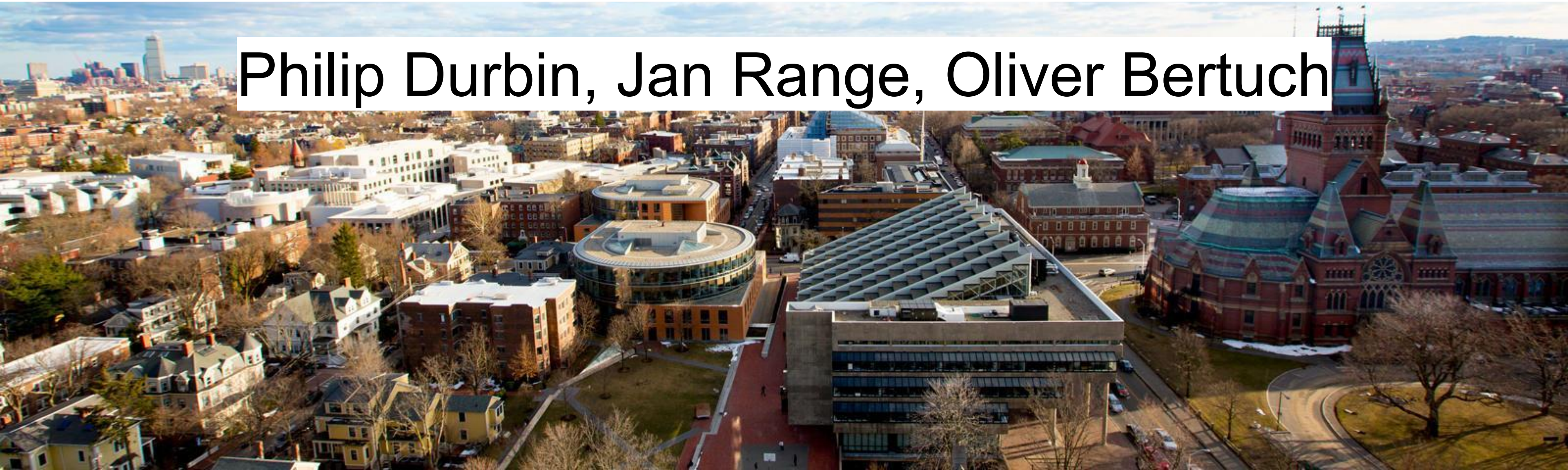
distribits
8a6d1c60e7fdf4194320245ae0c8df0ce682c677

The Dataverse® Project

Distributed Metadata and Data with Dataverse



Philip Durbin, Jan Range, Oliver Bertuch

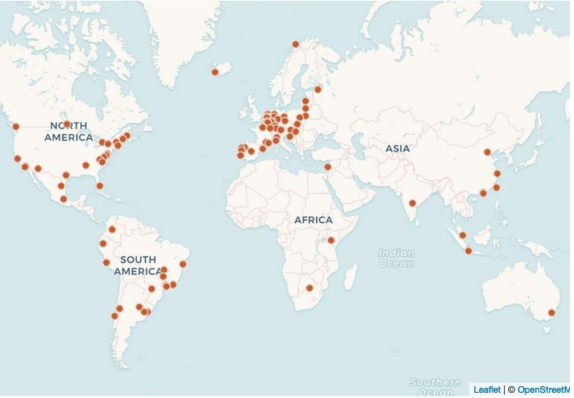
Agenda

- What is Dataverse?
- Distributed metadata: OAI-PMH (Harvesting)
- Distributed data
- Dataverse and DataLad
- Future architecture for a mixed-style storage configuration in Jülich DATA(1)
- Use of Dataverse APIs to manage data with Python (pyDataverse)

(1): Jülich DATA is the institutional data repository of Forschungszentrum Jülich GmbH, based on Dataverse.

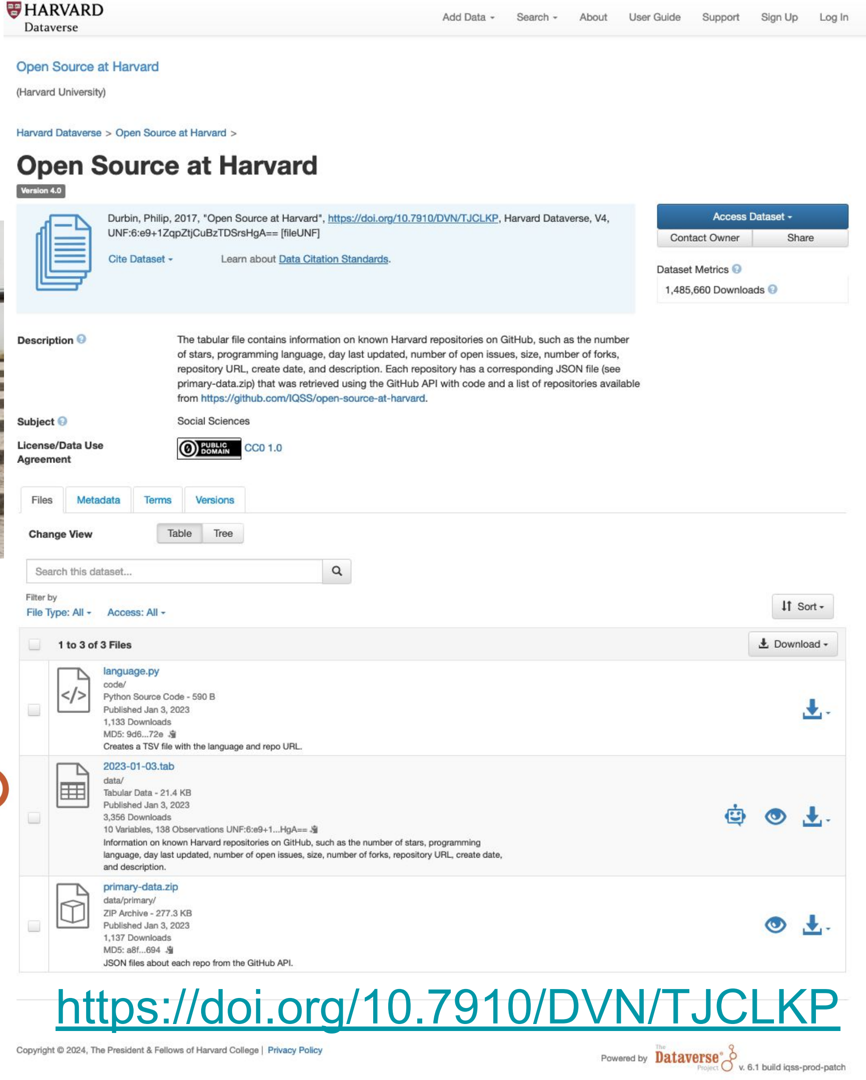
What is Dataverse?

What is Dataverse?


116 Installations

The Dataverse® Project

https://dataverse.org



Open Source at Harvard

(Harvard University)

Harvard Dataverse > Open Source at Harvard >

Open Source at Harvard

Version 4.0

Durbin, Philip, 2017, "Open Source at Harvard", https://doi.org/10.7910/DVN/TJCLKP, Harvard Dataverse, V4, UNF:6:e9+1ZqpZtjCuBzTDSrsHgA== [fileUNF]

Cite Dataset ▾ Learn about Data Citation Standards.

Dataset Metrics

1,485,660 Downloads

Description The tabular file contains information on known Harvard repositories on GitHub, such as the number of stars, programming language, day last updated, number of open issues, size, number of forks, repository URL, create date, and description. Each repository has a corresponding JSON file (see primary-data.zip) that was retrieved using the GitHub API with code and a list of repositories available from https://github.com/IQSS/open-source-at-harvard.

Subject Social Sciences

License/Data Use Agreement PUBLIC DOMAIN CC0 1.0

1 to 3 of 3 Files

language.py
code/
Python Source Code - 590 B
Published Jan 3, 2023
1,133 Downloads
MD5: 9d6...72e
Creates a TSV file with the language and repo URL.

2023-01-03.tab
data/
Tabular Data - 21.4 KB
Published Jan 3, 2023
3,356 Downloads
10 Variables, 138 Observations UNF:6:e9+1...HgA==
Information on known Harvard repositories on GitHub, such as the number of stars, programming language, day last updated, number of open issues, size, number of forks, repository URL, create date, and description.

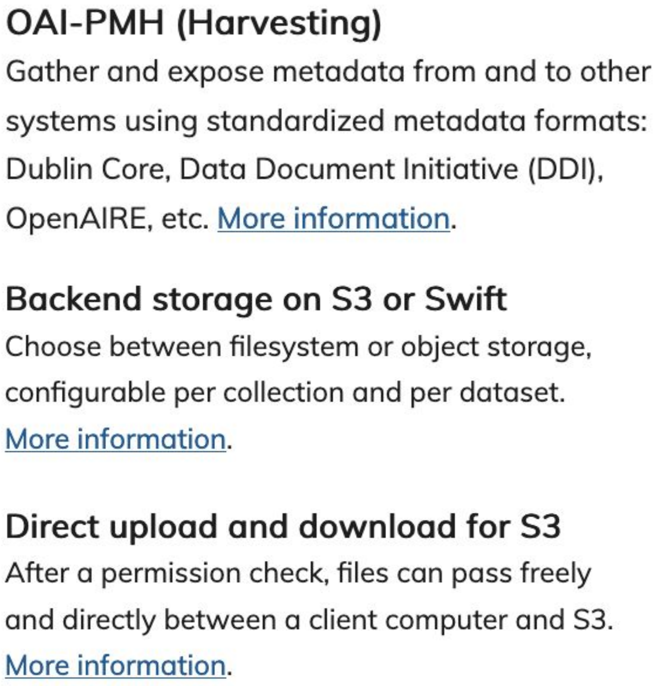
primary-data.zip
data/primary/
ZIP Archive - 277.3 KB
Published Jan 3, 2023
1,137 Downloads
MD5: a8f...694
JSON files about each repo from the GitHub API.

https://doi.org/10.7910/DVN/TJCLKP

Copyright © 2024, The President & Fellows of Harvard College | Privacy Policy

Powered by The Dataverse Project v. 6.1 build iqss-prod-patch

Features



OAI-PMH (Harvesting)

Gather and expose metadata from and to other systems using standardized metadata formats: Dublin Core, Data Document Initiative (DDI), OpenAIRE, etc. More information.

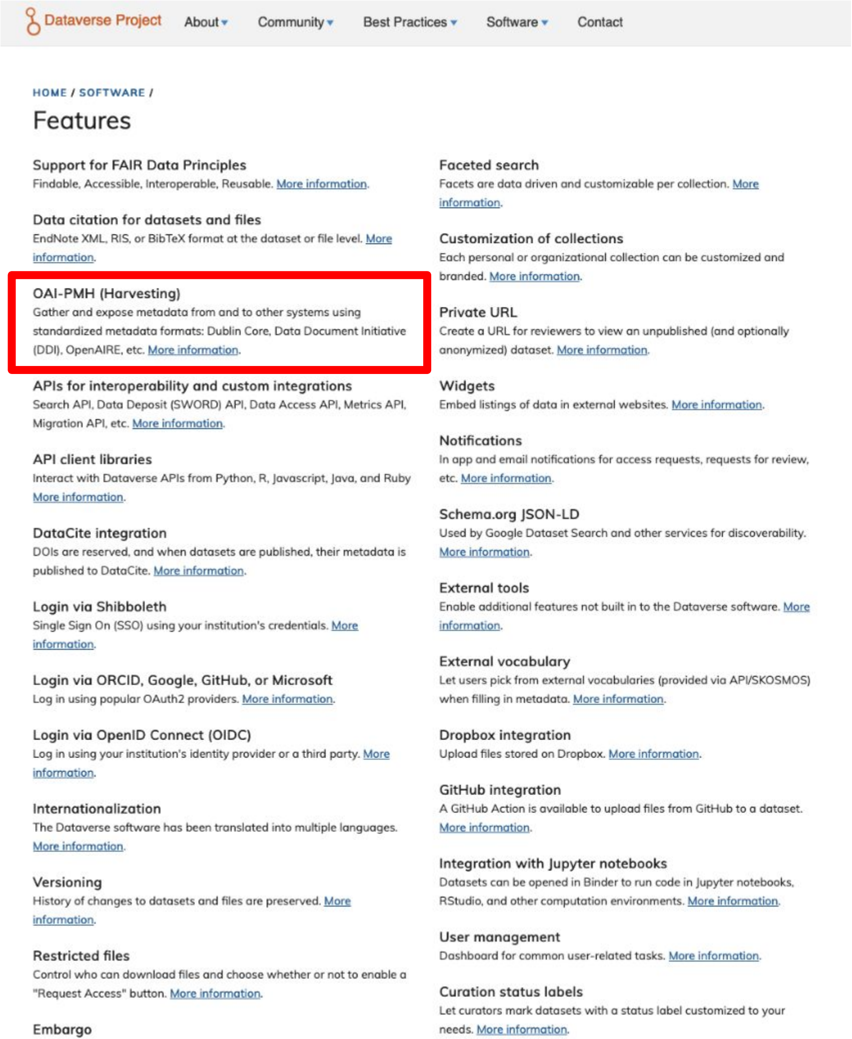
Backend storage on S3 or Swift

Choose between filesystem or object storage, configurable per collection and per dataset. More information.

Direct upload and download for S3

After a permission check, files can pass freely and directly between a client computer and S3. More information.

https://dataverse.org/software-features



HOME / SOFTWARE /

Features

Support for FAIR Data Principles
Findable, Accessible, Interoperable, Reusable. More information.

Data citation for datasets and files
EndNote XML, RIS, or BibTeX format at the dataset or file level. More information.

OAI-PMH (Harvesting)
Gather and expose metadata from and to other systems using standardized metadata formats: Dublin Core, Data Document Initiative (DDI), OpenAIRE, etc. More information.

APIs for interoperability and custom integrations
Search API, Data Deposit (SWORD) API, Data Access API, Metrics API, Migration API, etc. More information.

