Cheap and FAIR:

Building a Serverless Research Data Repository

Introduction

Andrea Zonca & Rick Wagner
zonca@sdsc.edu & rick@sdsc.edu
SDSC

Gateways 2024, September 30, 2024, Online

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

UC San Diego

Outline

Presentation

Hands-on

Break

- Introduction & Setup
- NB 1: GitHub Pages & JupyterHub
- What is FAIR Data?
- NB 2: Working with Globus Collections
- Identifiers and Landing Pages
- NB 3: Creating a Data Repository Catalog
- 15 Minute Break
- NB 4: Validating the Catalog & Programmatic Data Access
- NB 5: Enhancing the Catalog
- Making the SRDR Interactive
  - NB 6: Public Data
  - NB 7: Restricted Data

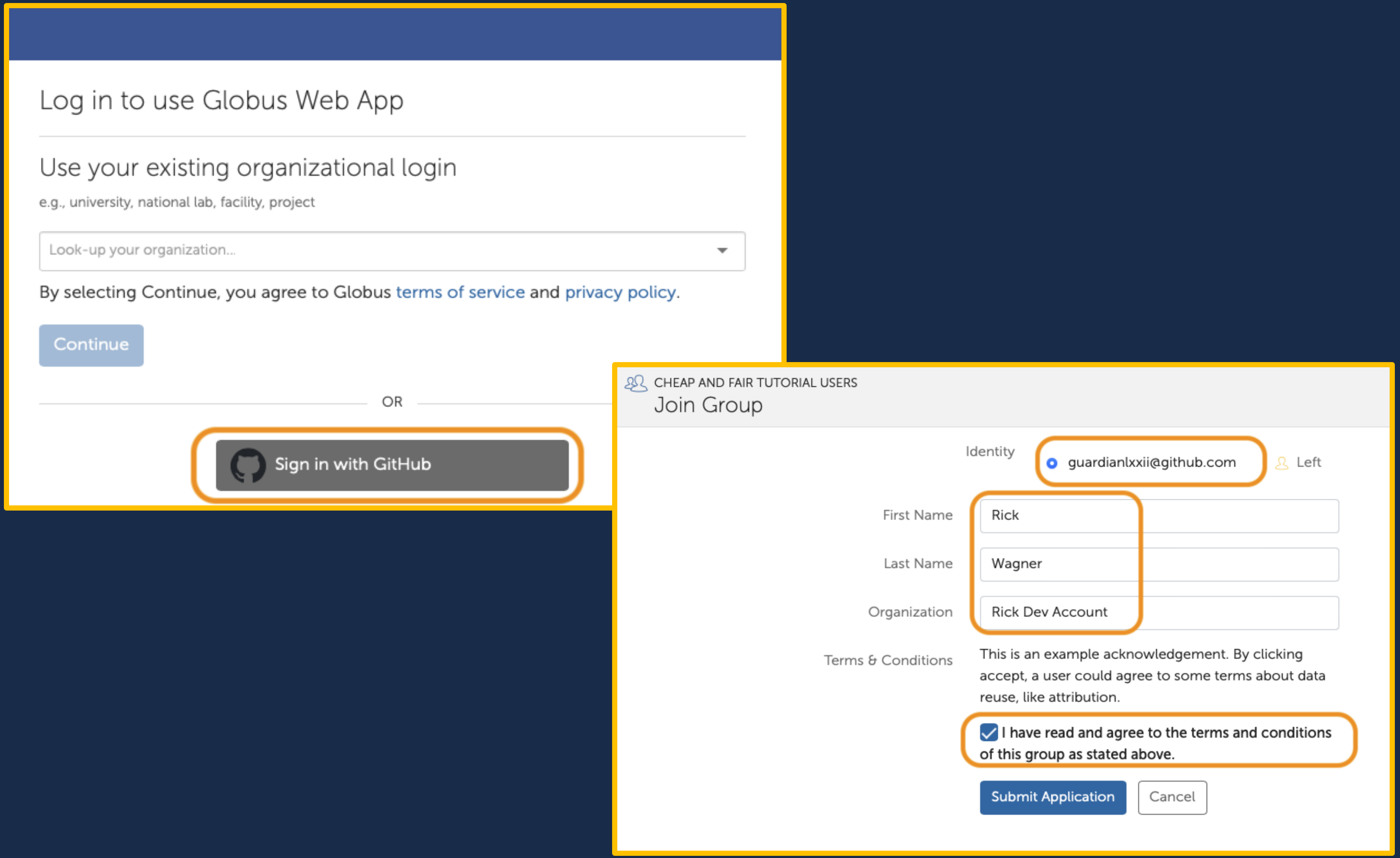
Initial Technical Requirements

- Internet connection
  - If you’re listening to this, you probably have one
- Web browser
- For communication, use whatever suits you
  - Chat
  - Raise hand, voice, etc.
  - Rick & Andrea will alternate helping attendees
  - Turn your camera on if you’re comfortable
- As we proceed, we’ll ask if you’ve completed a step. Please use the yes and no reactions in Zoom.

