
Modularity and Interoperability in Generalist Data Repositories

Gustavo Durand

Stefano M. Iacus



Introduction to Dataverse

Gustavo Durand

Tech Lead / Architect of the Dataverse Project

IQSS, Harvard University

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
- Core team
 - @ IQSS - developers, designers, UX/UI, metadata specialists, curation team, leadership team
 - key contributors from the community

Dataverse Features

<https://dataverse.org/software-features>

- Main goal of core code is to focus on publishing (citing, sharing, versioning, etc.), FAIR Data principles
- Robust APIs to allow interoperability with “external tools” and other repositories / software

Dataverse Community

Dataverse Community

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

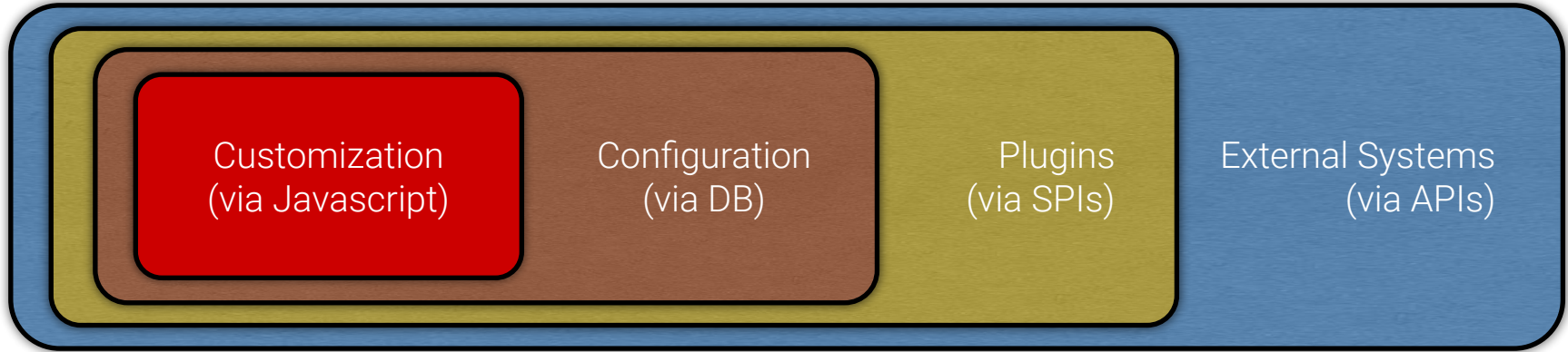
Dataverse Community

- 100 (self reporting) installations around the world



Modularity in Dataverse

Dataverse Ecosystem



Customization (via Javascript)

Customization (via Javascript)

Dataverse allows admins to customize their installations with HTML/Javascript in a few areas.

Customization (via Javascript)

- **Branding**
 - Dataverse provides configurable options for easy-to-add (and maintain) custom branding for your Dataverse installation via a custom home page, header, footer and / or style sheet
- **External Vocabulary Support**
 - Dataverse supports the integration of browser-based scripts that can alter the metadata entry and display user interfaces on a per metadata field basis
 - ORCID, e.g. author name field
 - CrossRef FundReg, e.g. Funding Info, Agency field
 - ROR, e.g. author affiliation field
 - Skomos - many vocabularies, e.g. keyword field

Configuration (via DB)

Configuration (via DB)

Several areas of functionality are defined by configuration via the database, rather than in the code itself, allowing the same code to be deployed by different institutions with different needs.

Configuration (via DB)

- **Custom Metadata definitions**
 - All metadata fields are defined outside of the code itself (currently in .tsv files) that describe the different attributes for that field.
 - These files, in turn, can be imported in via API and the DatasetFieldType definitions are stored in the database
- **Roles**
 - A role consists of a set of permissions that a user is allowed to perform
 - these sets of permissions are stored in the database (defined in external json files).
- **Authentication Providers**
 - Dataverse supports configuration of authentication providers via the api, with the configuration stored in the database
- **Import**
 - the base Dublin Core format, is implemented as a prototype of an expandable, programmatic model
 - instead of coding an XML format parser, a mapping of the external format fields to the internal Dataverse metadata structure can be defined in the database
 - A better model for import would be the Plugin model

Plugins (via SPIs)

Plugins (via SPIs)

Plugins allow developers to extend the functionality of the core code without having to make a separate fork of the repository. In Dataverse, we enable this via the SPI (Service Provider Interface) model.

Plugins (via SPIs)

- **Increasingly modular variations of SPI**
 - Dynamic loading
 - Dynamic loading from an external jar
 - Creating an Interface Library
 - Supporting separate execution

Plugins (via SPIs)

- **Export**
 - Metadata exporters take information provided by Dataverse about the contents of a Dataset, i.e. its metadata and list of included files and their metadata, and generate a file conforming to a specific community metadata format including all of that information or the subset that matches the format
 - V5.14 - can be built using dataverse-spi jar and deployed by dropping into a specified directory on the Dataverse server
- **PID Provider**
 - Dataverse currently supports the use of third-part PIDs as the way to persistently identify datasets and, optionally, individual files
- **Workflow Steps**
 - Actions that can be chained together and run via API/publication triggers (next section)
 - Some existing steps can call remote services
- **Archival Bag Creation**
 - Dataverse has an extensible mechanism to capture the full metadata and data contents of a published dataset version into an archival zipped BagIt Bag following the Research Data Alliance recommendations for Bag structure and contents
- **Data Stores**
 - New ways to store data files (e.g. local files, S3, Swift, Globus, Remote, ...)