API client libraries
Interact with Dataverse APIs from Python, R, Javascript, Java, and Ruby More information.

DataCite integration
DOIs are reserved, and when datasets are published, their metadata is published to DataCite. More information.

Login via Shibboleth
Single Sign On (SSO) using your institution's credentials. More information.

Login via ORCID, Google, GitHub, or Microsoft
Log in using popular OAuth2 providers. More information.

Login via OpenID Connect (OIDC)
Log in using your institution's identity provider or a third party. More information.

Internationalization
The Dataverse software has been translated into multiple languages. More information.

Versioning
History of changes to datasets and files are preserved. More information.

Restricted files
Control who can download files and choose whether or not to enable a "Request Access" button. More information.

Embargo

Faceted search
Facets are data driven and customizable per collection. More information.

Customization of collections
Each personal or organizational collection can be customized and branded. More information.

Private URL
Create a URL for reviewers to view an unpublished (and optionally anonymized) dataset. More information.

Widgets
Embed listings of data in external websites. More information.

Notifications
In app and email notifications for access requests, requests for review, etc. More information.

Schema.org JSON-LD
Used by Google Dataset Search and other services for discoverability. More information.

External tools
Enable additional features not built in to the Dataverse software. More information.

External vocabulary
Let users pick from external vocabularies (provided via API/SKOSMOS) when filling in metadata. More information.

Dropbox integration
Upload files stored on Dropbox. More information.

GitHub integration
A GitHub Action is available to upload files from GitHub to a dataset. More information.

Integration with Jupyter notebooks
Datasets can be opened in Binder to run code in Jupyter notebooks, RStudio, and other computation environments. More information.

User management
Dashboard for common user-related tasks. More information.

Curation status labels
Let curators mark datasets with a status label customized to your needs. More information.

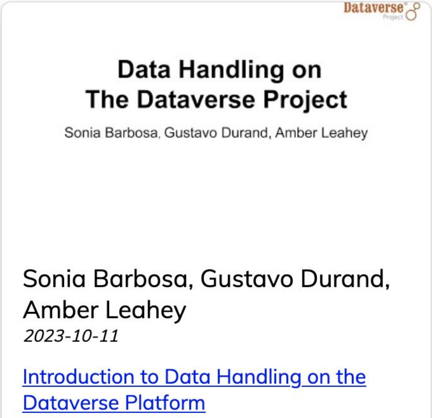
DataverseTV: https://dataverse.org/dataversetv

Introduction to Data Management and FAIR Data

Katie Mika
Data Services Librarian
Harvard Library & Institute for Quantitative Social Science
katherine_mika@harvard.edu

Katie Mika
2023-09-20

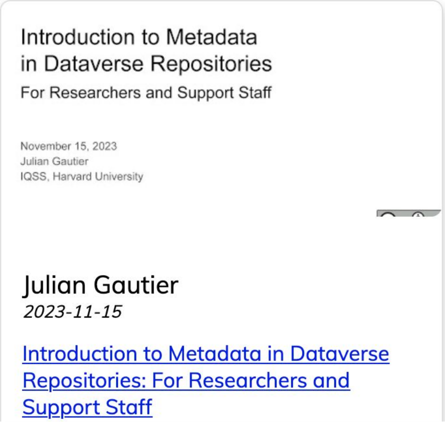
Introduction to Data Management & FAIR


Data Handling on The Dataverse Project

Sonia Barbosa, Gustavo Durand, Amber Leahey

Sonia Barbosa, Gustavo Durand, Amber Leahey
2023-10-11

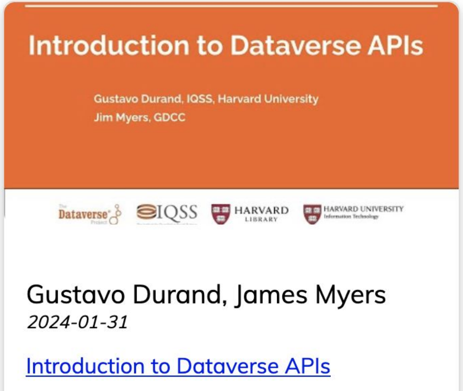
Introduction to Data Handling on the Dataverse Platform


Introduction to Metadata in Dataverse Repositories
For Researchers and Support Staff

November 15, 2023
Julian Gautier
IQSS, Harvard University

Julian Gautier
2023-11-15

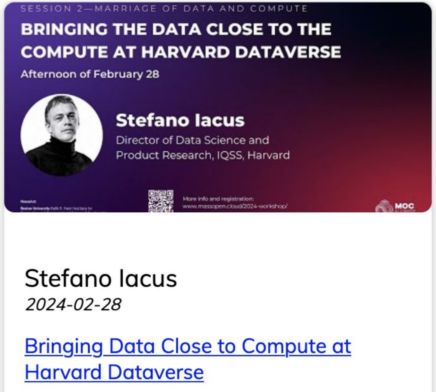
Introduction to Metadata in Dataverse Repositories: For Researchers and Support Staff


Introduction to Dataverse APIs

Gustavo Durand, IQSS, Harvard University
Jim Myers, GDCC

Gustavo Durand, James Myers
2024-01-31

Introduction to Dataverse APIs

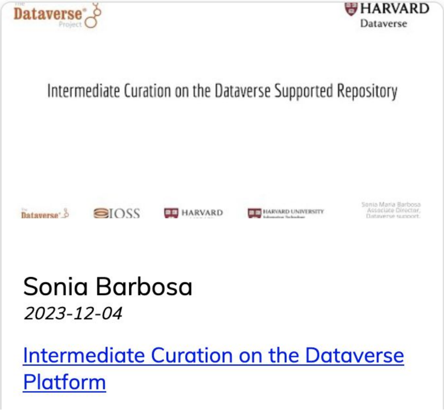

SESSION 2—MARRIAGE OF DATA AND COMPUTE
BRINGING THE DATA CLOSE TO THE COMPUTE AT HARVARD DATAVERSE
Afternoon of February 28

Stefano Iacus
Director of Data Science and Product Research, IQSS, Harvard

More info and registration:
www.massopen.cloud/2024-workshop/

Stefano Iacus
2024-02-28

Bringing Data Close to Compute at Harvard Dataverse



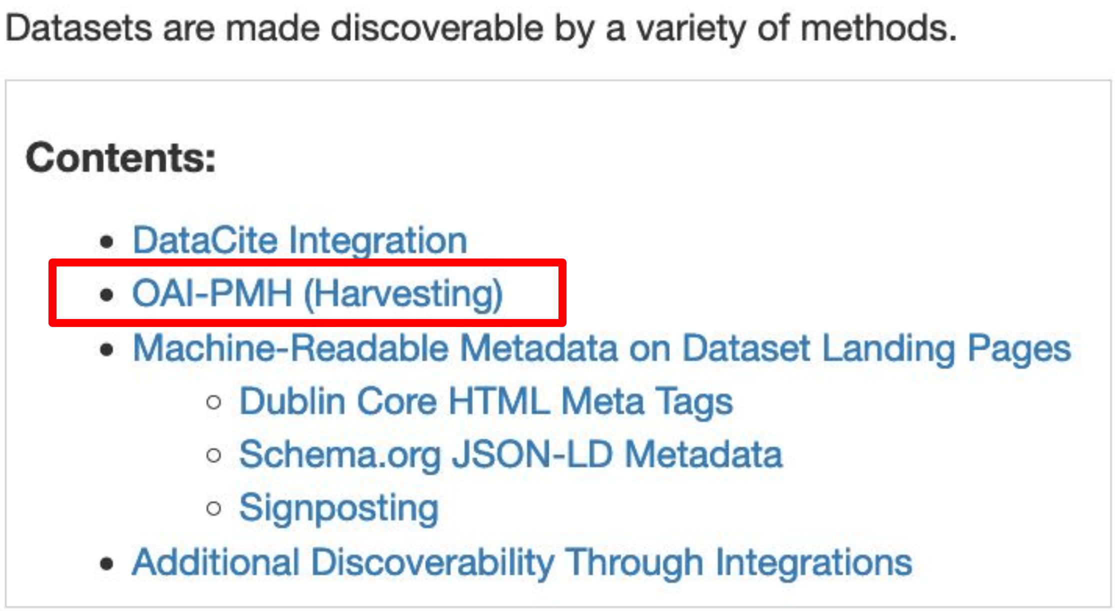
Intermediate Curation on the Dataverse Supported Repository

Sonia Barbosa
2023-12-04

Intermediate Curation on the Dataverse Platform

Distributed Metadata

Dataset discoverability



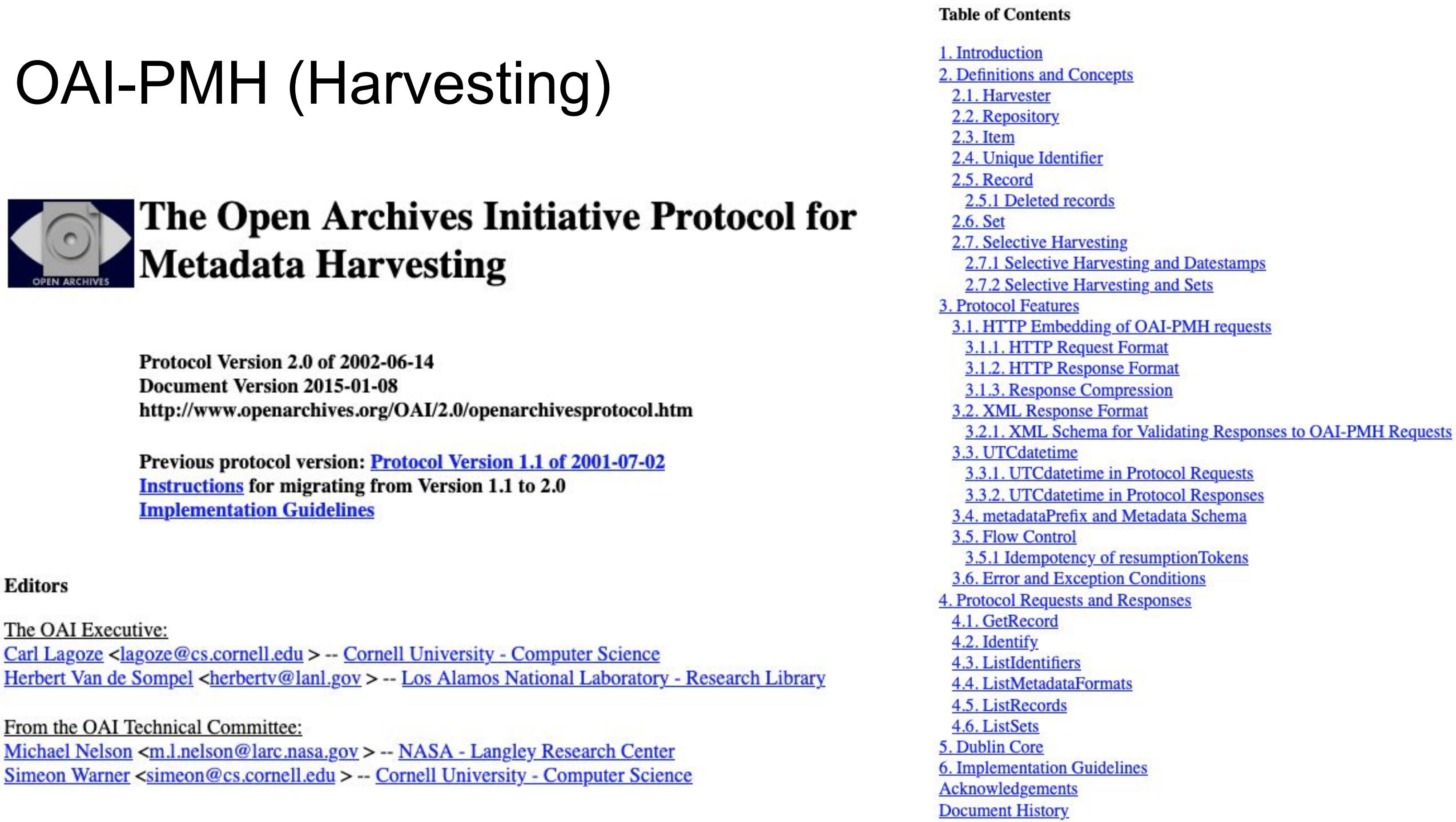
Datasets are made discoverable by a variety of methods.

