Dataverse's Approach to Technical Community Engagement

The Dataverse® Project

Gustavo Durand
Tech Lead / Architect

What we'll cover

- Introduction to Dataverse
- Dataverse Community
  - GDCC
- Contributing to Dataverse
  - Contributing to the Core Code
  - Extending via SPIs
  - Connecting via APIs
- The Future of Dataverse

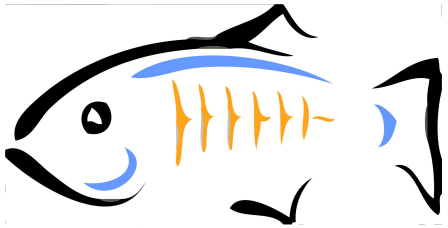
Introduction to Dataverse

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
- 10 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)
- *DataTags for Sensitive Data*

Dataverse Features - Users

- Multiple Sign In options
  - Native
  - Shibboleth
  - OAuth (ORCID, Github, Google, Microsoft)
  - Open ID Connect
- Dataverses within Dataverses
- Branding
- Widgets

Dataverse Features - Workflows

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

Dataverse Features - Interoperability

- APIs
  - SWORD
  - Native
  - Metrics
- Harvesting (OAI-PMH)
  - Client
  - Server
- Modular External Tools
  - Explore vs Configure
  - Scope: Dataset / Datafile

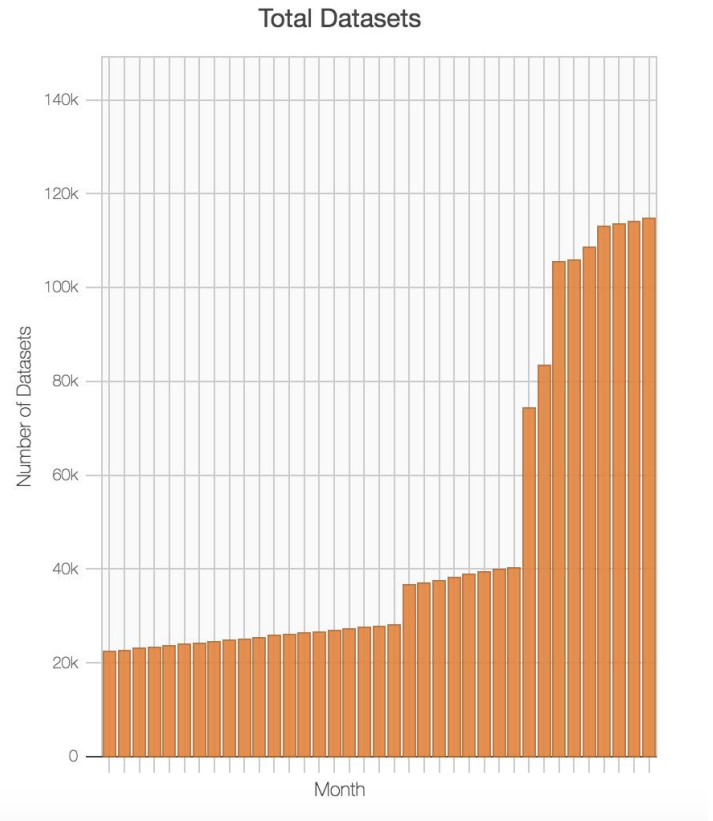
Dataverse Community

Dataverse Community

- **52** installations around the world

The Data (dataverse.org/metrics)

- 52 installations
- 5,900 Dataverses*
- 127,000 Datasets*
- 547,000 Files*
- 11,400,000 File Downloads*



* metrics collected from 27 installations
(running 4.9 and newer)

Dataverse Community

- 100+ Contributors
- Hundreds of members of the Dataverse Community - developers, researchers, librarians, data scientists
  - Workshops & Trainings
  - UI/UX Testing & Interviews
  - Global Dataverse Community Consortium
  - Dataverse Google Group
  - Dataverse Community Calls
  - Dataverse Community Meeting