External Systems (via APIs)

External Systems (via APIs)

From Dataverse 4 onward, APIs have been a major focus of the software and a majority of the functionality that is available via the UI is also available via API.

This allows external developers to develop other applications, which we often refer to as **external tools**, using whatever technology is most effective for their purpose.

Client Libraries

- Several client libraries have been created to help developers interact with Dataverse APIs from other languages, including:
 - Python
 - R
 - Javascript

External Tools

- Tools that talk to Dataverse
 - generally, used to deposit data into Dataverse (via Deposit API)
 - usually don't require anything special to be set up in the Dataverse repository
- Tools that Dataverse talks to
 - user starts on Dataverse and is directed to the external tool
 - have predefined areas in the UI where these would plug into (Explore tools)
 - require manifest files
- Tools that do both
 - user starts on Dataverse and is directed to the external tool
 - also have predefined areas in the UI where these would plug into (Configure tools)
 - require manifest files
 - will also send something back to Dataverse, so need an API token that has “write” privileges

External Workflows

- The Dataverse Software can also perform two sequences of actions when datasets are published
 - **PrePublishDataset** trigger
 - useful for having an external system prepare a dataset for being publicly accessed or to start an approval process.
 - **PostPublishDataset** trigger
 - might be used for sending notifications about the newly published dataset.

A couple of concrete cases

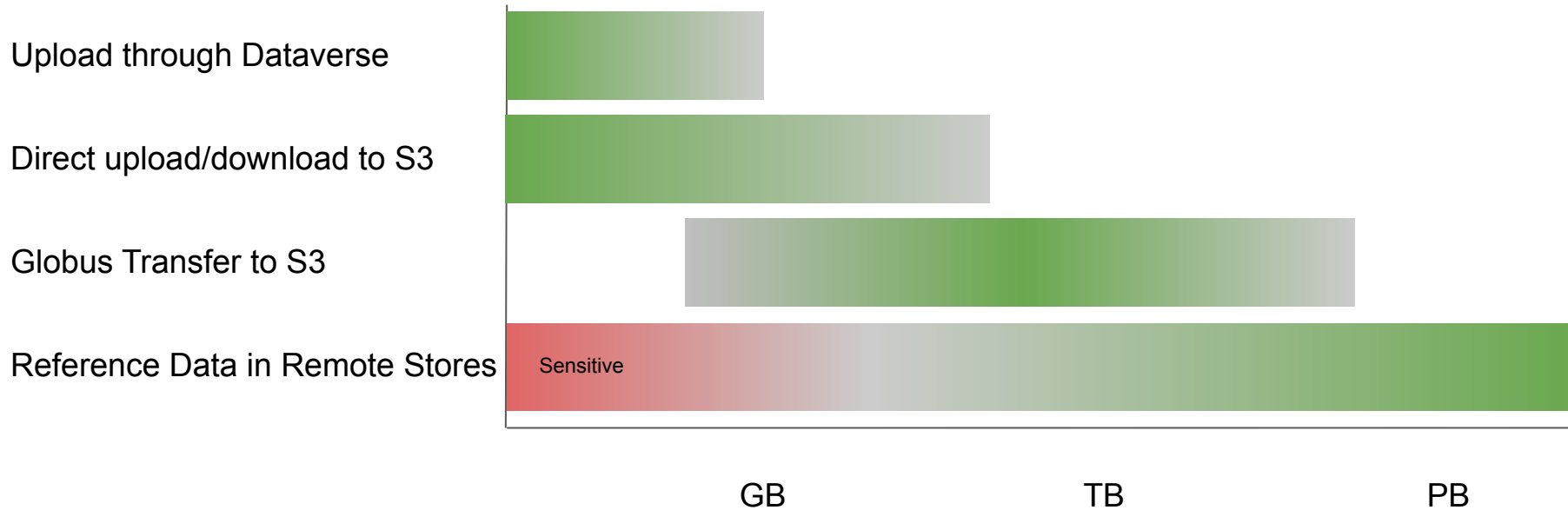
Stefano M. Iacus

Managing Director Dataverse Project

IQSS, Harvard University

**Really big data with Globus
("external tool" plugin)**

Really big data support in Dataverse (overview from 2022)



Previous Globus Work (v5.12)

Use of the Globus S3 connector to allow transfer of data to/from Dataverse via Globus

- Dataverse configured with S3 store
- Uses **external** JavaScript Dataverse-Globus transfer app

Remote Overlay store to reference remote data via URL

Originally created by Borealis group, updated/merged in v5.12 with support from Harvard Data Commons