Contents:

- DataCite Integration
- OAI-PMH (Harvesting)
- Machine-Readable Metadata on Dataset Landing Pages
  - Dublin Core HTML Meta Tags
  - Schema.org JSON-LD Metadata
  - Signposting
- Additional Discoverability Through Integrations

https://guides.dataverse.org/en/6.2/admin/discoverability.html

OAI-PMH (Harvesting)



The Open Archives Initiative Protocol for Metadata Harvesting

Protocol Version 2.0 of 2002-06-14
Document Version 2015-01-08
http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

Previous protocol version: Protocol Version 1.1 of 2001-07-02
Instructions for migrating from Version 1.1 to 2.0
Implementation Guidelines

Editors

The OAI Executive:
Carl Lagoze <lagoze@cs.cornell.edu > -- Cornell University - Computer Science
Herbert Van de Sompel <herbertv@lanl.gov > -- Los Alamos National Laboratory - Research Library

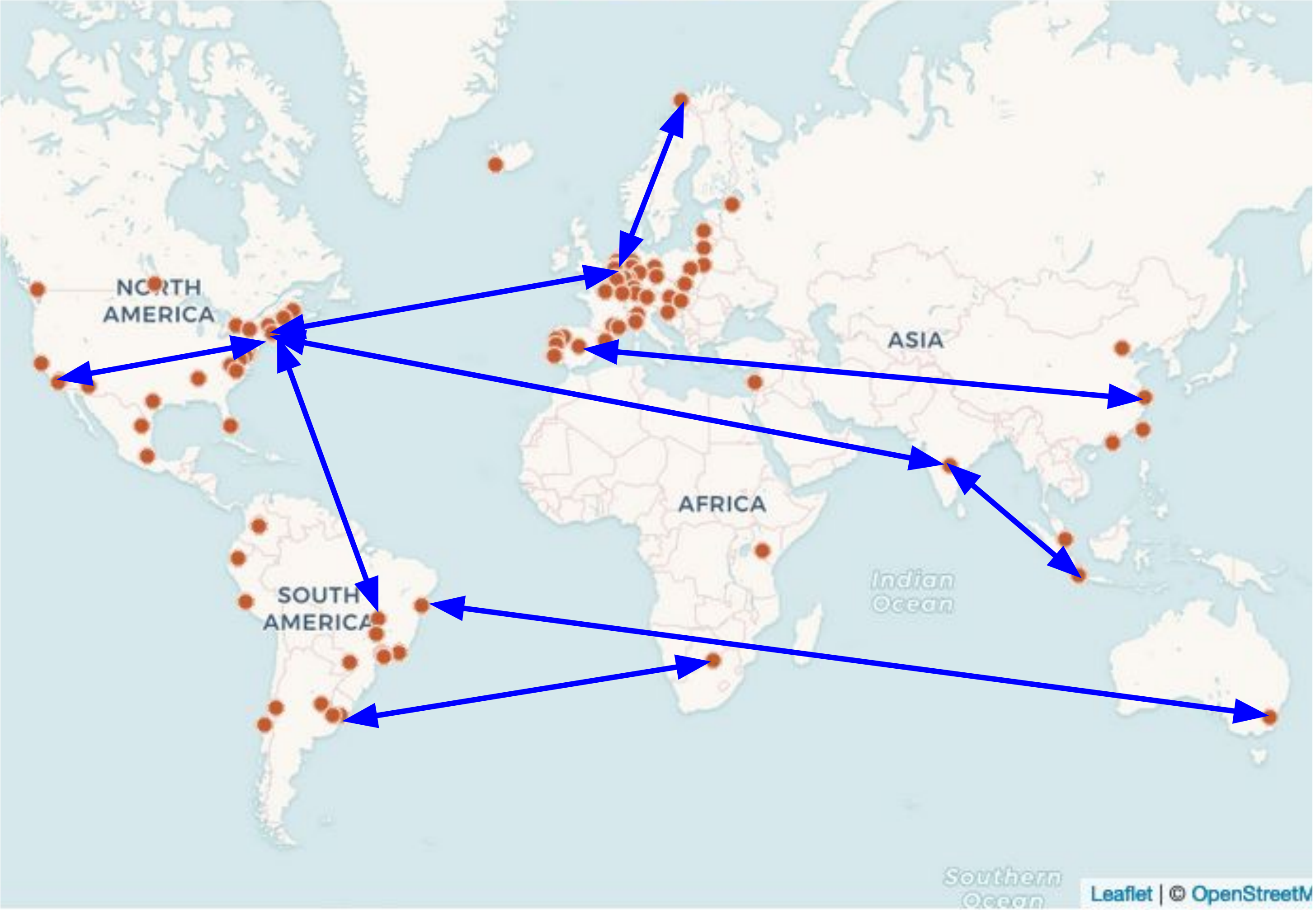
From the OAI Technical Committee:
Michael Nelson <m.l.nelson@larc.nasa.gov > -- NASA - Langley Research Center
Simeon Warner <simeon@cs.cornell.edu > -- Cornell University - Computer Science

Table of Contents

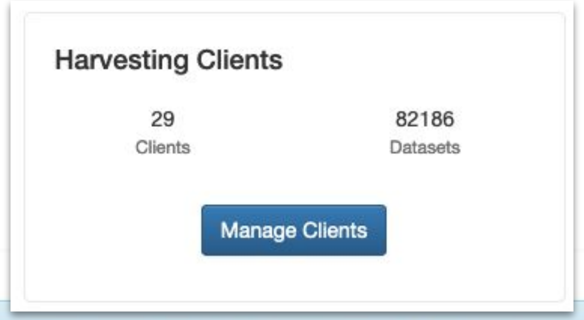
- 1. Introduction
- 2. Definitions and Concepts
  - 2.1. Harvester
  - 2.2. Repository
  - 2.3. Item
  - 2.4. Unique Identifier
  - 2.5. Record
    - 2.5.1 Deleted records
  - 2.6. Set
  - 2.7. Selective Harvesting
    - 2.7.1 Selective Harvesting and Datestamps
    - 2.7.2 Selective Harvesting and Sets
- 3. Protocol Features
  - 3.1. HTTP Embedding of OAI-PMH requests
    - 3.1.1. HTTP Request Format
    - 3.1.2. HTTP Response Format
    - 3.1.3. Response Compression
  - 3.2. XML Response Format
    - 3.2.1. XML Schema for Validating Responses to OAI-PMH Requests
  - 3.3. UTCdatetime
    - 3.3.1. UTCdatetime in Protocol Requests
    - 3.3.2. UTCdatetime in Protocol Responses
  - 3.4. metadataPrefix and Metadata Schema
  - 3.5. Flow Control
    - 3.5.1 Idempotency of resumptionTokens
  - 3.6. Error and Exception Conditions
- 4. Protocol Requests and Responses
  - 4.1. GetRecord
  - 4.2. Identify
  - 4.3. ListIdentifiers
  - 4.4. ListMetadataFormats
  - 4.5. ListRecords
  - 4.6. ListSets
- 5. Dublin Core
- 6. Implementation Guidelines
- Acknowledgements
- Document History

http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

116 Installations



Harvesting clients

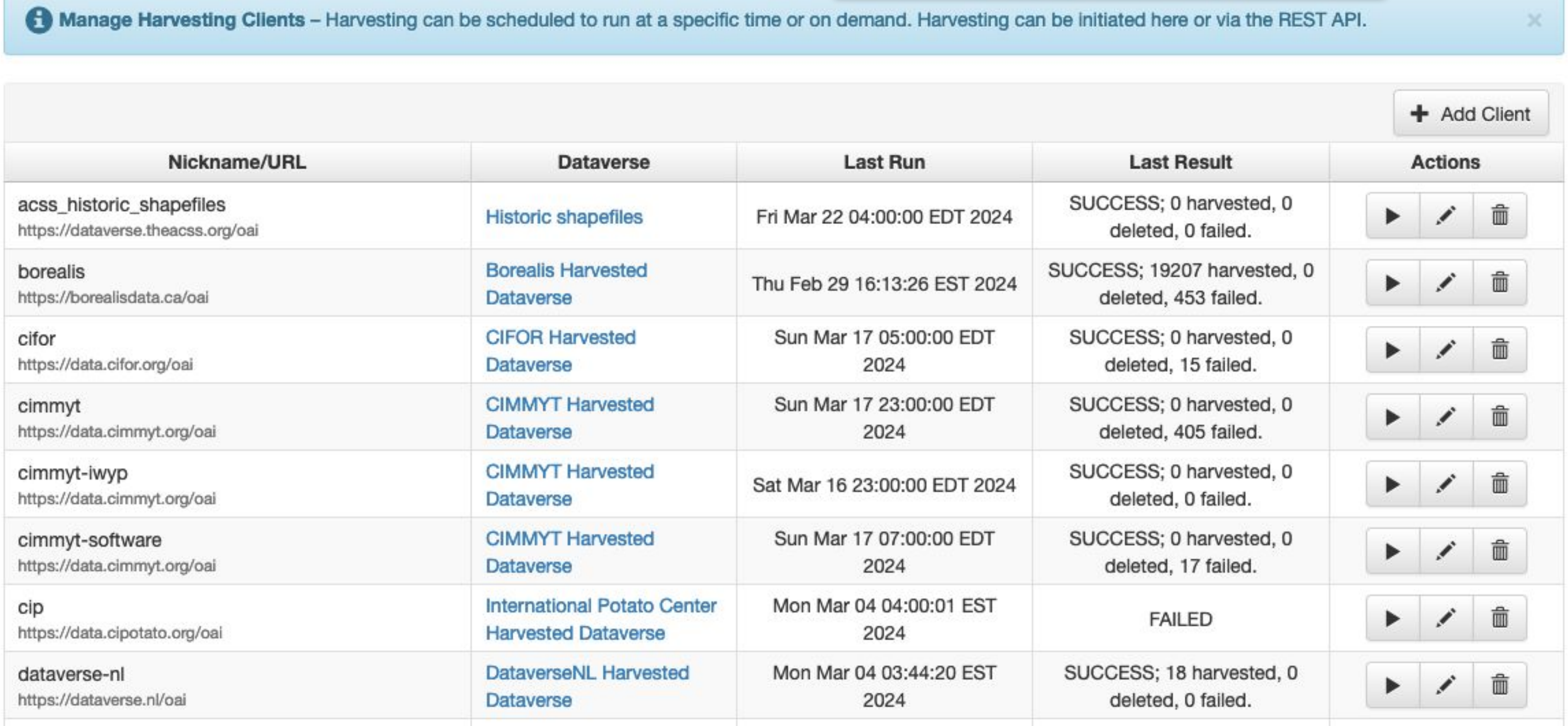


Harvesting Clients

29 Clients

82186 Datasets

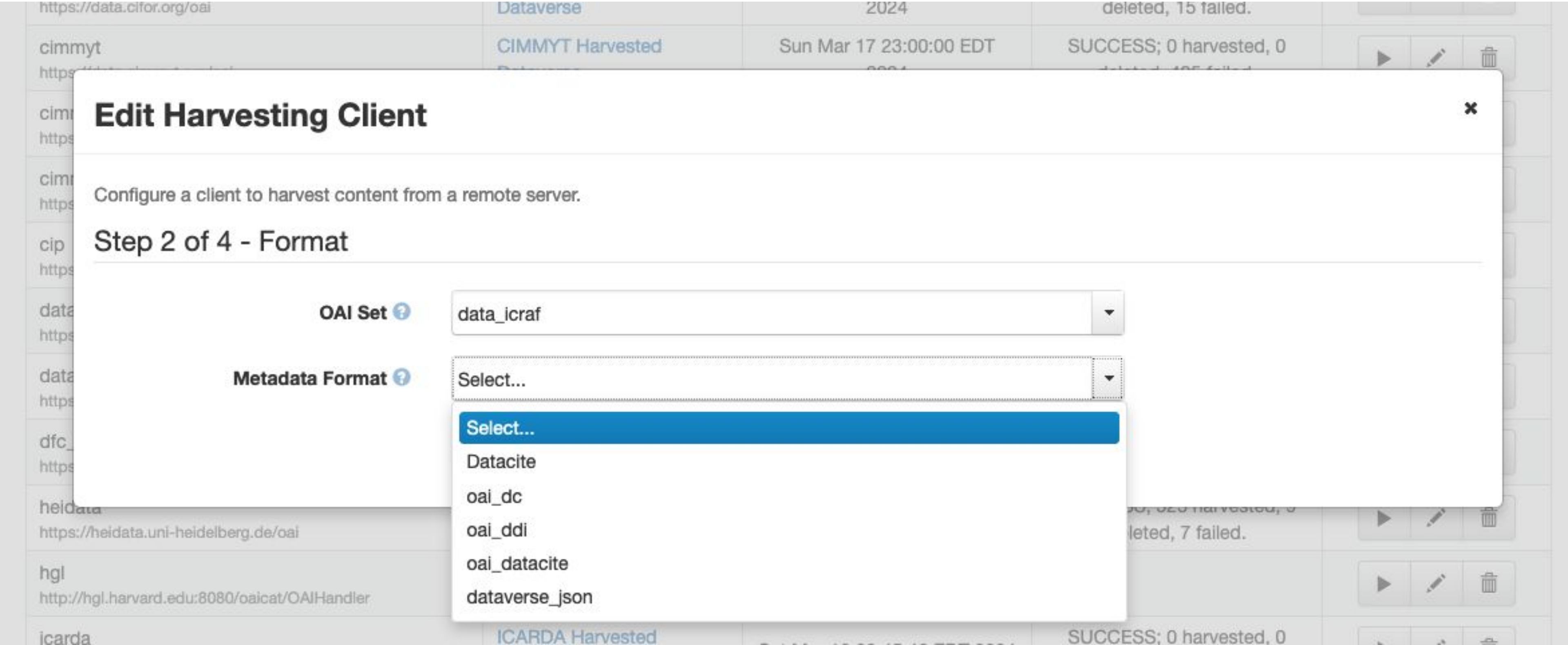
Harvard Dataverse > Dashboard >



Manage Harvesting Clients – Harvesting can be scheduled to run at a specific time or on demand. Harvesting can be initiated here or via the REST API.

Nickname/URL	Dataverse	Last Run	Last Result	Actions
acss_historic_shapefiles https://dataverse.theacss.org/oai	Historic shapefiles	Fri Mar 22 04:00:00 EDT 2024	SUCCESS; 0 harvested, 0 deleted, 0 failed.	▶ ✎ 🗑
borealis https://borealisdata.ca/oai	Borealis Harvested Dataverse	Thu Feb 29 16:13:26 EST 2024	SUCCESS; 19207 harvested, 0 deleted, 453 failed.	▶ ✎ 🗑
cifor https://data.cifor.org/oai	CIFOR Harvested Dataverse	Sun Mar 17 05:00:00 EDT 2024	SUCCESS; 0 harvested, 0 deleted, 15 failed.	▶ ✎ 🗑
cimmyt https://data.cimmyt.org/oai	CIMMYT Harvested Dataverse	Sun Mar 17 23:00:00 EDT 2024	SUCCESS; 0 harvested, 0 deleted, 405 failed.	▶ ✎ 🗑
cimmyt-iwyp https://data.cimmyt.org/oai	CIMMYT Harvested Dataverse	Sat Mar 16 23:00:00 EDT 2024	SUCCESS; 0 harvested, 0 deleted, 0 failed.	▶ ✎ 🗑
cimmyt-software https://data.cimmyt.org/oai	CIMMYT Harvested Dataverse	Sun Mar 17 07:00:00 EDT 2024	SUCCESS; 0 harvested, 0 deleted, 17 failed.	▶ ✎ 🗑
cip https://data.cipotato.org/oai	International Potato Center Harvested Dataverse	Mon Mar 04 04:00:01 EST 2024	FAILED	▶ ✎ 🗑
dataverse-nl https://dataverse.nl/oai	DataverseNL Harvested Dataverse	Mon Mar 04 03:44:20 EST 2024	SUCCESS; 18 harvested, 0 deleted, 0 failed.	▶ ✎ 🗑

Harvesting formats (Dublin Core, DDI, DataCite, etc.)