The Dataverse Cup 🏆

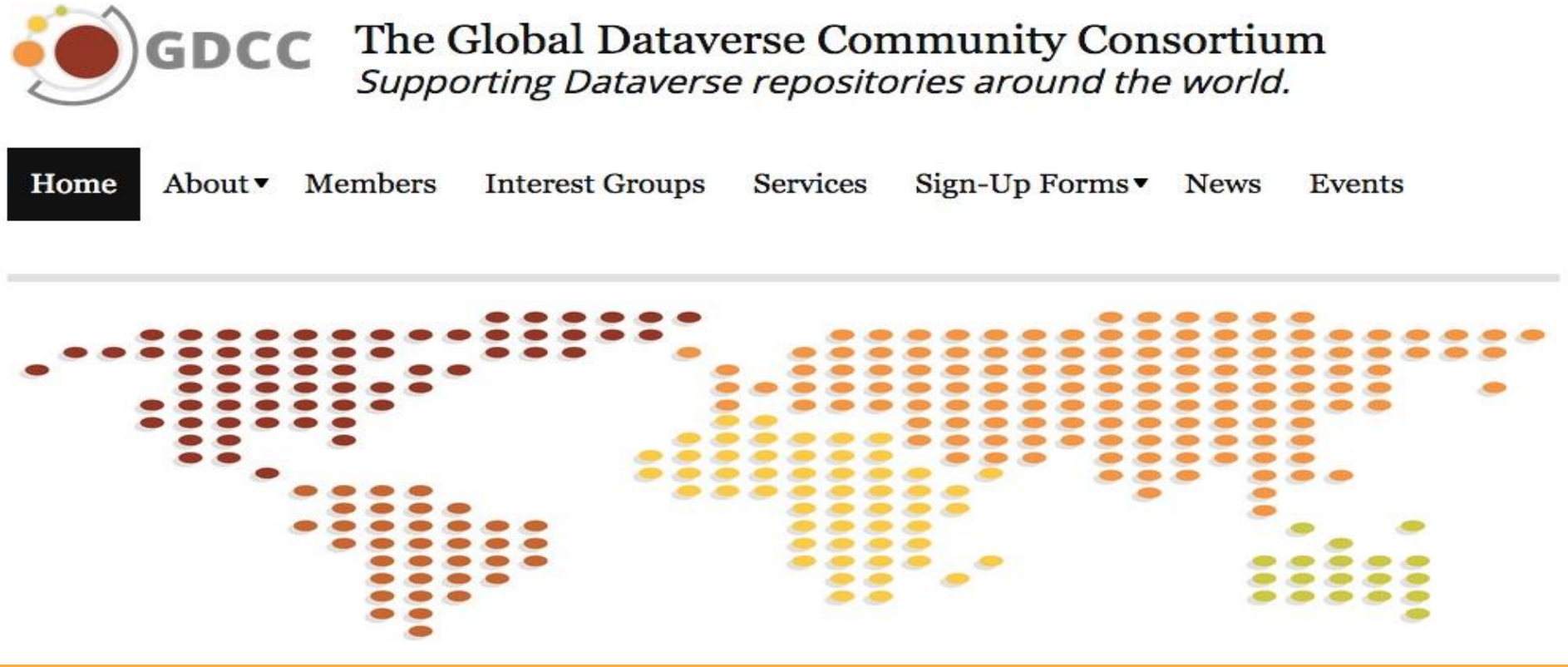
Global Dataverse Community Consortium

Global Dataverse Community Consortium

- Supporting Dataverse repositories around the world

The Global Dataverse Community Consortium (GDCC) is dedicated to providing international organization to existing Dataverse community efforts, and will provide a collaborative venue for institutions to leverage economies of scale in support of Dataverse repositories around the world.