Mostly Jim Mayer's contribution

Limitations

- Added cost for connector
- Doesn't handle restriction/embargoes
- **Not well suited to parallel transfer to tape storage**
 - Requires parallel S3 endpoints
 - **Assumes direct and immediate access via S3**
- **Can't reference remote (too large to transfer) files directly using Globus (only URL)**

Updated Design

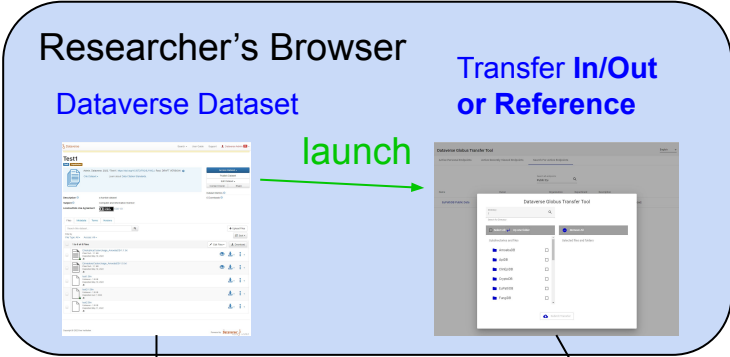
- Build upon the RemoteOverlay store design to reference files on Globus endpoints, including file/tape endpoints
 - Allow referencing a remote endpoint without transfer to Dataverse
 - **Use Globus API (rather than S3)** for store interactions
- Optionally **store files in separate sub-directories** to support **access control** needed for restriction/embargoes
- Extend Dataverse-Globus transfer app for transfer in/out (no download from Dataverse UI)

Globus Transfer To Dataverse

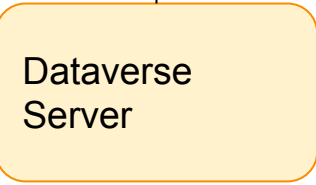
Researcher's Browser
Dataverse Dataset

Transfer In/Out or Reference

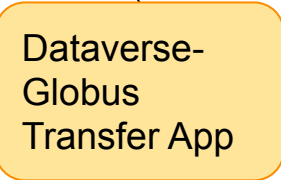
launch



Dataverse Server



Dataverse-Globus Transfer App

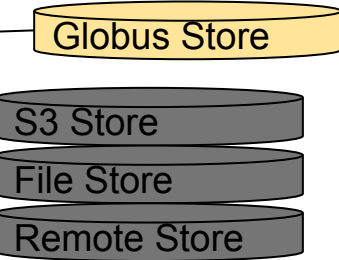


Globus Store

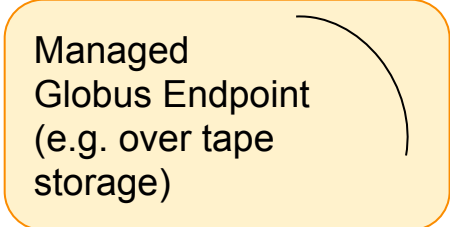
S3 Store

File Store

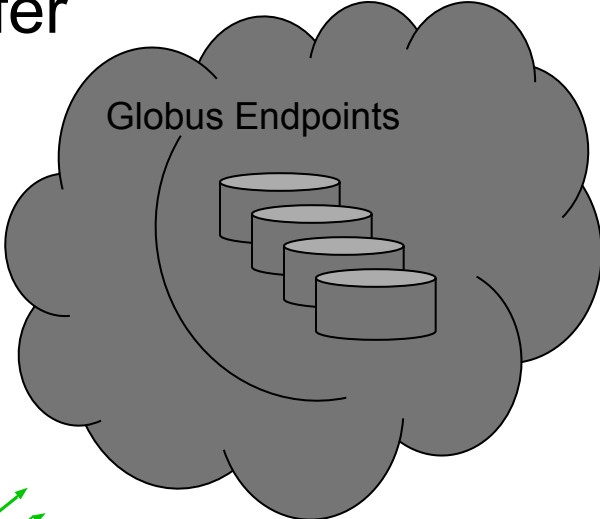
Remote Store



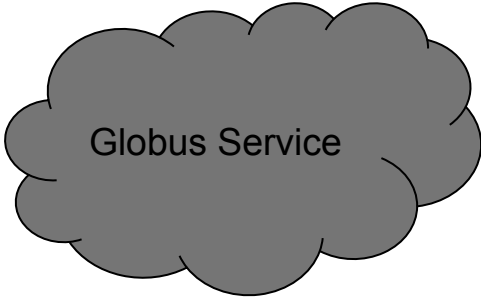
Managed Globus Endpoint (e.g. over tape storage)



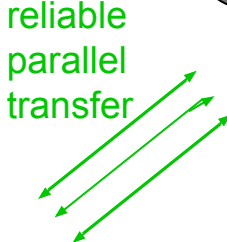
Globus Endpoints



Globus Service



reliable parallel transfer



External vocabularies (JavaScript plugin)

External Vocab Support Demo

Compound Skosmos Keyword Demo

Vocabulary: unesco

Vocabulary URL: http://skos.um.es/unescothes/CSD00

Term: Precious metals

Term URL: http://skos.um.es/unescothes/C03108

Creator Demo: https://orcid.org/0000-0001-8462-650X

Skosmos Demo: http://skos.um.es/unescothes/C03107

http://aims.fao.org/aos/agrovoc/c_6161

Skosmos Demo

ARDC Topic

Vocabulary: unesco

Term: Precambrian, http://skos.um.es/unescothes/C03107

Precious metals, http://skos.um.es/unescothes/C03108

Precipitation, http://skos.um.es/unescothes/C03109

Measuring Instruments (Precision Instruments), http://skos.um.es/unescothes/C02444

Forecasting (Prediction)

Input via the Skosmos JavaScript: the **original** input field is **hidden** and **replaced** by **both a vocabulary selector** and a **dynamic term selection list**.

