse calling custom code
- APIs - custom code calling Dataverse

Example Collaborations (Core)

- SBGrid Data
  - Large Data and Support
- Massachusetts Open Cloud
  - Big Data Storage and Compute Access (OpenStack)
- Provenance
  - W3C PROV
- Australian Data Archive (ADA)
  - Use Guestbook for Request Access

Example Collaborations (SPIs)

- SBGrid Data
  - Pre Publish Workflows
- DANS/CIMMYT/GESIS
  - Handles
  - da|ra

Example Collaborations (APIs)

- File Access APIs (External Tools)
  - Harvard SEAS - TwoRavens
  - Scholars Portal - Data Explorer
  - QDR - File Previewers for pdfs, images, videos
- Deposit APIs
  - Open Journal Systems - OJS Plugin
- Client Libraries
  - ResearchSpace - Java
  - AUSSDA - python - pyDataverse

The Future of Dataverse

Dataverse Roadmap

https://www.iq.harvard.edu/roadmap-dataverse-project

- Strategic Goals
- Implementation, Planning, Future

External Tools for Datasets

- Ability to launch an External Tool at the Dataset level
- Examples
  - Code Ocean / Reproducibility tools
  - Compute (for big data)
- Infrastructure development is in the current sprint at IQSS

DataTags

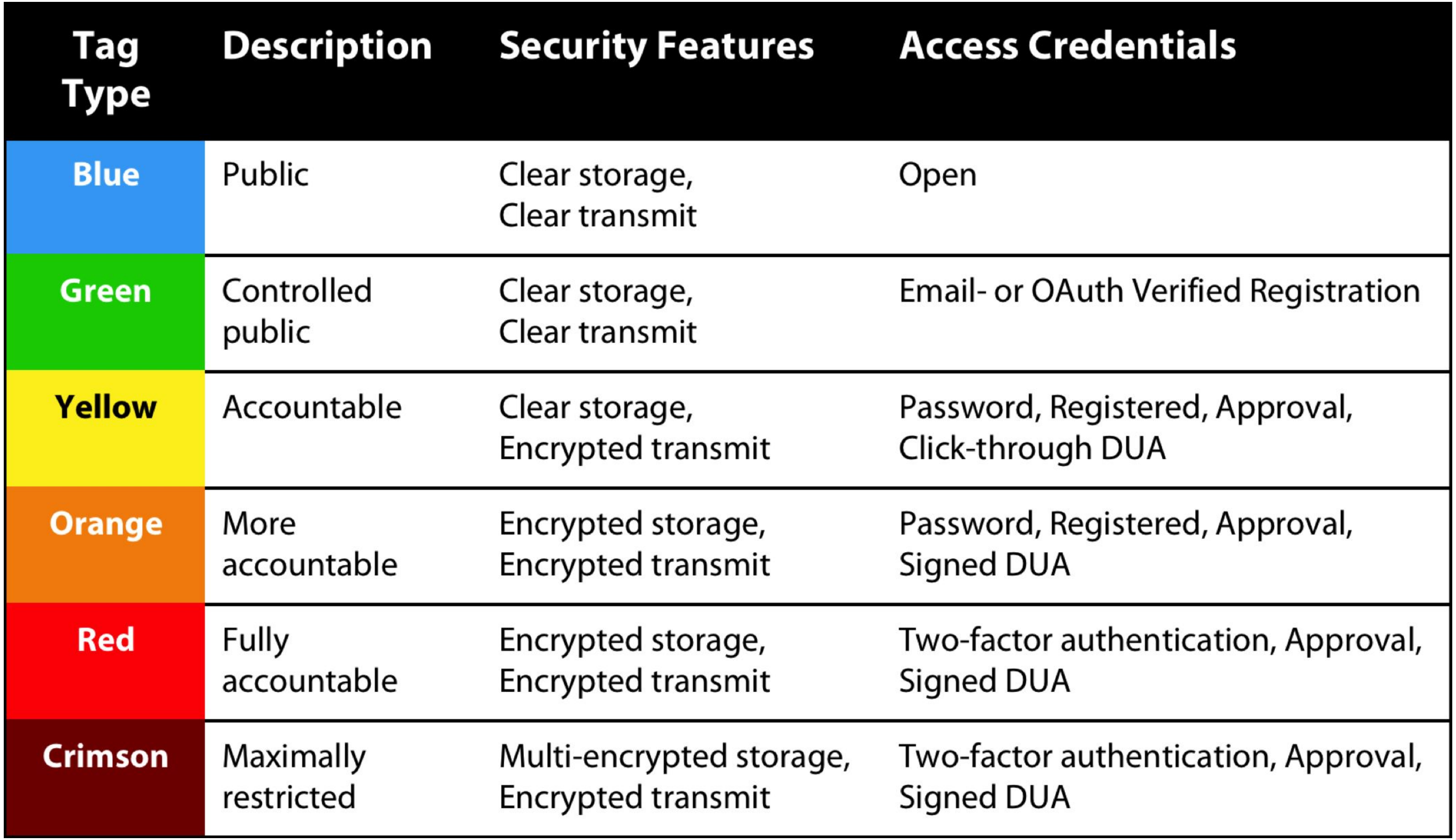
DataTags

Share Sensitive Data with Confidence

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 level 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

Dataverse & DataTags