http://DataverseCommunity.Global

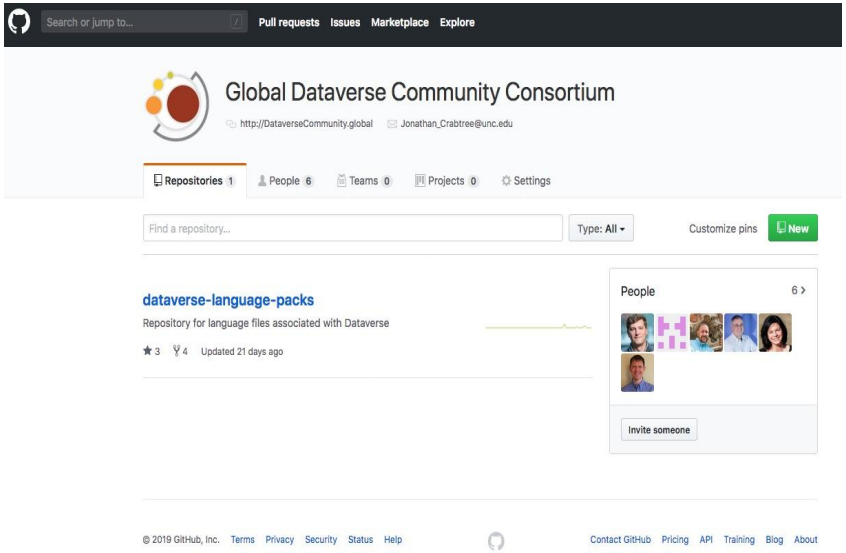


Global Dataverse Community Consortium

Australian Data Archive	Australia
Consorcio Madrono	Spain
DANS	Netherlands
DataverseNO	Norway
Fudan University	China
Gottingen eResearch Alliance	Germany
Harvard University	United States
International Centre for Research in Agroforestry	Kenya
Johns Hopkins University	United States
Nanyang Technological University	Singapore
Syracuse University	United States
Texas Digital Library	United States
University of California Los Angeles	United States
University of Campinas	Brazil
University of North Carolina Chapel Hill	United States
University of Virginia	United States

Membership Expanding

Initial Services



New Potential Services?

Collaborative Code Development

Shared Programming Staff

Joint Documentation Initiative

Collaborative Code Testing

Joint Funding Applications

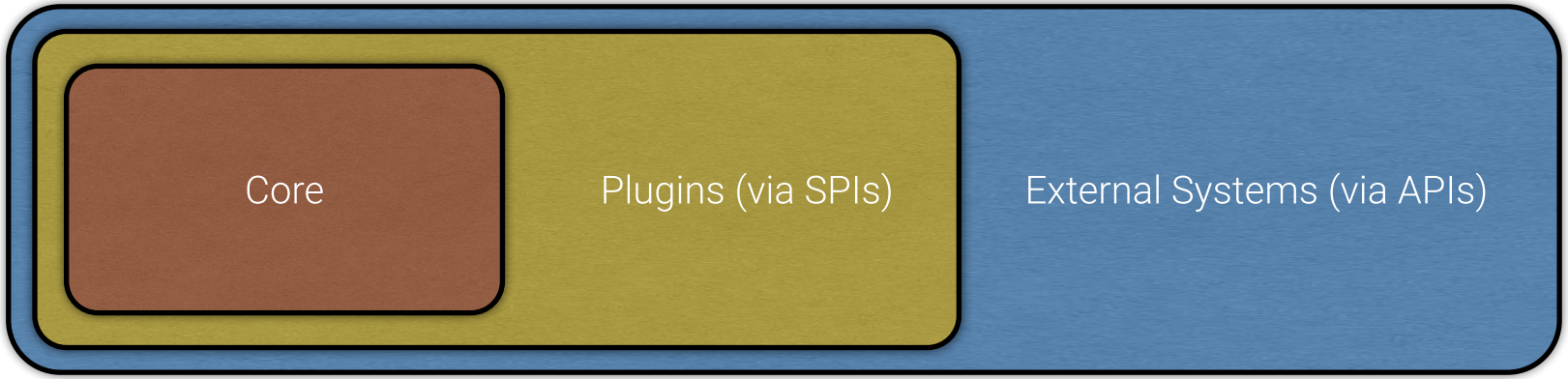
Shared Community Policies

?????????