Edit Harvesting Client

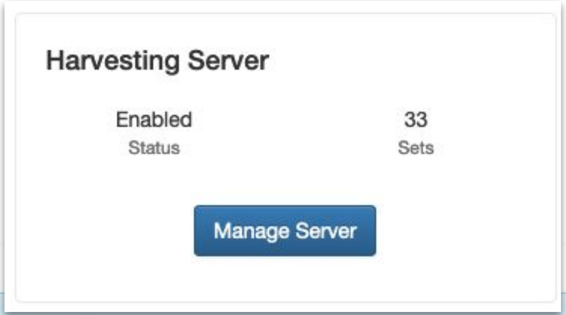
Configure a client to harvest content from a remote server.

Step 2 of 4 - Format

OAI Set: data_icraf

Metadata Format: Select...

Harvesting server (sets)

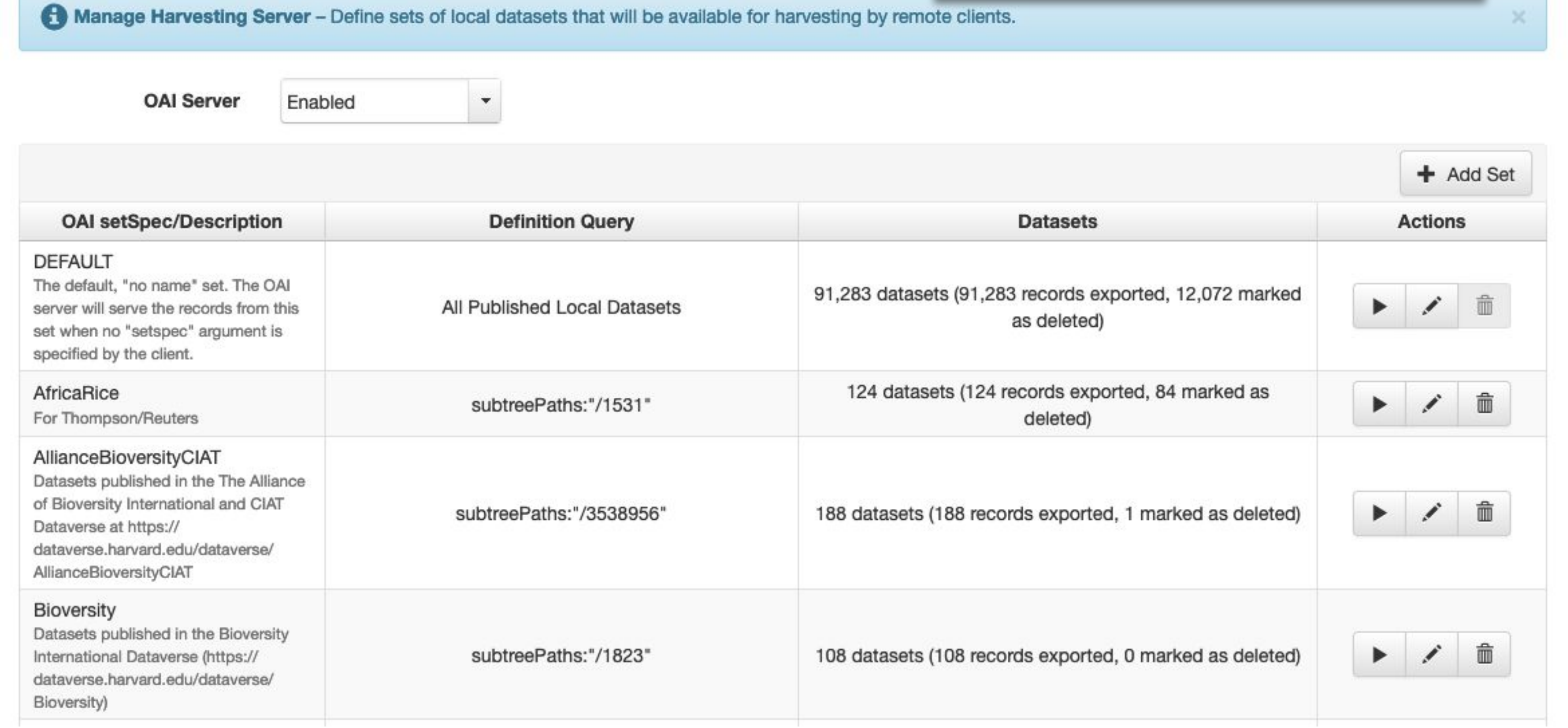


Harvesting Server

Enabled
Status

33
Sets

Harvard Dataverse > Dashboard >

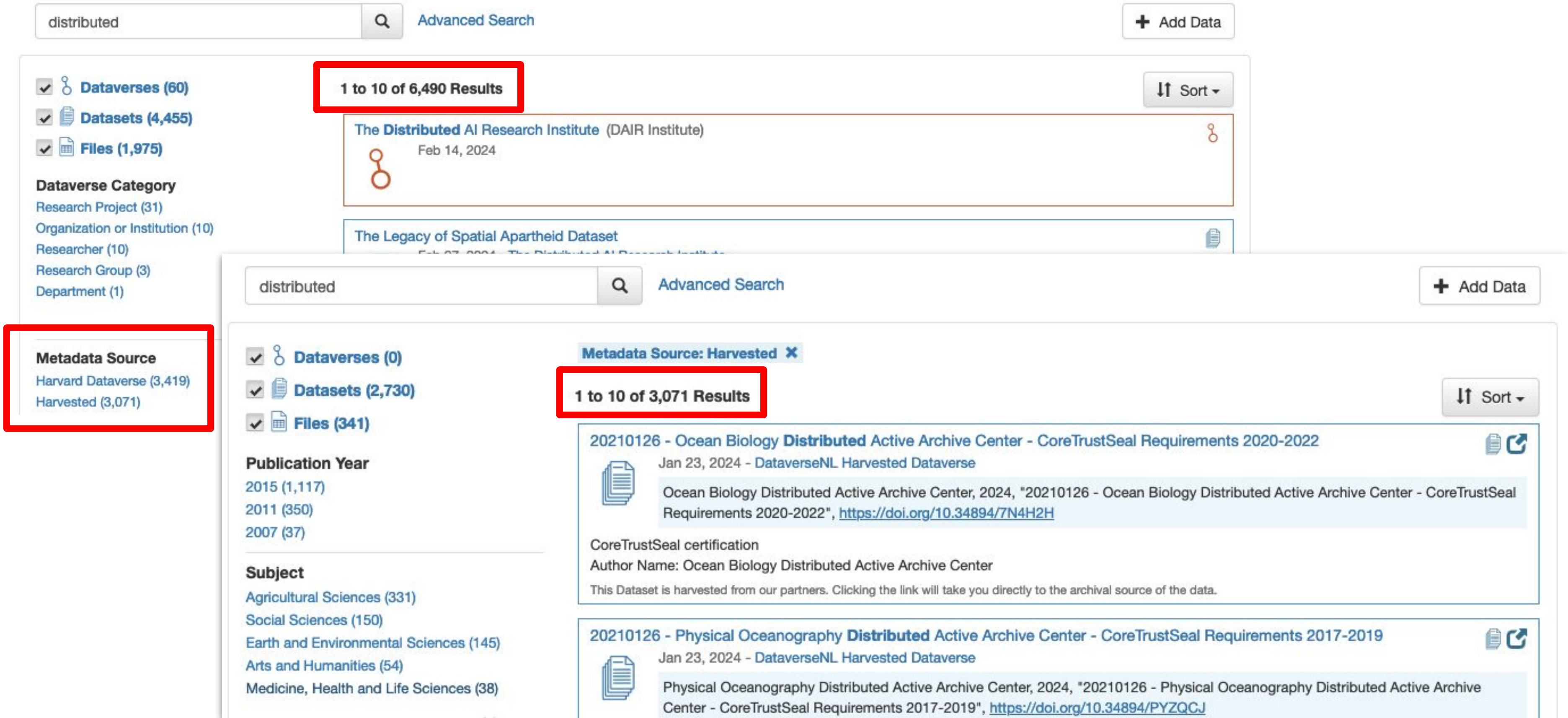


Manage Harvesting Server – Define sets of local datasets that will be available for harvesting by remote clients.

OAI Server: Enabled

OAI setSpec/Description	Definition Query	Datasets	Actions
DEFAULT The default, "no name" set. The OAI server will serve the records from this set when no "setspec" argument is specified by the client.	All Published Local Datasets	91,283 datasets (91,283 records exported, 12,072 marked as deleted)	▶ ✎ 🗑
AfricaRice For Thompson/Reuters	subtreePaths:"/1531"	124 datasets (124 records exported, 84 marked as deleted)	▶ ✎ 🗑
AllianceBioversityCIAT Datasets published in the The Alliance of Bioversity International and CIAT Dataverse at https://dataverse.harvard.edu/dataverse/AllianceBioversityCIAT	subtreePaths:"/3538956"	188 datasets (188 records exported, 1 marked as deleted)	▶ ✎ 🗑
Bioversity Datasets published in the Bioversity International Dataverse (https://dataverse.harvard.edu/dataverse/Bioversity)	subtreePaths:"/1823"	108 datasets (108 records exported, 0 marked as deleted)	▶ ✎ 🗑

More results through harvesting



☑ Dataverses (60)
☑ Datasets (4,455)
☑ Files (1,975)

Dataverse Category

Research Project (31)
Organization or Institution (10)
Researcher (10)
Research Group (3)
Department (1)

Metadata Source

Harvard Dataverse (3,419)
Harvested (3,071)

1 to 10 of 6,490 Results

The **Distributed** AI Research Institute (DAIR Institute)
Feb 14, 2024

The Legacy of Spatial Apartheid Dataset

☑ Dataverses (0)
☑ Datasets (2,730)
☑ Files (341)

Publication Year

2015 (1,117)
2011 (350)
2007 (37)

Subject

Agricultural Sciences (331)
Social Sciences (150)
Earth and Environmental Sciences (145)
Arts and Humanities (54)
Medicine, Health and Life Sciences (38)

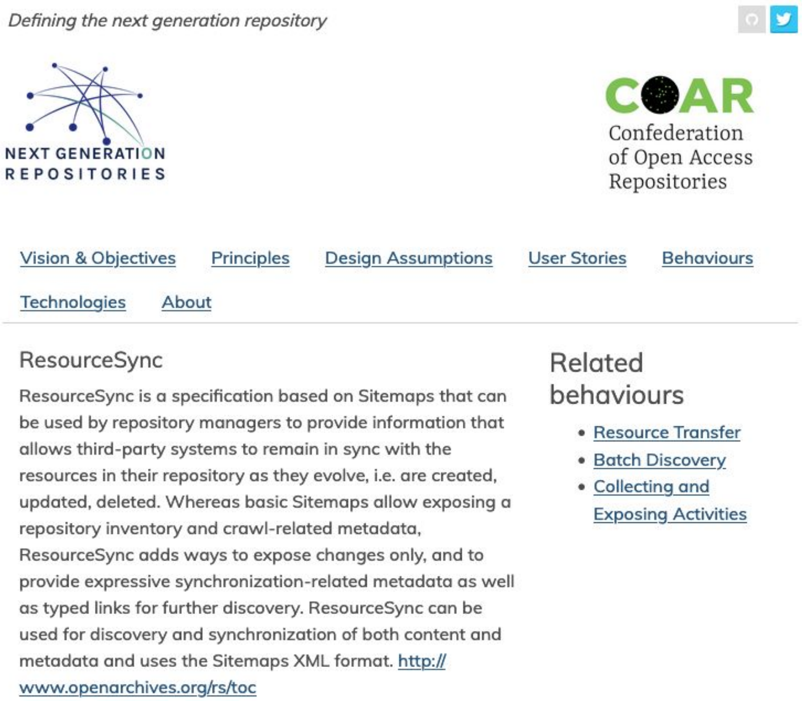
Metadata Source: Harvested ✖

1 to 10 of 3,071 Results

20210126 - Ocean Biology **Distributed** Active Archive Center - CoreTrustSeal Requirements 2020-2022
Jan 23, 2024 - DataverseNL Harvested Dataverse
Ocean Biology Distributed Active Archive Center, 2024, "20210126 - Ocean Biology Distributed Active Archive Center - CoreTrustSeal Requirements 2020-2022", https://doi.org/10.34894/7N4H2H
CoreTrustSeal certification
Author Name: Ocean Biology Distributed Active Archive Center
This Dataset is harvested from our partners. Clicking the link will take you directly to the archival source of the data.

20210126 - Physical Oceanography **Distributed** Active Archive Center - CoreTrustSeal Requirements 2017-2019
Jan 23, 2024 - DataverseNL Harvested Dataverse
Physical Oceanography Distributed Active Archive Center, 2024, "20210126 - Physical Oceanography Distributed Active Archive Center - CoreTrustSeal Requirements 2017-2019", https://doi.org/10.34894/PYZQCJ

2002? Anything newer?