- Implementation underway with experts from across the University
- Staged implementation of less sensitive DataTags first

Differential Privacy

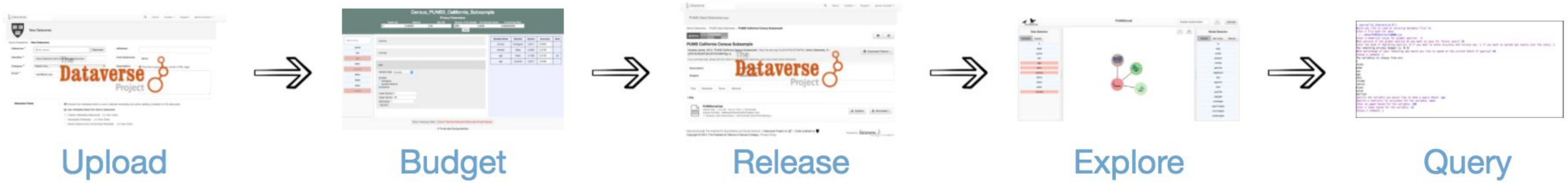
$$Pr[T(M(X)) = 1] \leq e^{\epsilon} Pr[T(M(X')) = 1] + \delta, \qquad \forall\, T.$$

Differential Privacy is a formal, mathematical conception of privacy preservation.

It **guarantees** that any reported result does not reveal information about any one single individual, regardless of auxiliary information.

PSI (Differential Privacy)

Private data Sharing Interface



Upload → Budget → Release → Explore → Query

- **upload** private data to a secured Dataverse archive,
- decide / **budget** what statistics they would like to release about that data
- **release** privacy preserving versions of those statistics to the repository
- that can be **explored** through a curator interface without releasing the raw data
- including interactive **queries**.

Dataverse & PSI

- Prototype of integration with the PSI tool
- Ability to store multiple versions of metadata; external tools able to access the different versions based on user

TRSA

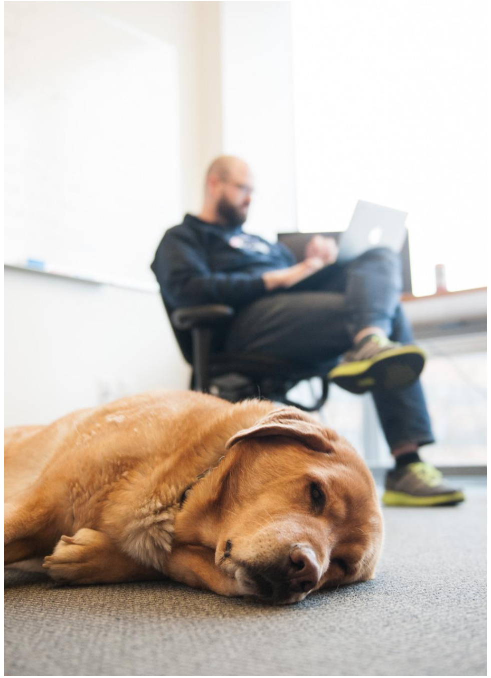
- **T**rusted **R**emote **S**torage **A**gents
  - Agent - Dataverse can communicate with this
  - Storage - especially for sensitive or big data
  - Remote - Dataverse does not control access
  - Trusted - service agreement guarantees