Contributing to Dataverse

Community Development

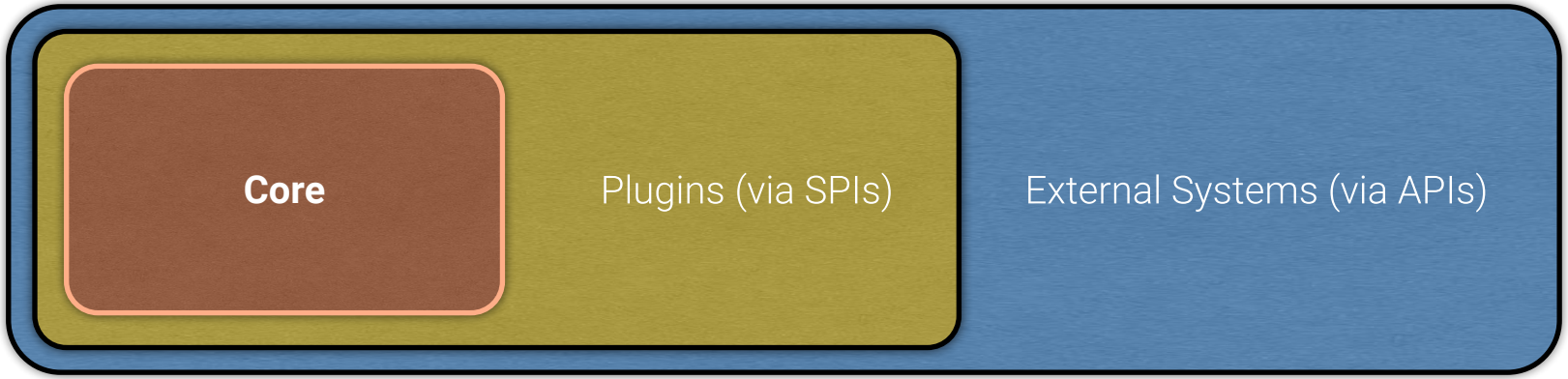
Dataverse Ecosystem



References for Developers

- http://guides.dataverse.org/
  - Developer's Guide
  - Style Guide
  - API Guide
  - Installation Guide (for External Tools)