Skosmos Demo

Precambrian, http://skos.um.es/unescothes/C03107

ARDC Topic

Vocabulary: unesco

Term: Select a term

agrovoc

Selecting an alternate vocabulary hosted on a Skosmos server.

Skosmos Demo

Precambrian, http://skos.um.es/unescothes/C03107

ARDC Topic

Vocabulary: agrovoc

Term: pre-emption rights, http://aims.fao.org/aos/agrovoc/c_9c081fdc




preagricultural sector, http://aims.fao.org/aos/agrovoc/c_6158

prebiotics, http://aims.fao.org/aos/agrovoc/c_4ddb48b

precipitation, http://aims.fao.org/aos/agrovoc/c_6161

precipitation deficit, http://aims.fao.org/aos/agrovoc/c_25308

Selecting a term from the chosen vocabulary. The Skosmos JavaScript dynamically populates the list with terms matching what the user has typed.

-  **Dataverses (2)**
-  **Datasets (6)**
-  **Files (8)**

Dataverse Category
Department (2)

Publication Year
2022 (5)

Publication Status
Published (4)
Draft (3)
Unpublished (3)

Author Name
Admin, Dataverse (5)
Myers, Jim (1)

Subject
Computer and Information Science (4)
Earth and Environmental Sciences (1)
Mathematical Sciences (1)
Social Sciences (1)

Deposit Date
2022 (6)

Skosmos Demo
Precambrian (2)
precipitation (1)

One of the **metadata fields promoted** as a **search facet** on the Dataverse collection page. The JavaScript replaces the URLs with a human readable form.

```
<input id="datasetForm:tabView:j_idt1613:1:j_idt1616:2:fieldvaluelist:1:cvocInputText"
  name="datasetForm:tabView:j_idt1613:1:j_idt1616:2:fieldvaluelist:1:cvocInputText"
  type="text" value="http://aims.fao.org/aos/agrovoc/c_6161" class="ui-inputfield ui-
  inputtext ui-widget ui-state-default ui-corner-all form-control ui-state-filled" data-
  cvoc-parent="skosterm" data-cvoc-managedfields="{}" data-cvoc-vocabs="
  {&quot;unesco&quot;;
  {&quot;vocabularyUri&quot;;&quot;http://skos.um.es/unescothes/CS000&quot;;&quot;uriSpa
  ce&quot;;&quot;http://skos.um.es/unescothes/&quot;;&quot;agrovoc&quot;;
  {&quot;vocabularyUri&quot;;&quot;http://aims.fao.org/vest-
  registry/kos/agrovoc&quot;;&quot;uriSpace&quot;;&quot;http://aims.fao.org/aos/agrovoc/
  &quot;}}" data-cvoc-filter="" data-cvoc-service-url="https://skosmos.dev.finto.fi/"
  data-cvoc-protocol="skosmos" data-cvoc-allowfreetext="true" lang="en" aria-
  label="Additional Entry" role="textbox" aria-readonly="false" aria-disabled="false"
  wtx-context="E1612EF8-35A8-4BA6-A10C-351E6575A332">
```

To **enable JavaScripts** to find the fields they should manage, Dataverse annotates them with appropriate **data-* attributes**. The HTML source for the input example above includes attributes that specify “skosmos” as the protocol, and indicate which Skosmos server should be called and which vocabularies to allow, to allow free-text entries, and to display terms in English.

```

[{"field-name": "compoundDemo",
"term-uri-field": "compoundDemoTermURI",
"cvoc-uri": "https://skosmos.dev.finto.fi/",
"js-uri": "https://gdcc.github.io/dataverse-external-vocab-support/scripts/skosmos.js",
"protocol": "skosmos",
"retrieval-uri": "https://skosmos.dev.finto.fi/rest/v1/data?uri={0}",
"term-parent-uri": "",
"allow-free-text": false,
"languages": "en, fr, es, ru",
"vocabsts": {
  "unesco": {
    "vocabularyUri": "http://skos.um.es/unescothes/C5000",
    "uriSpace": "http://skos.um.es/unescothes/"
  }
},
"managed-fields": {
  "vocabularyName": "compoundDemoVocabulary",
  "termName": "compoundDemoTerm",
  "vocabularyUri": "compoundDemoVocabularyURI"
},
"retrieval-filtering": {
  "@context": {
    "termName": "https://schema.org/name",
    "vocabularyName": "https://dataverse.org/schema/vocabularyName",
    "vocabularyUri": "https://dataverse.org/schema/vocabularyUri",
    "lang": "@language",
    "value": "@value"
  },
  "@id": {
    "pattern": "{0}",
    "params": ["@id"]
  },
  "termName": {
    "pattern": "{0}",
    "params": ["/graph/uri=@id/prefLabel"]
  },
  "vocabularyName": {
    "pattern": "{0}",
    "params": ["/graph/type=skos:ConceptScheme/prefLabel"]
  },
  "vocabularyUri": {
    "pattern": "{0}",
    "params": ["/graph/type=skos:ConceptScheme/uri"]
  }
}
},
]