✓ ✗

Setup Steps

- Link to setup steps below
- Join GitHub if needed
- Globus:
  - Join with GitHub
  - Or Link GitHub identity
- Join Group to access resources



Log in to use Globus Web App

Use your existing organizational login

e.g., university, national lab, facility, project

Look-up your organization...

By selecting Continue, you agree to Globus terms of service and privacy policy.

Continue

OR

Sign in with GitHub

CHEAP AND FAIR TUTORIAL USERS

Join Group

Identity	● guardianlxxii@github.com Left
First Name	Rick
Last Name	Wagner
Organization	Rick Dev Account
Terms & Conditions	This is an example acknowledgement. By clicking accept, a user could agree to some terms about data reuse, like attribution.

☑ I have read and agree to the terms and conditions of this group as stated above.

Submit Application Cancel

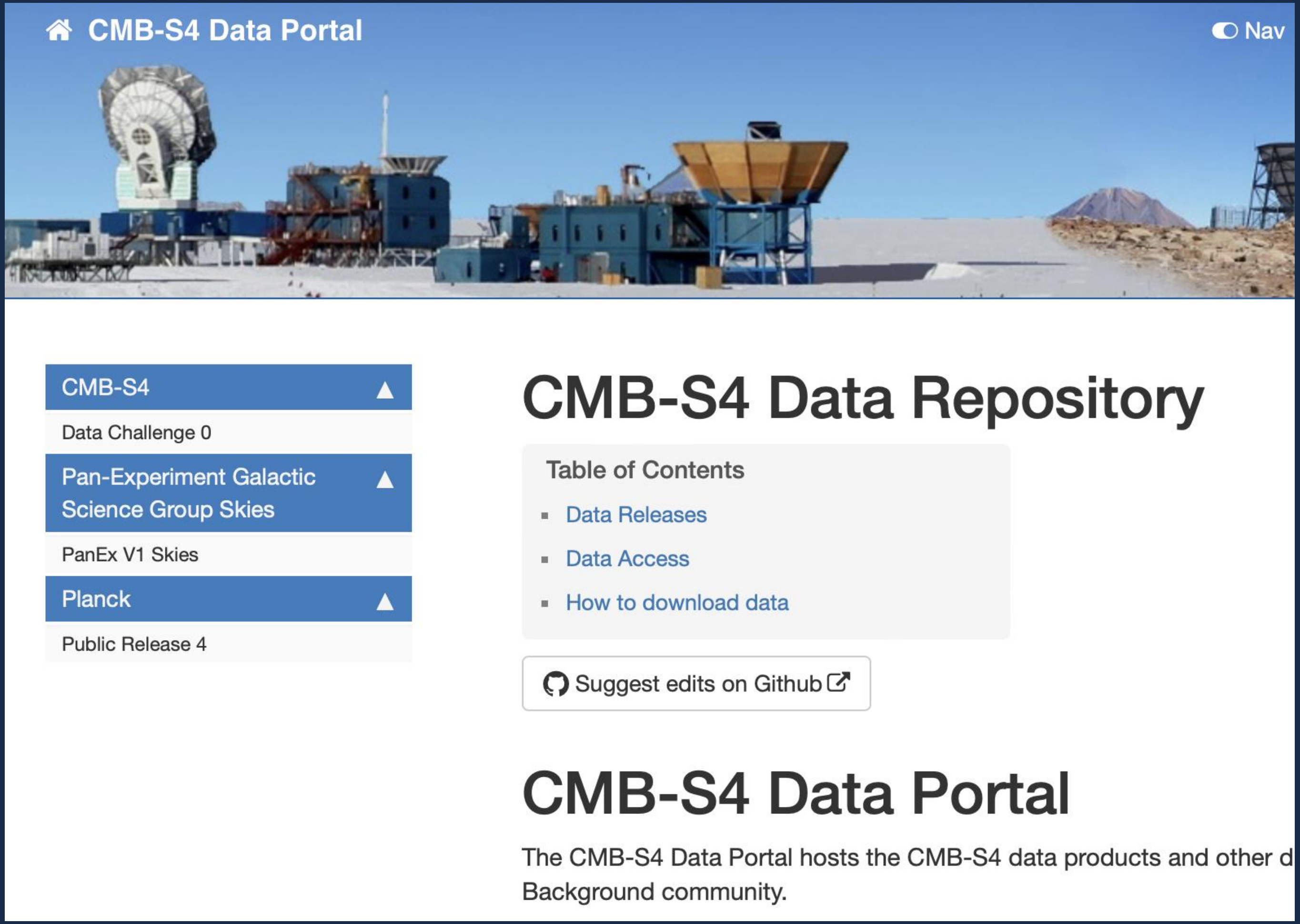
Look in chat

https://github.com/cheapandfair/cheapandfair-gateways-2024

Motivation: CMB S4 Data Repository

- Public & restricted data
- Ease of access
- Large (TBs) and small (MBs) datasets
- Future curation & publication
- Minimal operational overhead

https://data.cmb-s4.org



CMB-S4 Data Repository

Table of Contents

- Data Releases
- Data Access
- How to download data

CMB-S4 Data Portal

The CMB-S4 Data Portal hosts the CMB-S4 data products and other d
Background community.

A “Missing Middle” (a little hyperbolic)

The Dataverse® Project

Researcher projects currently have two data repository options

zenodo

DSPACE

DRYAD

Self-hosted & “on prem”