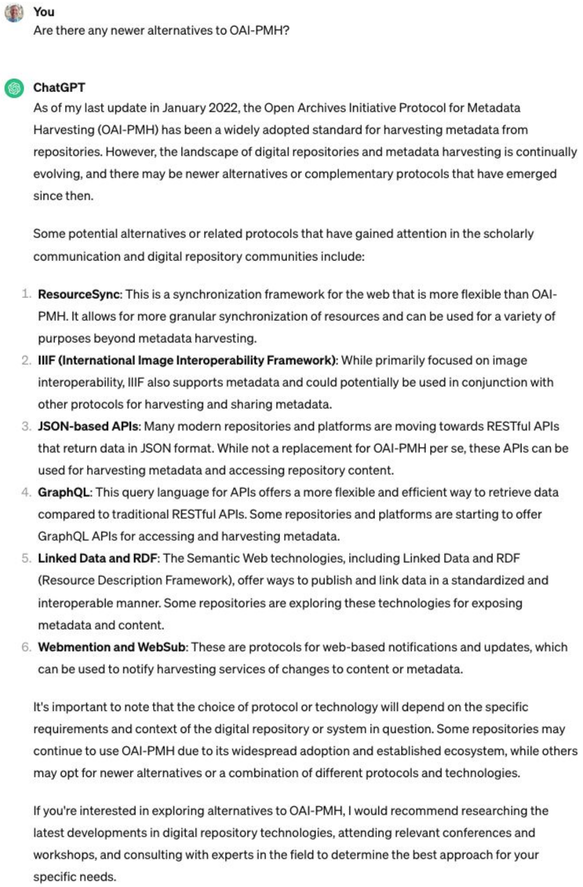
Defining the next generation repository

ResourceSync

ResourceSync is a specification based on Sitemaps that can be used by repository managers to provide information that allows third-party systems to remain in sync with the resources in their repository as they evolve, i.e. are created, updated, deleted. Whereas basic Sitemaps allow exposing a repository inventory and crawl-related metadata, ResourceSync adds ways to expose changes only, and to provide expressive synchronization-related metadata as well as typed links for further discovery. ResourceSync can be used for discovery and synchronization of both content and metadata and uses the Sitemaps XML format. http://www.openarchives.org/rs/toc

Related behaviours

- Resource Transfer
- Batch Discovery
- Collecting and Exposing Activities



You
Are there any newer alternatives to OAI-PMH?

ChatGPT
As of my last update in January 2022, the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) has been a widely adopted standard for harvesting metadata from repositories. However, the landscape of digital repositories and metadata harvesting is continually evolving, and there may be newer alternatives or complementary protocols that have emerged since then.

Some potential alternatives or related protocols that have gained attention in the scholarly communication and digital repository communities include:

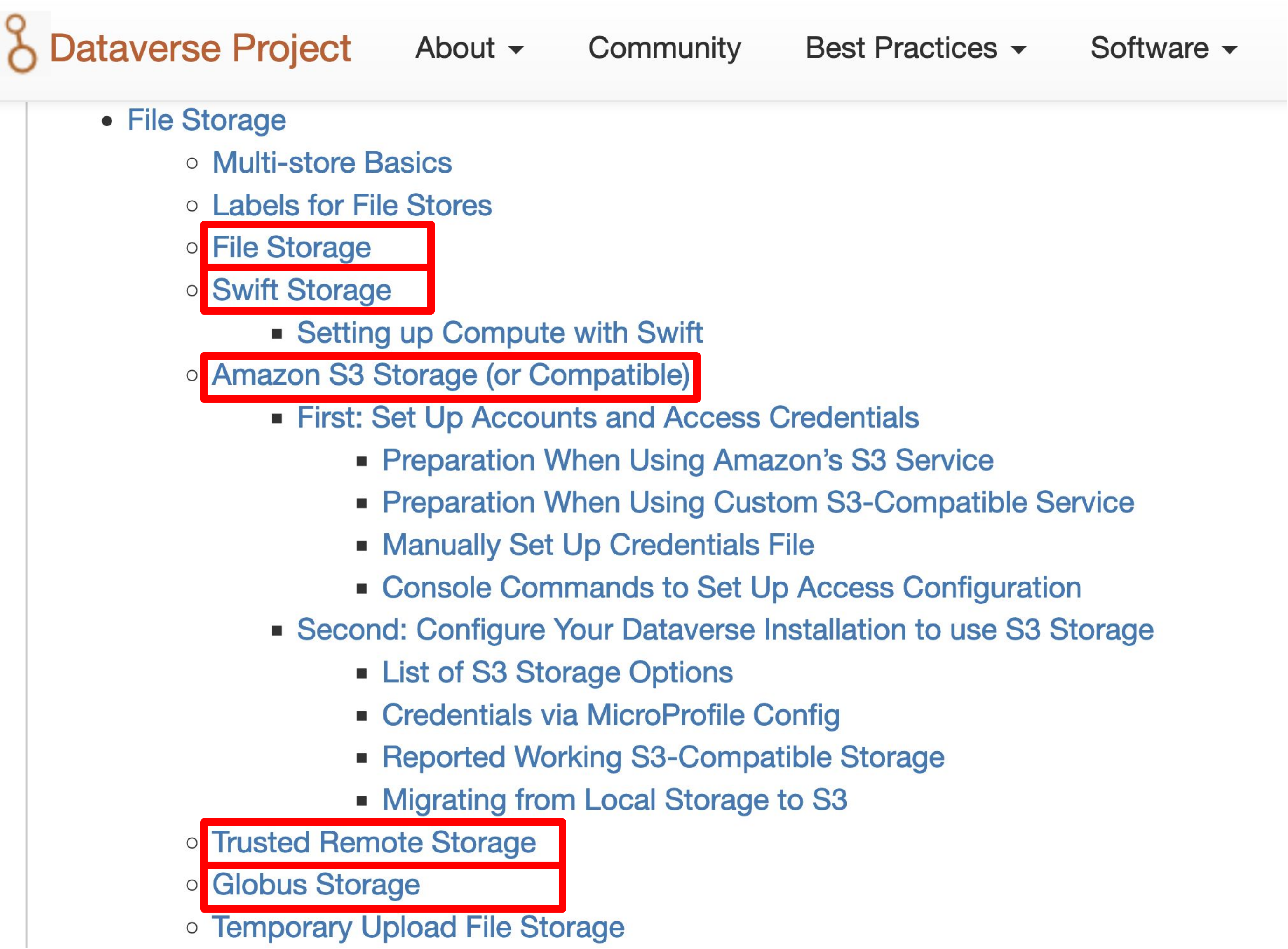
1. **ResourceSync**: This is a synchronization framework for the web that is more flexible than OAI-PMH. It allows for more granular synchronization of resources and can be used for a variety of purposes beyond metadata harvesting.
2. **IIIF (International Image Interoperability Framework)**: While primarily focused on image interoperability, IIIF also supports metadata and could potentially be used in conjunction with other protocols for harvesting and sharing metadata.
3. **JSON-based APIs**: Many modern repositories and platforms are moving towards RESTful APIs that return data in JSON format. While not a replacement for OAI-PMH per se, these APIs can be used for harvesting metadata and accessing repository content.
4. **GraphQL**: This query language for APIs offers a more flexible and efficient way to retrieve data compared to traditional RESTful APIs. Some repositories and platforms are starting to offer GraphQL APIs for accessing and harvesting metadata.
5. **Linked Data and RDF**: The Semantic Web technologies, including Linked Data and RDF (Resource Description Framework), offer ways to publish and link data in a standardized and interoperable manner. Some repositories are exploring these technologies for exposing metadata and content.
6. **Webmention and WebSub**: These are protocols for web-based notifications and updates, which can be used to notify harvesting services of changes to content or metadata.

It's important to note that the choice of protocol or technology will depend on the specific requirements and context of the digital repository or system in question. Some repositories may continue to use OAI-PMH due to its widespread adoption and established ecosystem, while others may opt for newer alternatives or a combination of different protocols and technologies.

If you're interested in exploring alternatives to OAI-PMH, I would recommend researching the latest developments in digital repository technologies, attending relevant conferences and workshops, and consulting with experts in the field to determine the best approach for your specific needs.

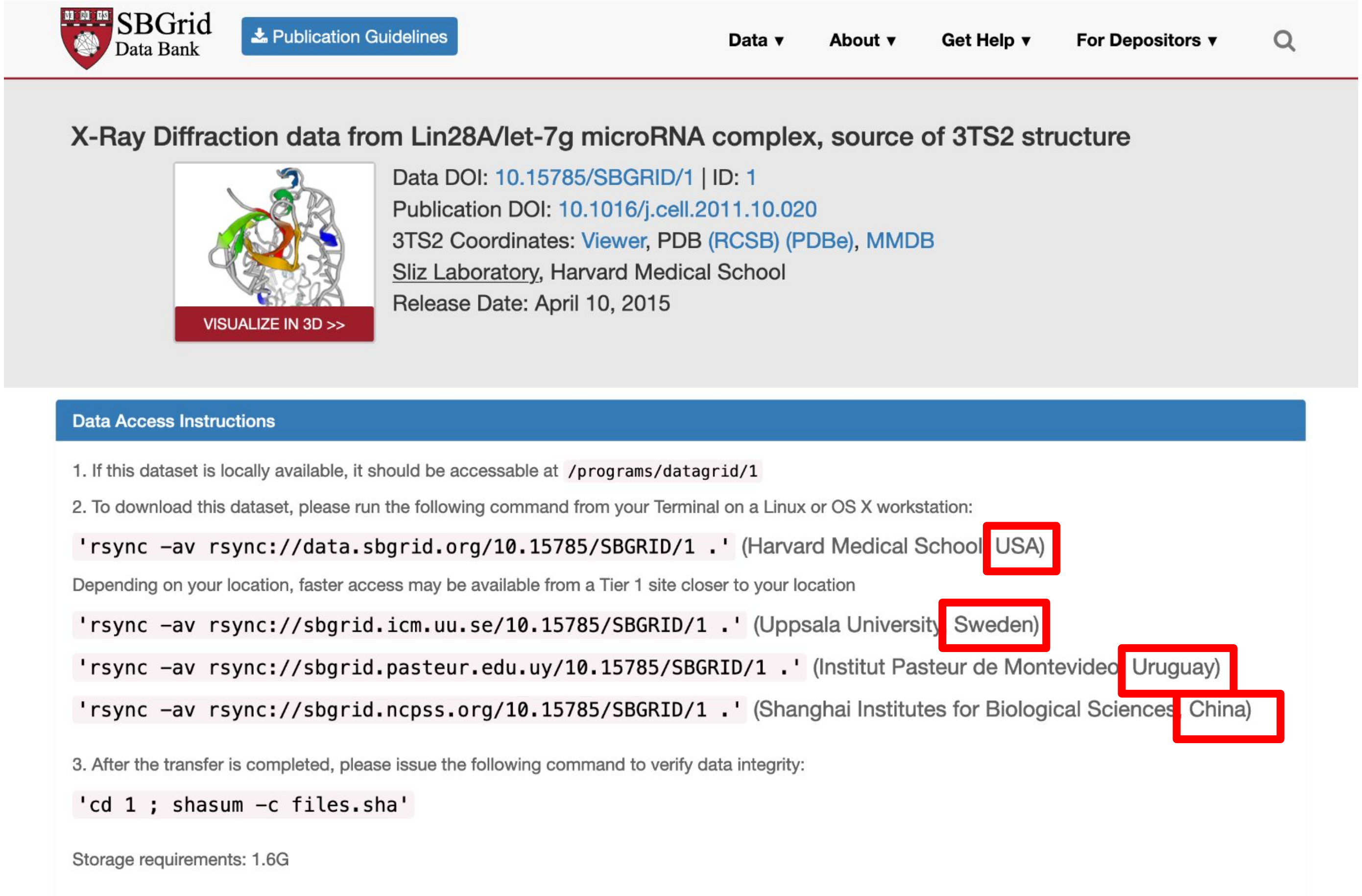
https://ngr.coar-repositories.org/technology/resourcesync/

Distributed Data



- File Storage
    - Multi-store Basics
    - Labels for File Stores
    - File Storage
    - Swift Storage
        - Setting up Compute with Swift
    - Amazon S3 Storage (or Compatible)
        - First: Set Up Accounts and Access Credentials
            - Preparation When Using Amazon’s S3 Service
            - Preparation When Using Custom S3-Compatible Service
            - Manually Set Up Credentials File
            - Console Commands to Set Up Access Configuration
        - Second: Configure Your Dataverse Installation to use S3 Storage
            - List of S3 Storage Options
            - Credentials via MicroProfile Config
            - Reported Working S3-Compatible Storage
            - Migrating from Local Storage to S3
    - Trusted Remote Storage
    - Globus Storage
    - Temporary Upload File Storage

https://guides.dataverse.org/en/6.2/installation/config.html#file-storage



X-Ray Diffraction data from Lin28A/let-7g microRNA complex, source of 3TS2 structure

Data DOI: 10.15785/SBGRID/1 | ID: 1
Publication DOI: 10.1016/j.cell.2011.10.020
3TS2 Coordinates: Viewer, PDB (RCSB) (PDBe), MMDB
Sliz Laboratory, Harvard Medical School
Release Date: April 10, 2015

Data Access Instructions

1. If this dataset is locally available, it should be accessable at `/programs/datagrid/1`
2. To download this dataset, please run the following command from your Terminal on a Linux or OS X workstation:

`'rsync -av rsync://data.sbgrid.org/10.15785/SBGRID/1 .'` (Harvard Medical School USA)

Depending on your location, faster access may be available from a Tier 1 site closer to your location

`'rsync -av rsync://sbgrid.icm.uu.se/10.15785/SBGRID/1 .'` (Uppsala University Sweden)

`'rsync -av rsync://sbgrid.pasteur.edu.uy/10.15785/SBGRID/1 .'` (Institut Pasteur de Montevideo Uruguay)

`'rsync -av rsync://sbgrid.ncpss.org/10.15785/SBGRID/1 .'` (Shanghai Institutes for Biological Sciences China)