Dataverse Ecosystem

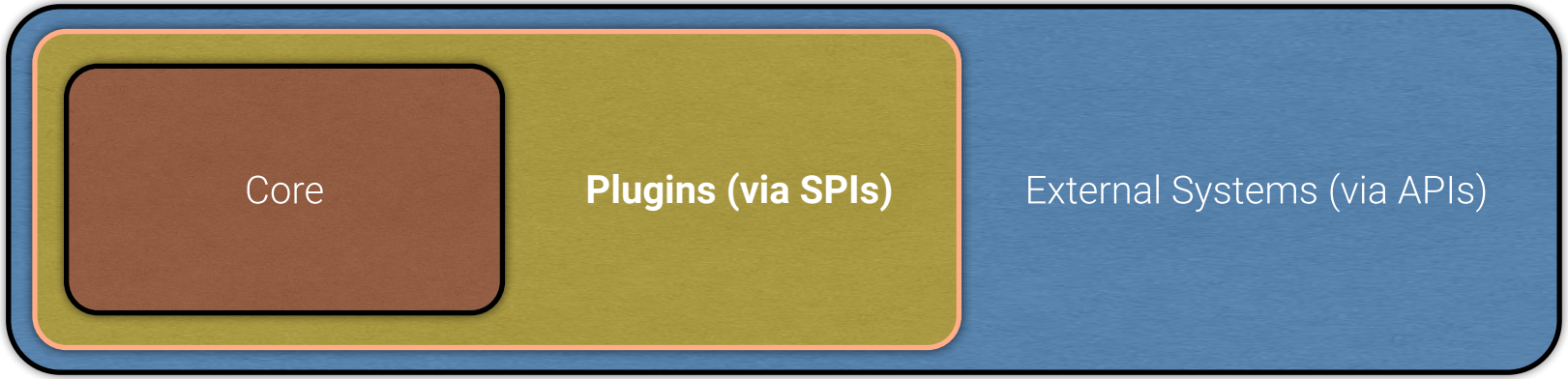


Contributing Code to the Core

- General Concepts
  - Let’s talk early and often! (Preview vs Review)
  - We like small batches, but we’ll follow your lead
- Process
  - Design
  - Development
  - Code Review
  - QA
  - (Iterate)
  - Merge

Example Collaborations (Core)

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


Core
Plugins (via SPIs)
External Systems (via APIs)

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

Service Provider Interface (SPI) is an API intended to be implemented or extended by a third party. It can be used to enable framework extension and replaceable components.

from Wikipedia, The Free Encyclopedia

Code is designed such that functionalities are implemented in cohesive modules with clear interface to the rest of the code

- **Internal service providers**
  - Extension via pull-requests
  - General use by multiple installations
- **External service providers**
  - Extension via .jar files
  - Specific use by one or few installations

Dataverse uses of SPIs

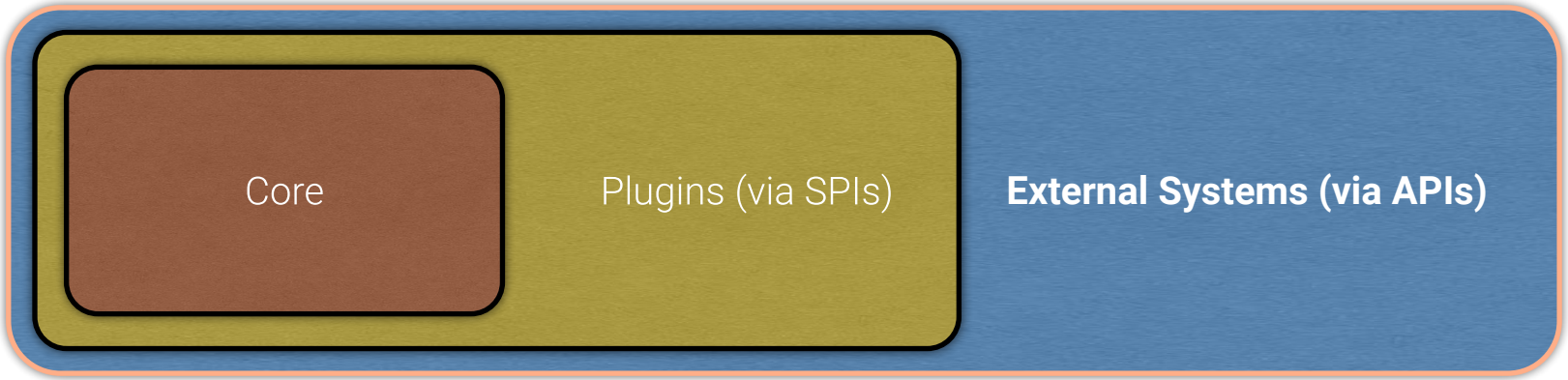
```
@PostConstruct
public void startup() {

    // First, set up the factories
    try {
        registerProviderFactory( new BuiltinAuthentic
        registerProviderFactory( new ShibAuthenticati
        registerProviderFactory( new OAuth2Authentica
    } catch (AuthorizationSetupException ex) {
        logger.log(Level.SEVERE, "Exception setting u
    }
```

- Identity Providers
- Exporters
- Workflow Step Providers
- Persistent Identifier Registration

Example Collaborations (SPIs)

- SBGrid Data
  - Pre Publish Workflows
- DANS/CIMMYT
  - Handles


Core
Plugins (via SPIs)
External Systems (via APIs)

Why APIs?