```

The **JSON configuration used by Dataverse to associate the Skosmos JavaScript with the new metadata field**. In addition to information that is copied to data-* attributes as discussed above, the configuration includes a link to the desired JavaScript so Dataverse can load it, the name of the metadata field that will be managed by the script. The retrieval-url and retrieval- filtering entries tell Dataverse how to retrieve JSON/JSON-LD information about the term which is then cached for later use.

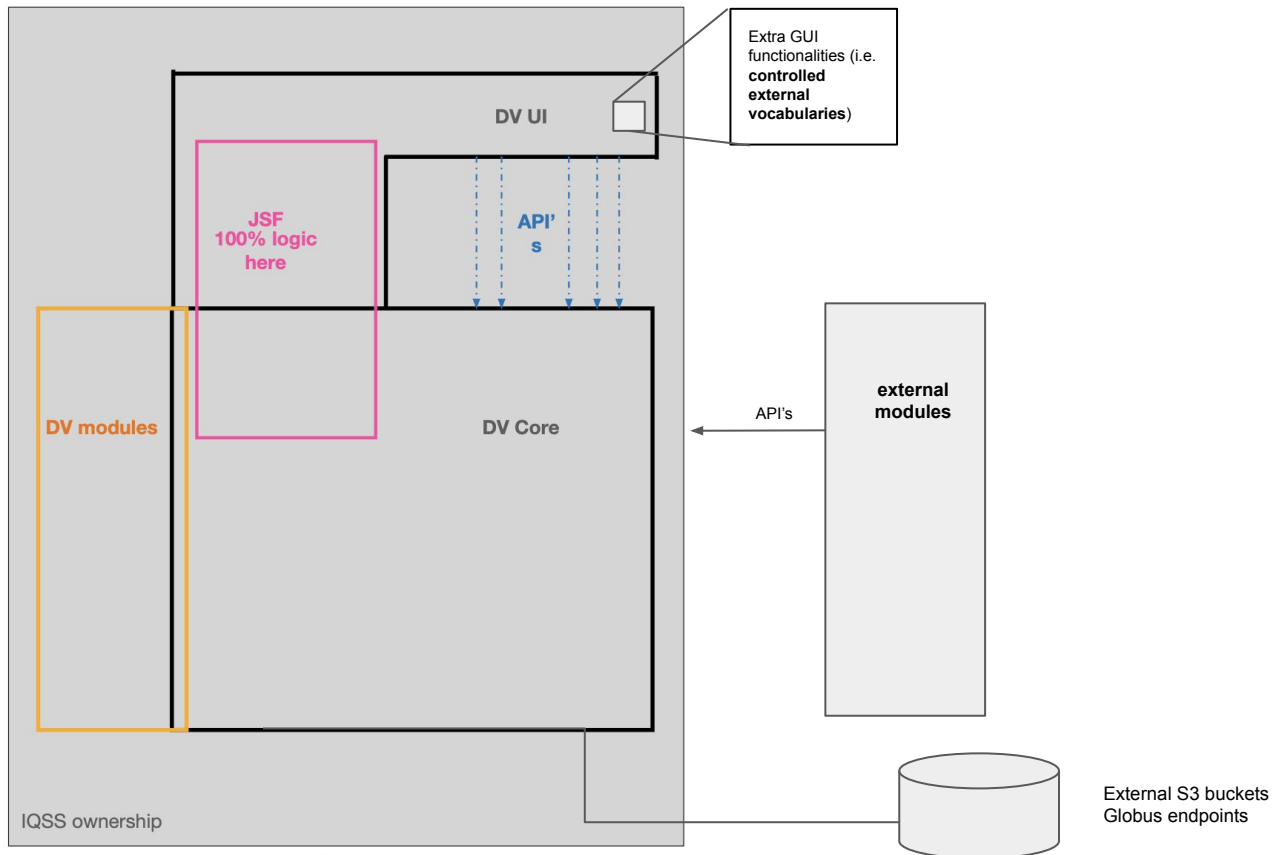
The Re-Arch Project

Goals

- **modernize** the application
- **separate backend and frontend** to increase **interoperability**
- Dataverse backend becomes an **API-first application**
- **extend modularization** of backend and frontend
- **speed up development** and implementation of new UI/UX ideas
- Native **accessibility** (A11y) and **internationalization** (i18n) support
- **empower the community**