3. After the transfer is completed, please issue the following command to verify data integrity:

`'cd 1 ; shasum -c files.sha'`

Storage requirements: 1.6G

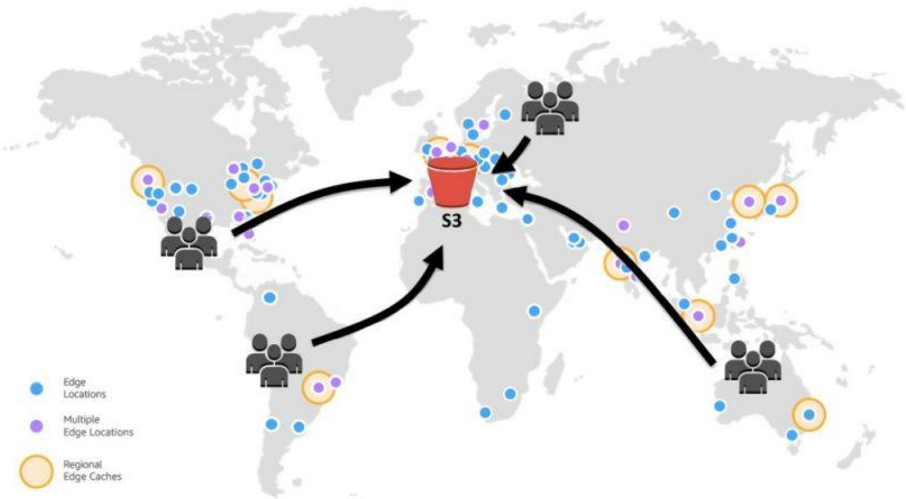
https://data.sbgrid.org/dataset/1

Dreams of distributed data

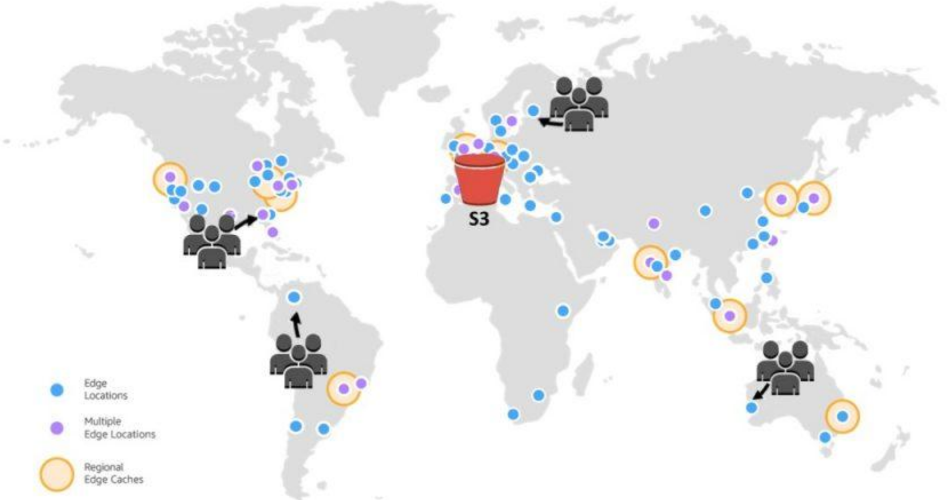
"An idea of: synchronizing the data between the Harvard Dataverse and the soon-to-be-up [Foobar Dataverse] (and possibly other major locations) to make a connected dataverse. I believe that it would enhance the usability of the system for the global research community and possibly distribute the traffic better."

Amazon CloudFront

before

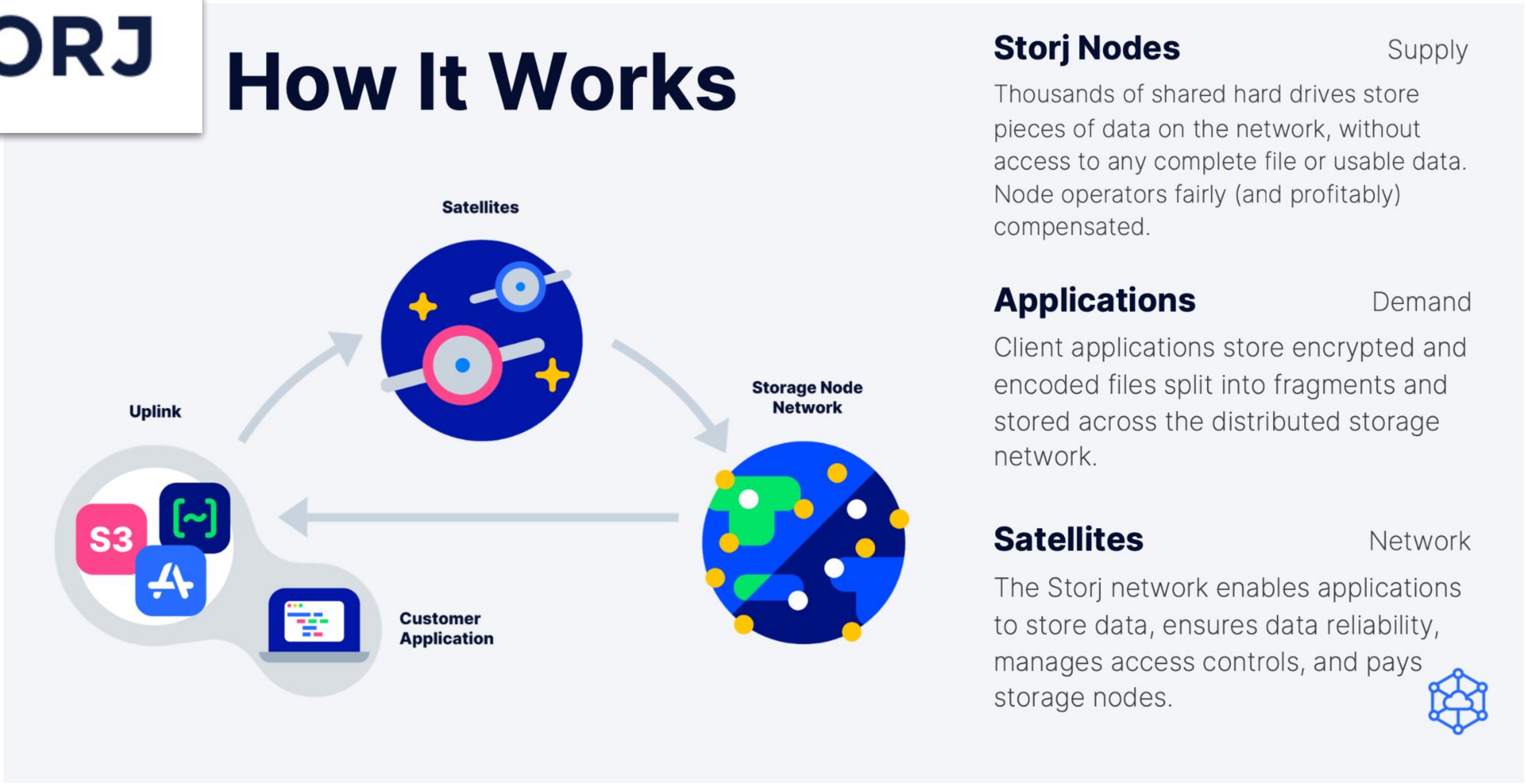


after



https://aws.amazon.com/blogs/networking-and-content-delivery/amazon-s3-amazon-cloudfront-a-match-made-in-the-cloud/

STORJ



How It Works

Storj Nodes Supply

Thousands of shared hard drives store pieces of data on the network, without access to any complete file or usable data. Node operators fairly (and profitably) compensated.

Applications Demand

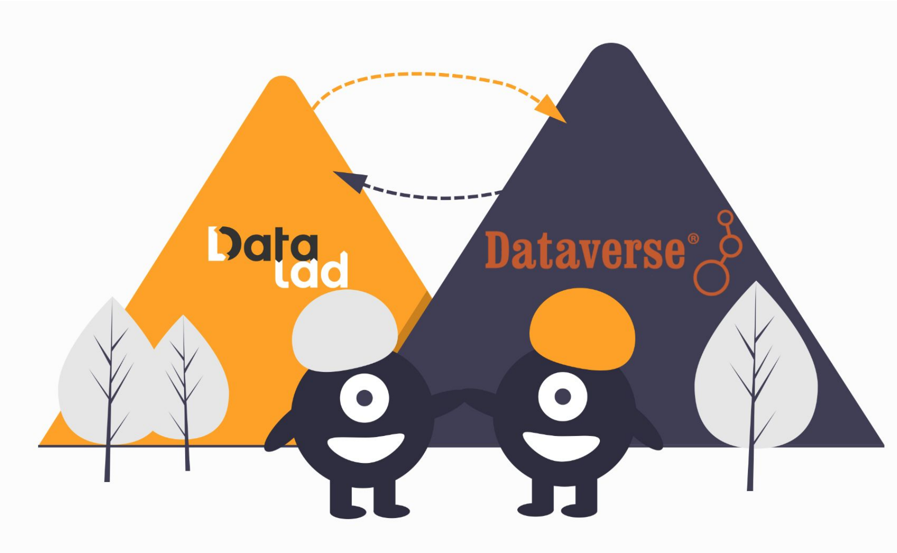
Client applications store encrypted and encoded files split into fragments and stored across the distributed storage network.

Satellites Network

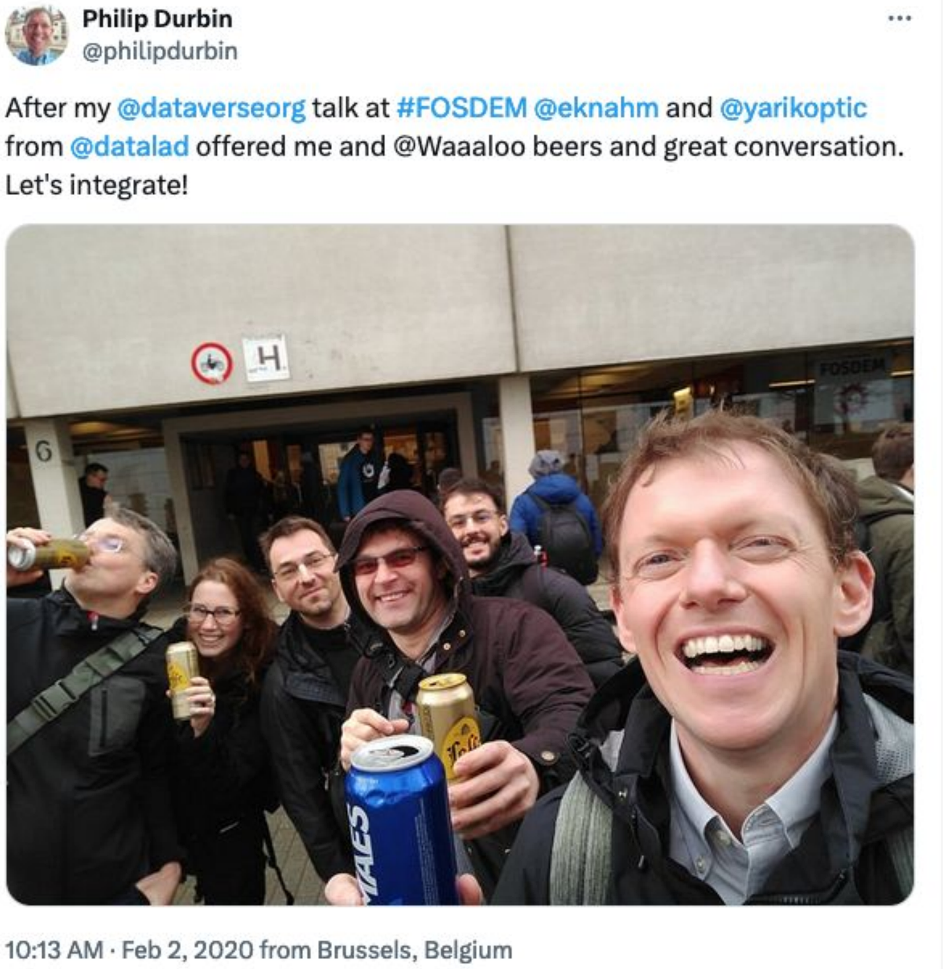
The Storj network enables applications to store data, ensures data reliability, manages access controls, and pays storage nodes.

https://osf.io/3txy4
https://www.youtube.com/watch?v=3ek7F_Dxcjk&t=2102s

Dataverse and DataLad



https://docs.datalad.org/projects/dataverse



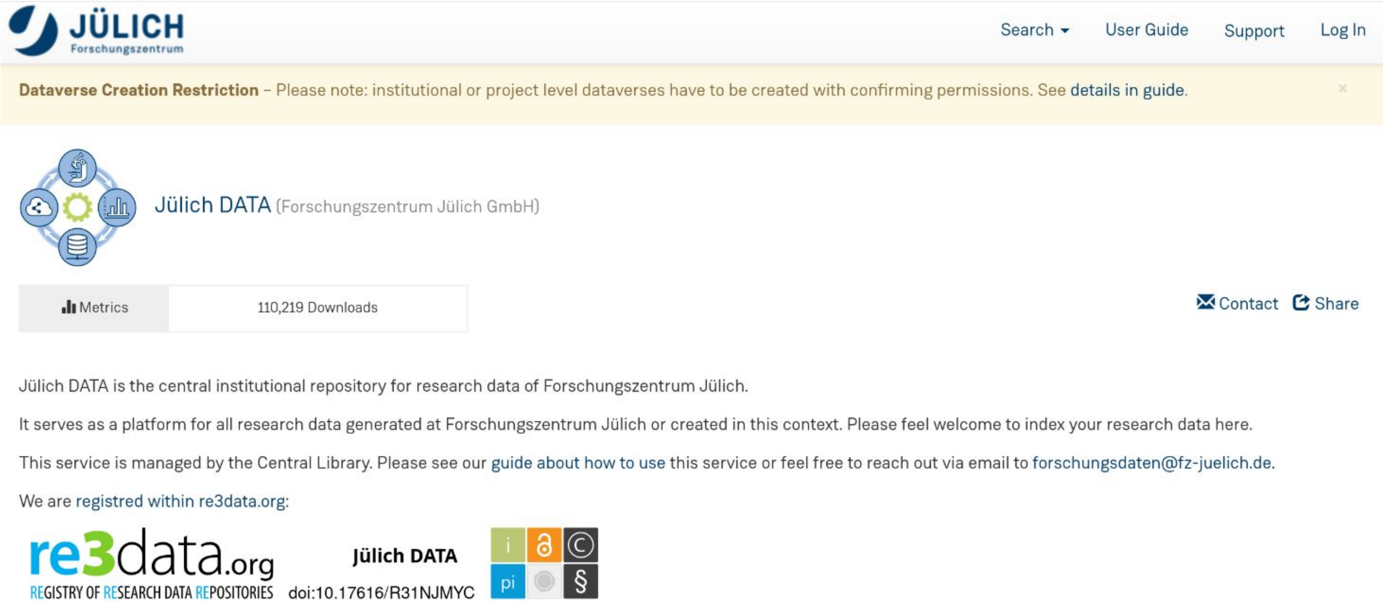
Philip Durbin
@philipdurbin

After my @dataverseorg talk at #FOSDEM @eknahm and @yarikoptic from @datalad offered me and @Waaaloo beers and great conversation. Let's integrate!