- **Design App as a platform**
- **Foster a developer community**
- Setup application state
  - pre-populating a fresh dev install (e.g. after dropping the db)
  - pre-populating test cases
- Inspect application state without the UI (which might not be there)
- **Allow for integration tests**

Dataverse uses of APIs

- Setup
- Admin
- **File Access**
- **Deposit**
- **Export**

External Tools

- Tools that talk to Dataverse
- Tools that Dataverse talks to
- Tools that do both

Tools that talk to Dataverse

- Generally use the Deposit API
- Don’t require anything special in Dataverse Core Code
- Examples
  - **OJS Plugin** - adds data sharing and preservation to the Open Journal Systems publication process.
  - **OSF Plugin** - Allows you to view, download, and upload files to and from a Dataverse dataset from an Open Science Framework project

Tools that Dataverse talks to

- Modular Explore Button on Dataset / Data File page
  - Sends user to external tool site
  - Available for different file types / datasets
- Requires a manifest file to be uploaded into Dataverse
- Examples:
  - **TwoRavens** - a system of interlocking statistical tools for data exploration, analysis, and meta-analysis
  - **Data Explorer** - a GUI which lists the variables in a tabular data file allowing searching, charting and cross tabulation analysis

Tools that do both

- Modular Configure Button on Dataset / Data File page
  - Sends user to external tool site
  - Uses Dataverse APIs to send something back to Dataverse
  - Currently only for ingested Tabular files
- Requires a manifest file to be uploaded into Dataverse
- Example:
  - **PSI** - a Private data Sharing Interface that allows researchers with sensitive datasets to make differentially private statistics about their data available through data repositories
  - **Data Curation Tool** - A tool for curating data by adding labels, groups, weights and other details to assist with informed reuse

External Tools Manifest

- External tools must be expressed in an external tool manifest file
- Can be uploaded to Dataverse via API

```
{
  "displayName": "Awesome Tool",
  "description": "The most awesome tool.",
  "type": "explore",
  "toolUrl": "https://awesometool.com",
  "toolParameters": {
    "queryParameters": [
      {
        "fileid": "{fileId}"
      },
      {
        "key": "{apiToken}"
      }
    ]
  }
}
```

Example Collaborations (APIs)

- File Access APIs (External Tools)
  - Harvard SEAS - TwoRavens
  - Scholars Portal - Data Explorer, Data Curation Tool
  - 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

The “Present” of Dataverse

Dataverse 4.19

- Released Wednesday!
- Release Highlights:
    - Open ID Connect Support (contributed by Oliver, FZJ)
    - Python Installer
    - Support for Glassfish upgrade
    - Full Release Notes: https://github.com/IQSS/dataverse/releases/tag/v4.19