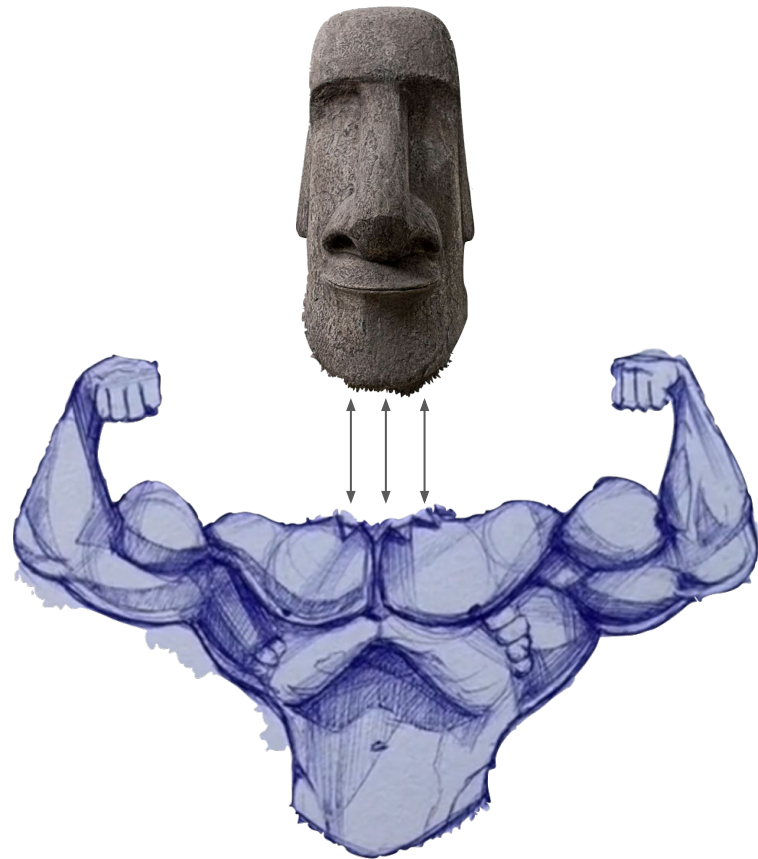
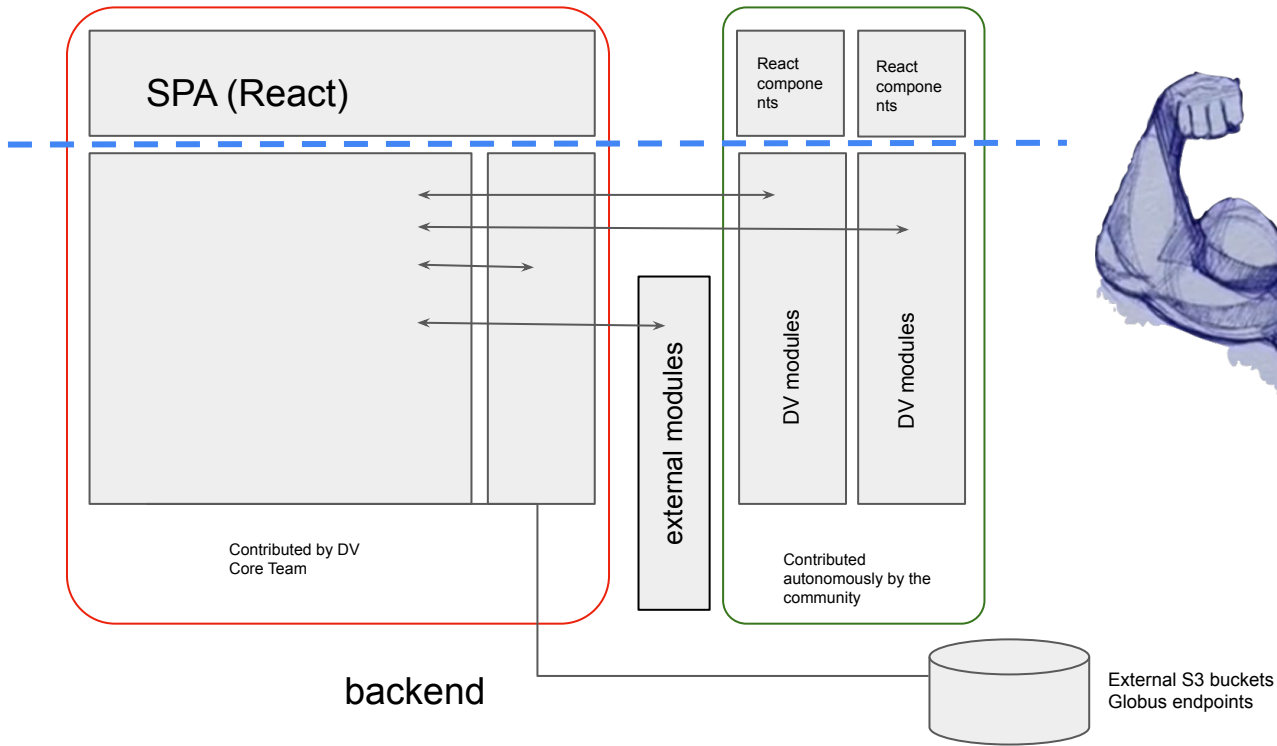
Re.
ar EXPOSITION
chi
tec
ture