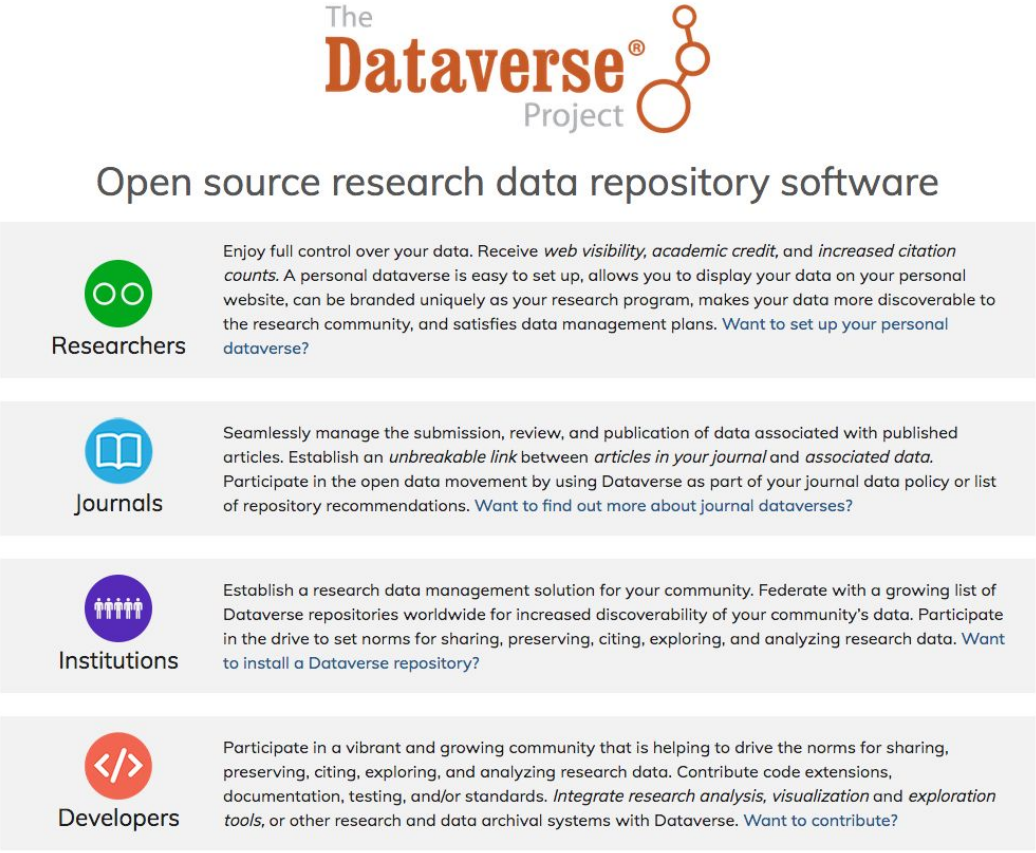
Dataverse & TRSA

- Developed by Odum with guidance for IQSS
- Prototype currently available; work to merge into core code has begun (needed APIs, UI / UX design)

Thank you!

@dataverseorg
gdurand@iq.harvard.edu
juliangautier@g.harvard.edu
support@dataverse.org





The Dataverse® Project

Open source research data repository software

Researchers

Enjoy full control over your data. Receive *web visibility, academic credit*, and *increased citation counts*. A personal dataverse is easy to set up, allows you to display your data on your personal website, can be branded uniquely as your research program, makes your data more discoverable to the research community, and satisfies data management plans. Want to set up your personal dataverse?

Journals

Seamlessly manage the submission, review, and publication of data associated with published articles. Establish an *unbreakable link* between *articles in your journal* and *associated data*. Participate in the open data movement by using Dataverse as part of your journal data policy or list of repository recommendations. Want to find out more about journal dataverses?

Institutions

Establish a research data management solution for your community. Federate with a growing list of Dataverse repositories worldwide for increased discoverability of your community's data. Participate in the drive to set norms for sharing, preserving, citing, exploring, and analyzing research data. Want to install a Dataverse repository?

Developers

Participate in a vibrant and growing community that is helping to drive the norms for sharing, preserving, citing, exploring, and analyzing research data. Contribute code extensions, documentation, testing, and/or standards. *Integrate research analysis, visualization* and *exploration tools*, or other research and data archival systems with Dataverse. Want to contribute?

https://dataverse.org
https://github.com/iqss/dataverse