10:13 AM · Feb 2, 2020 from Brussels, Belgium

Future architecture for a mixed-style storage configuration in Jülich DATA

What is Jülich DATA (now)?



Dataverse Creation Restriction – Please note: institutional or project level dataverses have to be created with confirming permissions. See details in guide.

Jülich DATA (Forschungszentrum Jülich GmbH)

Jülich DATA is the central institutional repository for research data of Forschungszentrum Jülich.

It serves as a platform for all research data generated at Forschungszentrum Jülich or created in this context. Please feel welcome to index your research data here.

This service is managed by the Central Library. Please see our guide about how to use this service or feel free to reach out via email to forschungsdaten@fz-juelich.de.

We are registred within re3data.org:

- Status quo: institutional data registry since 2020
- Hosting Central Library, Open Science Team for FZJ (~2.5k researchers)
- Zero storage policy / ideal

Publication personas: current use cases

Alice

Uploads to and publishes in external repository

Registers in Jülich DATA with added institutional metadata

Main reason: reporting for **€€€** from Helmholtz

Bobby

Metadata-only publication in Jülich DATA

Data is referenced only for valid reasons (privacy, size, policy) using a URI to storage location

Devin

In denial, does not publish anything

¯_(ツ)_/¯

Lessons learned

- Only 150 (published) datasets?!
- Metadata input twice (publication and registration) makes users unhappy

- Open Science Team is well established addressee for RDM questions
- We need to turn people away looking for publication options, especially >300GB and no/small community repos
- Chicken & Egg: “No easy solution available? Build our own or don’t publish.”

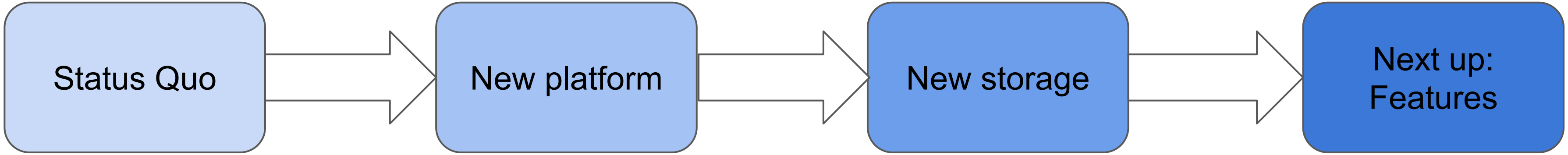
- Discoverability of files for metadata-only publications is zero
- FZJ joined SF-DORA in 2023, we want software publications, too

Infrastructure Updates

Infrastructure Update Challenges

- Kubernetes platform
- Built in 2020
- Optimized for registry

- How much storage?
- Structured interviews w/ institute data stewards



Status Quo

New platform

New storage

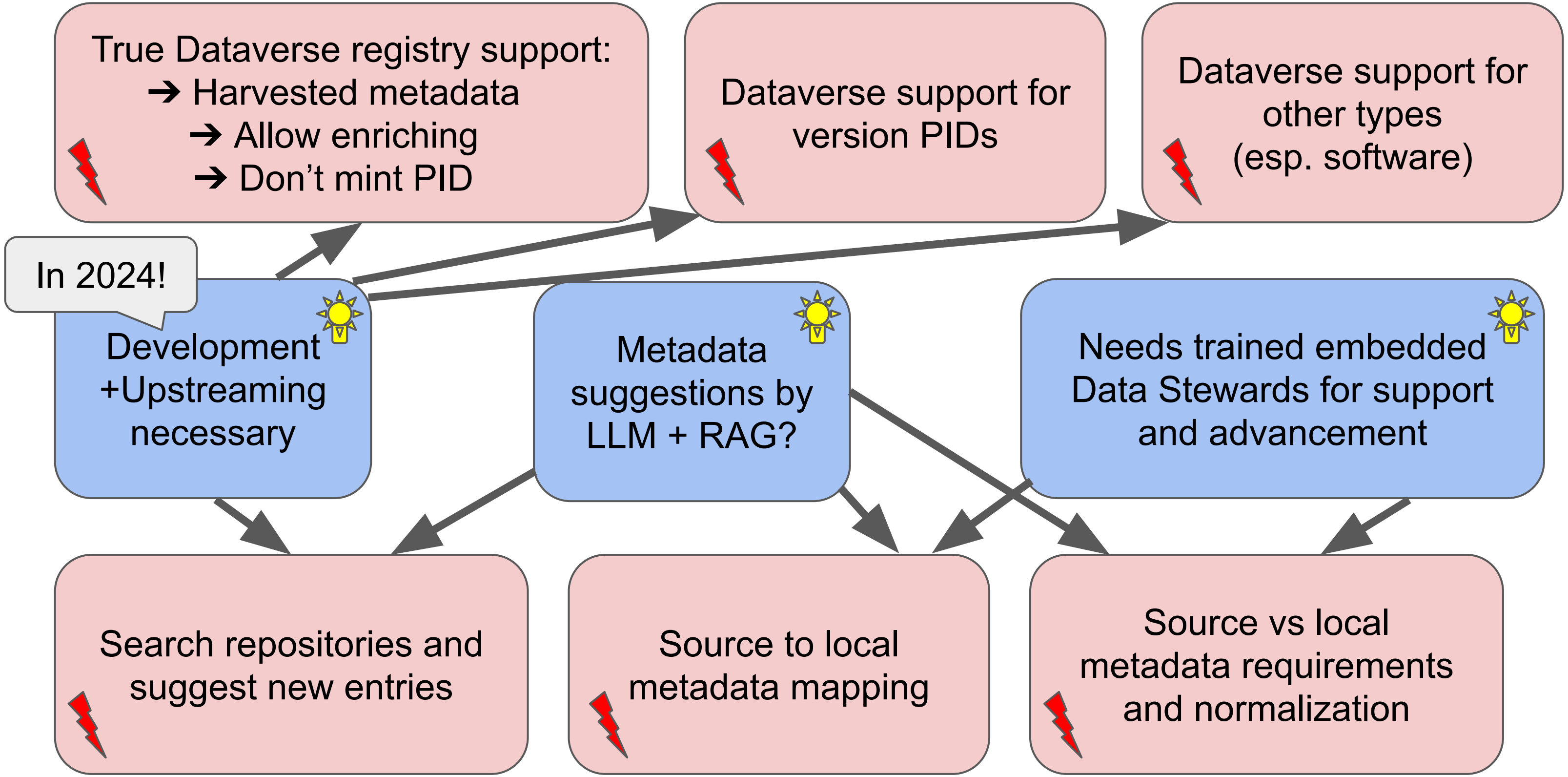
Next up:
Features

- Terraform new K8s
- Migrate to contributed upstream containers
- Contribute K8s operator for management

- T.B.D. by autumn 2024
- Onto more adventures!

Feature Updates

Challenges to support Alice (Registration)



True Dataverse registry support:
- ➔ Harvested metadata
  - ➔ Allow enriching
  - ➔ Don't mint PID

Dataverse support for version PIDs

Dataverse support for other types (esp. software)

In 2024!

Development +Upstreaming necessary

Metadata suggestions by LLM + RAG?

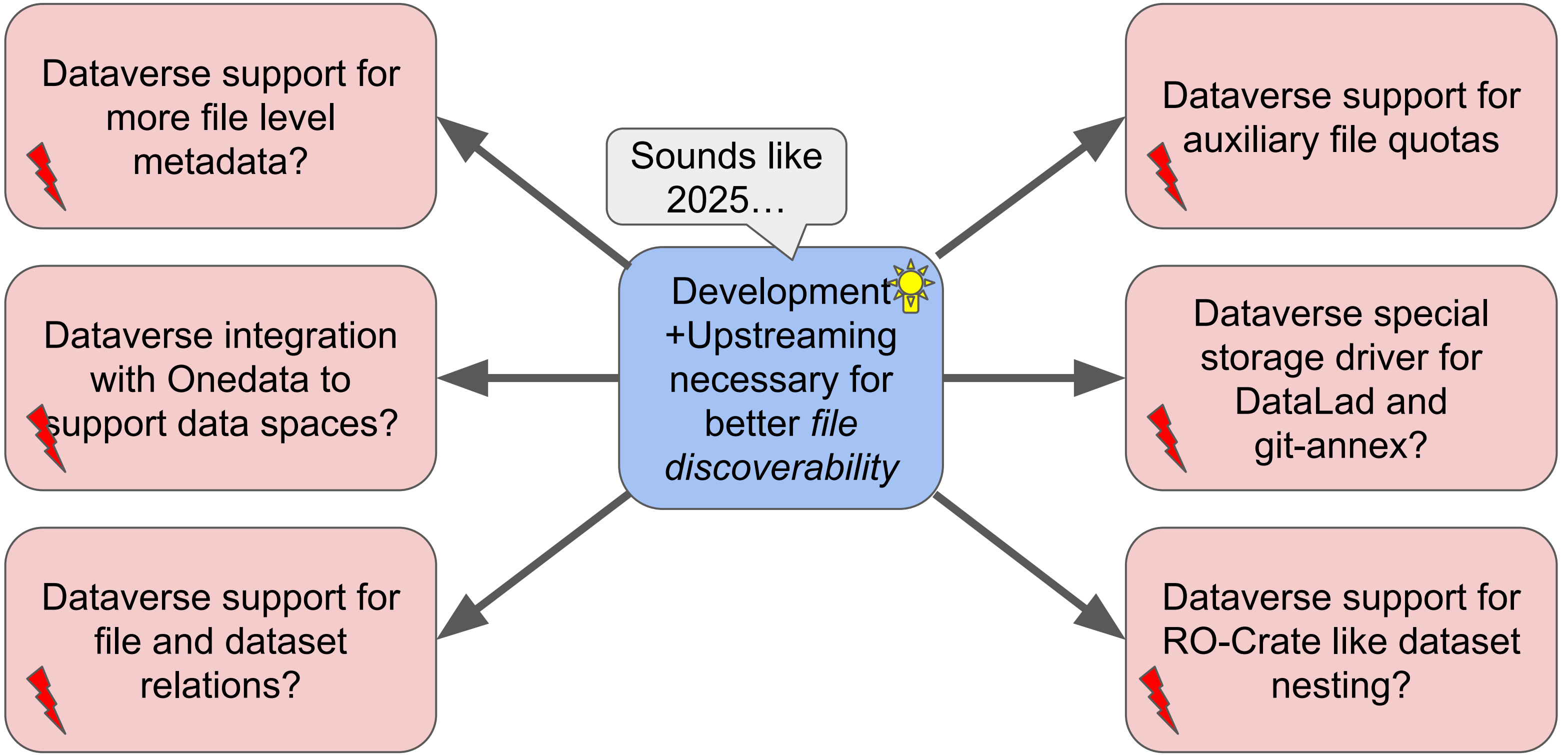
Needs trained embedded Data Stewards for support and advancement

Search repositories and suggest new entries

Source to local metadata mapping

Source vs local metadata requirements and normalization

Challenges to support Bobby (mostly Metadata Publication)



Sounds like 2025…

Development +Upstreaming necessary for better *file discoverability*

Dataverse support for more file level metadata?

Dataverse integration with Onedata to support data spaces?

Dataverse support for file and dataset relations?

Dataverse support for auxiliary file quotas

Dataverse special storage driver for DataLad and git-annex?

Dataverse support for RO-Crate like dataset nesting?

Publication personas: adding a third use case

Alice

Uploads to and publishes in community repository

Registers in Jülich DATA

Bobby

Metadata publication in Jülich DATA

(Most) data is referenced only for valid reasons (privacy, size, policy)

Charlie

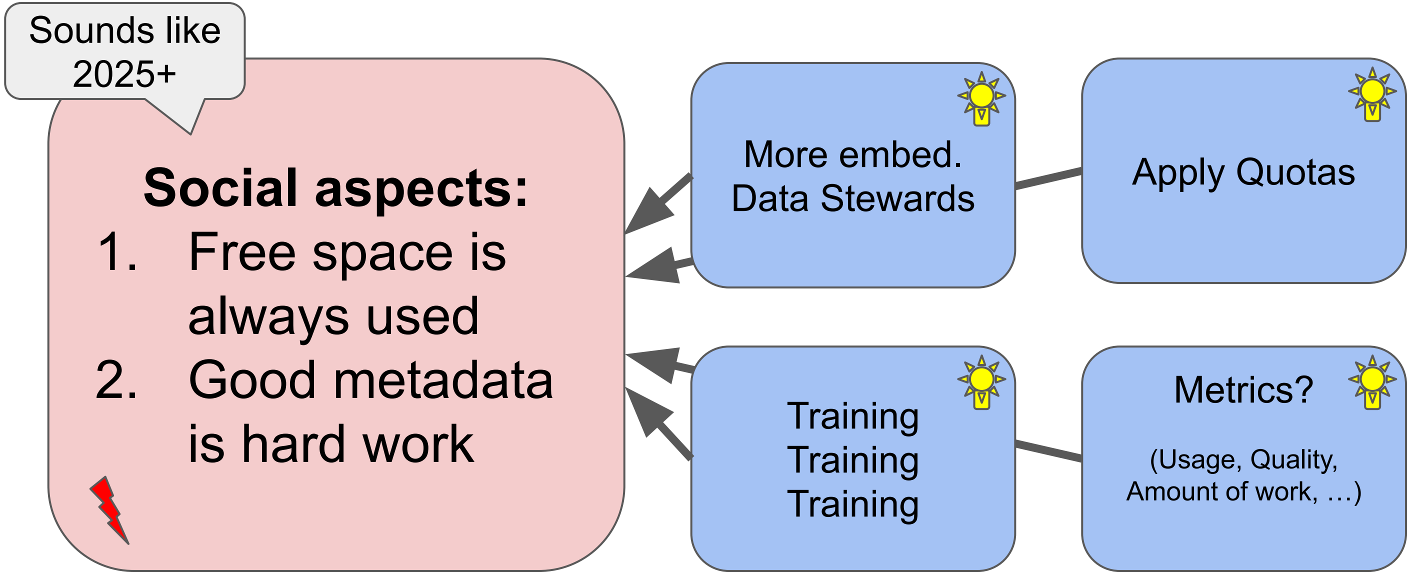
Uploads to and publishes using Jülich DATA

Devin

In denial, does not publish anything

¯_(ツ)_/¯

Challenges for Charlie (Publication)



Sounds like 2025+

Social aspects:

1. Free space is always used
2. Good metadata is hard work

More embed. Data Stewards

Apply Quotas

Training Training Training

Metrics?