Monolith (though robust Java) application

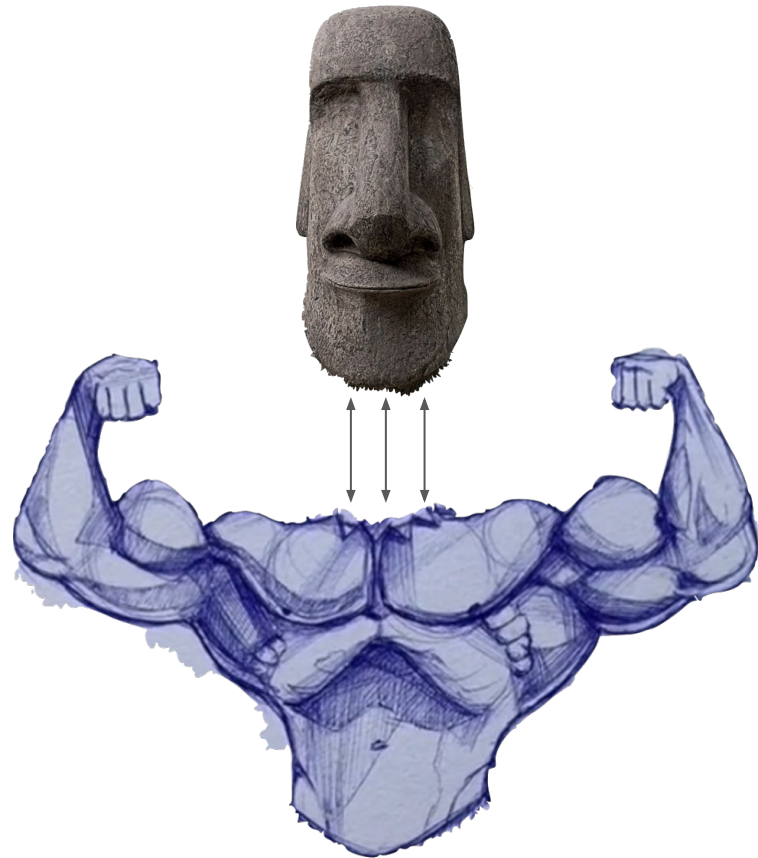
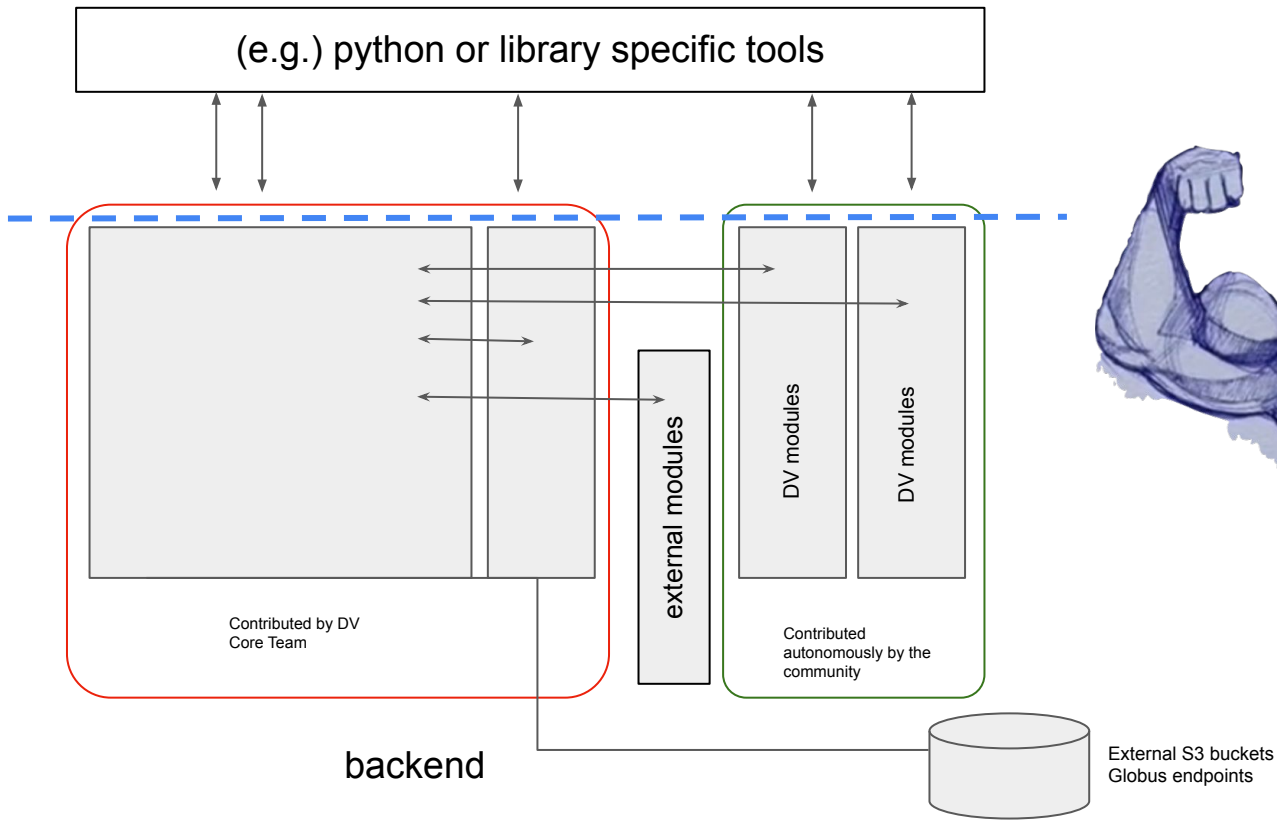