Dataverse Roadmap

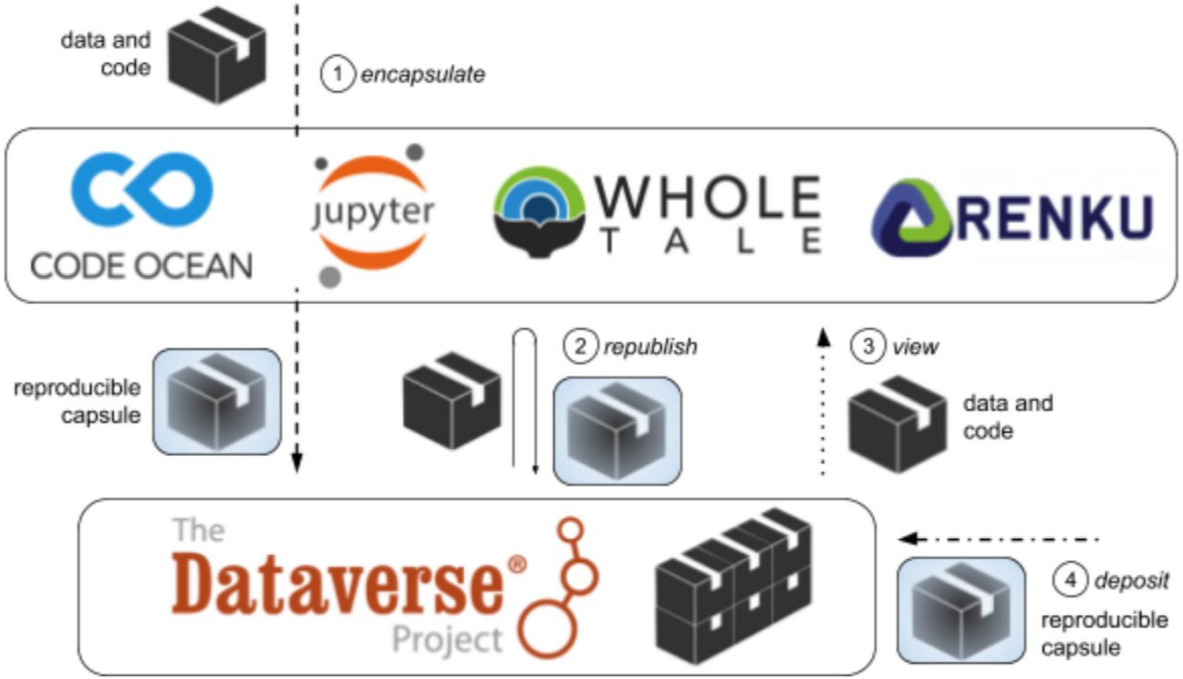
https://www.iq.harvard.edu/roadmap-dataverse-project

- Strategic Goals
- Implementation, Planning, Future

Dataset Page / File Page Redesign

- a more modular, scalable, accessible, and responsive experience
- informed by present and future use cases
- **Validately** remote prototype usability test
  - This version has been simplified and tuned for researchers
  - Anyone in the community is welcome to participate
  - https://validately.com/unmoderated/6716a601-2d91-11ea-a2f1-42010af00531

Support for Capsules



DataTags

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.

OpenDP/PSI (Differential Privacy)

Private data Sharing Interface

- **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 & OpenDP

- 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

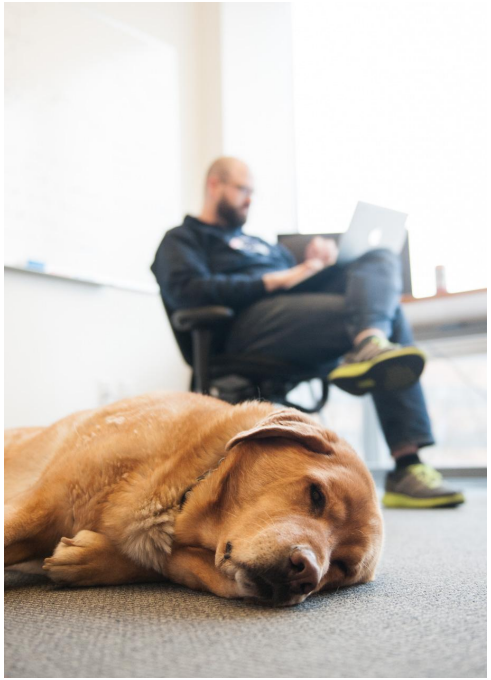
Trusted Remote Storage Agents

- the ability to have a user designate an external location for files rather than to upload them during the ingest process
- By policy, these agents should have an MOU with archive so they agree to maintain object to prevent DOI violations or dead links
- In collaboration with Odum

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!



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