(Usage, Quality, Amount of work, …)

Computation next to data?

Control institute archive storage?

Storage tiering with tape libs?

Globus integration? (€€€€€€)

University of Stuttgart
Cluster of Excellence „Data-integrated Simulation Sciences“

SimTech

Use of Dataverse APIs to manage data with Python

Jan Range, Philip Durbin, Oliver Bertuch | Distribits| 05-04-2024

About PyDataverse

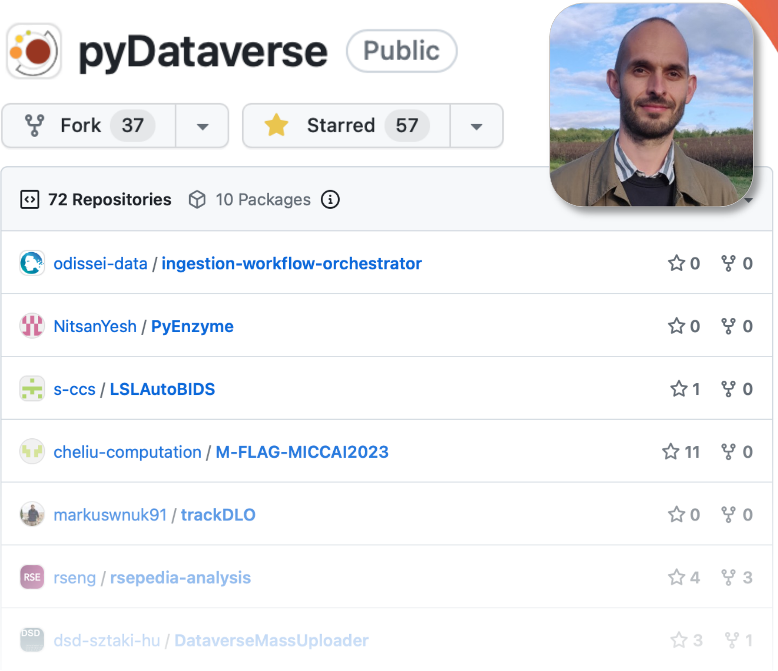
Overview

Python library to interface Dataverse

- Developed by Stefan Kasberger (until ~2021)
- Implements a subset of essential Dataverse endpoints

Popular among Datanauts

- 59 stars on GitHub (best on GDCC)
- 72 repositories utilize pyDataverse



pyDataverse Public

Fork 37 | Starred 57

72 Repositories 10 Packages

odissei-data / ingestion-workflow-orchestrator ☆ 0 forks 0
NitsanYesh / PyEnzyme ☆ 0 forks 0
s-ccs / LSLAutoBIDS ☆ 1 forks 0
cheliu-computation / M-FLAG-MICCAI2023 ☆ 11 forks 0
markuswnuk91 / trackDLO ☆ 0 forks 0
rseng / rsepedia-analysis ☆ 4 forks 3
dsd-sztaki-hu / DataverseMassUploader ☆ 3 forks 1

About PyDataverse

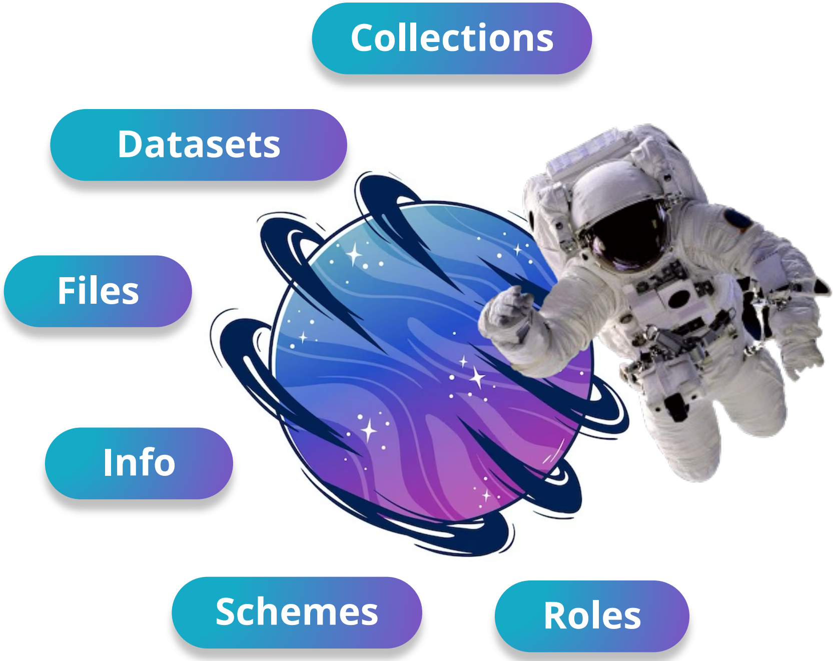
Overview

Low-level API for Dataverse

- Direct implementation of endpoints
- Maximum control and flexibility
- Supports async requests

Essentials for Datanauts

- Global management and infos
- CRUD operations for any element



Collections

Datasets

Files

Info

Schemes

Roles

Caveat of PyDataverse

Low-level to the core

Requires deeper knowledge

- Understanding of the native API
- Combination of multiple endpoints for high-level concepts (e.g. Dataset handling)

Example: Dataset Creation

- Requires two endpoints to be called
- Complex Dataverse JSON to add metadata

```
{
  "metadataBlocks": {
    "citation": {
      "displayName": "Citation Metadata",
      "name": "citation",
      "fields": [
        {
          "typeName": "title",
          "multiple": false,
          "typeClass": "primitive",
          "value": "My dataset"
        },
        {
          "typeName": "author",
          "multiple": true,
          "typeClass": "compound",
          "value": [
            {
              "authorName": {
                "typeName": "authorName",
                "multiple": false,
                "typeClass": "primitive",
                "value": "John Doe"
              }
            }
          ]
        }]}}}
```

Caveat of PyDataverse

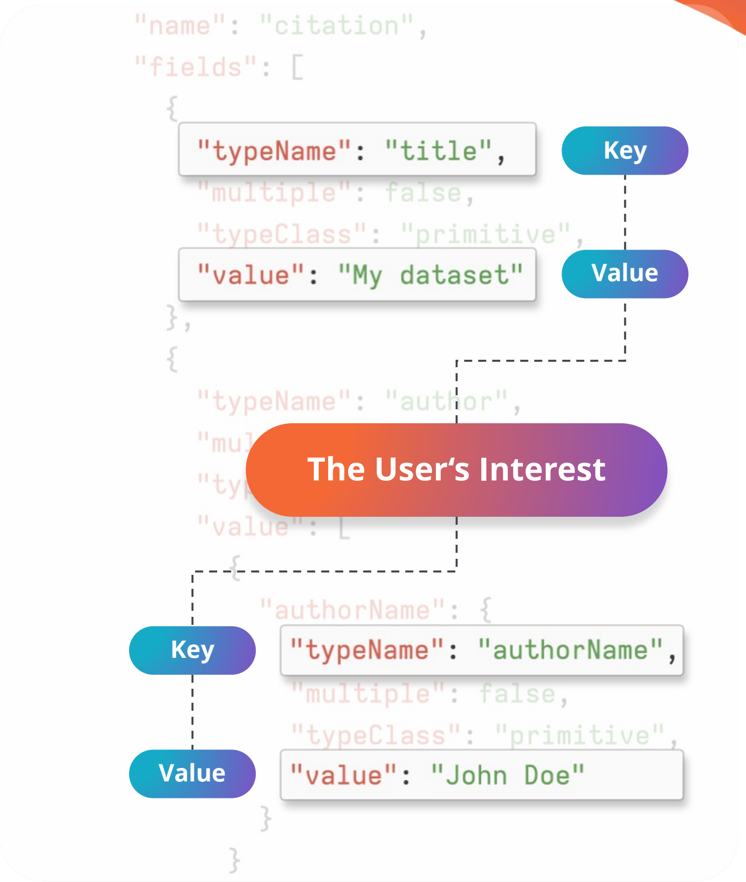
Low-level to the core

Requires deeper knowledge

- Understanding of the native API
- Combination of multiple endpoints for high-level concepts (e.g. Dataset handling)

Example: Dataset Creation

- Requires two endpoints to be called
- Complex Dataverse JSON to add metadata



```
"name": "citation",
"fields": [
  {
    "typeName": "title",
    "multiple": false,
    "typeClass": "primitive",
    "value": "My dataset"
  },
  {
    "typeName": "author",
    "mul
    "ty
    "value": [
      {
        "authorName": {
          "typeName": "authorName",
          "multiple": false,
          "typeClass": "primitive",
          "value": "John Doe"
        }
      }
```

pyDataverse

High-level API

Dynamic Metadata Objects

(EasyDataverse)

Data File Upload to S3

(DVUploader)

Low-level API

Dataverse API Endpoints

(URLs and parameters)

Model Schemes

(Payloads and responses)

Testing framework

Unit Tests

Integration

End-To-End

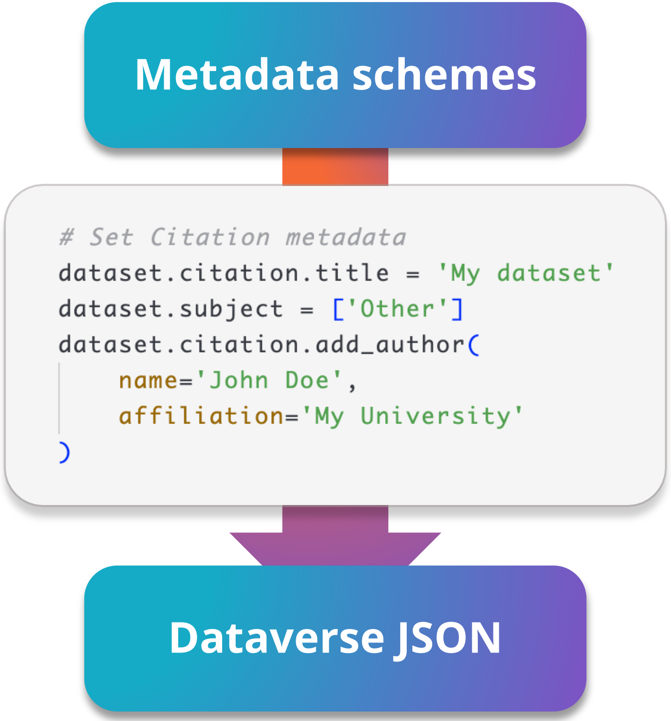
High-level concepts

Dynamic metadata objects

- Fetches metadata schemes on connection
- Transforms schemes into Python classes
- Stores technical details in attribute metadata

High-level recipe

- Combines multiple endpoints into a simple interface
- Improves user-friendliness and convenience
- Built-in type and schema validation



Metadata schemes

```
# Set Citation metadata
dataset.citation.title = 'My dataset'
dataset.subject = ['Other']
dataset.citation.add_author(
    name='John Doe',
    affiliation='My University'
)
```

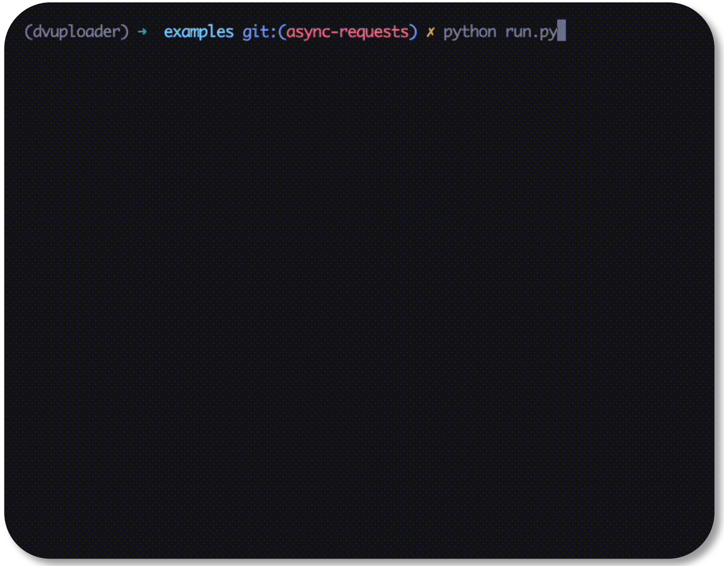
Dataverse JSON

High-level concepts
Parallelized upload to S3

- Facilitates direct upload to an S3 bucket
- Complex procedure of multiple endpoints
- Avoids duplicate files by checksum lookup

High-level recipe

- Abstracts procedure into a simple interface
- Improves user-friendliness and usability
- Asynchronous process to maximize performance



```
(dvuploader) ➜  examples git:(async-requests) ✗ python run.py
```

Practical example

Dataset handling with S3

1 Create dataset at demo.dataverse.org

- Using metadata block objects
- Uploading multiple files to S3

2 Update the dataset and add more files

→ Google Colab Notebook

Join us!

PyDataverse Working Group

We are open to anyone who wants to participate and contribute!

- → Bi-weekly meetings
- → Recordings, notes and upcoming dates available at https://py.gdcc.io

💬 *#python* on Zulip → Here