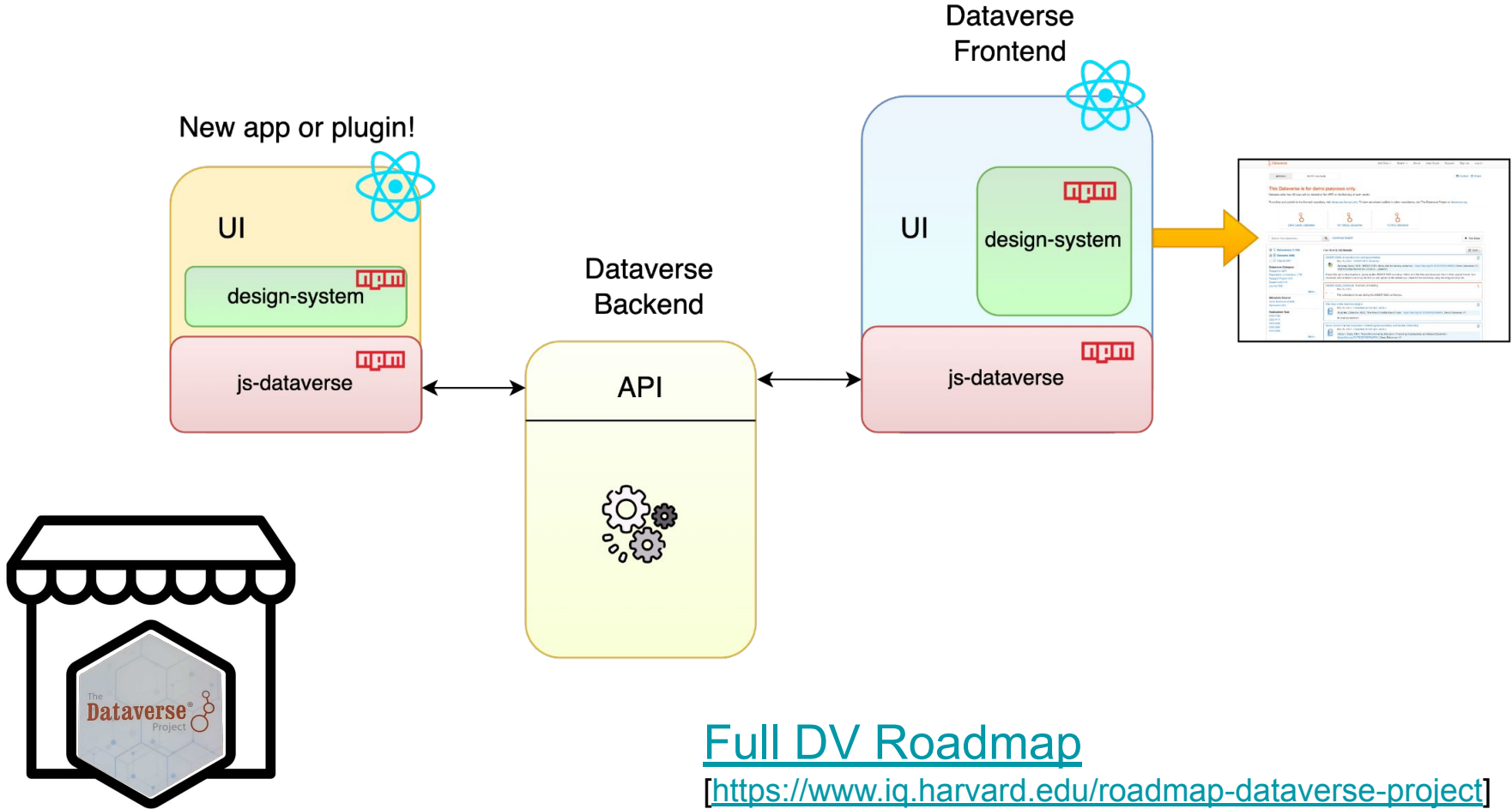
DV as an headless API-first application

frontend



DV as an headless API-first application





DV plugin marketplace

[Full DV Roadmap](https://www.iq.harvard.edu/roadmap-dataverse-project)

[\[https://www.iq.harvard.edu/roadmap-dataverse-project\]](https://www.iq.harvard.edu/roadmap-dataverse-project)

Thank you

Dataverse Community Meeting 2024
CIMMYT in Texcoco, Mexico



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>