- Lot of features & extensible
- Requires significant operational support
- Institutional scale solutions

Public or shared repositories

- Zero management, free/“pre-paid”
- Fixed features (e.g., metadata schema)
- Largely open data solutions

Generalist Repository Comparison Chart `https://doi.org/10.5281/zenodo.7946938`

A Data Repository Is...

a set of
datasets

organized in and
accessible from a
collection

described by and
discoverable
in a catalog

managed by one or
more policies

Premise: A data repository's *quality* is based its policies and how well they are implemented.

SRDR* Components

Datasets: One or more files in a single folder and its subfolders.

```
datasetA/
        file1.csv
        file2.png
        manifest.json
        metadata.json
        folderB/
                file3.fits
                file4.dat
```

Collection: Datasets organized in Globus Guest Collection

Provides HTTPS URLs and per-datasets access control.

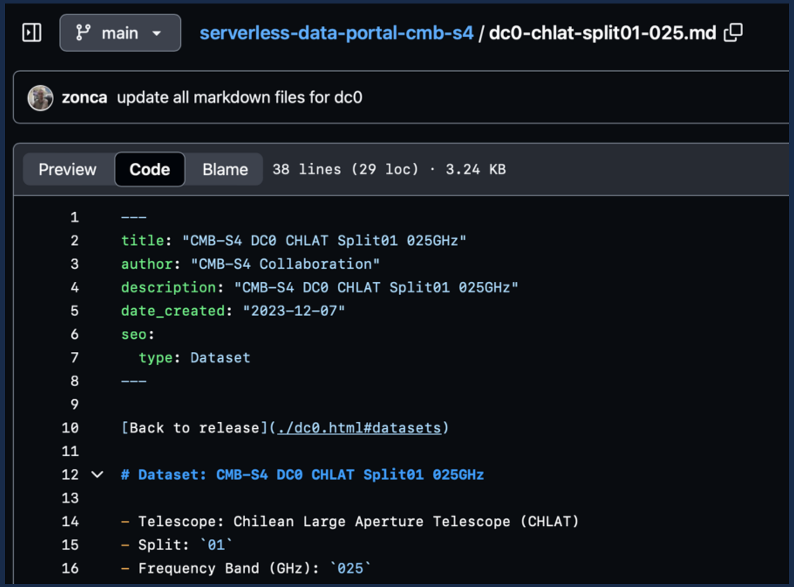
```
collection/
        datarelease1/
                datasetA/
                datasetB/
        datarelease2/
                datasetC/
                datasetD/
```

*SRDR: Serverless Research Data Repository

SRDR Components

Catalog: GitHub Pages site with dataset landing pages and lists of all datasets

Policies: permissions, retention, metadata standards



serverless-data-portal-cmb-s4 / dc0-chlat-split01-025.md

zonca update all markdown files for dc0

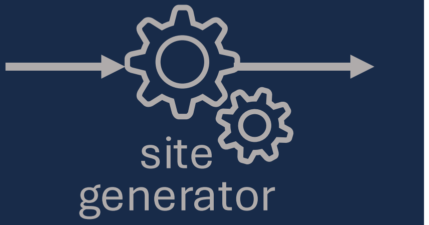
38 lines (29 loc) · 3.24 KB

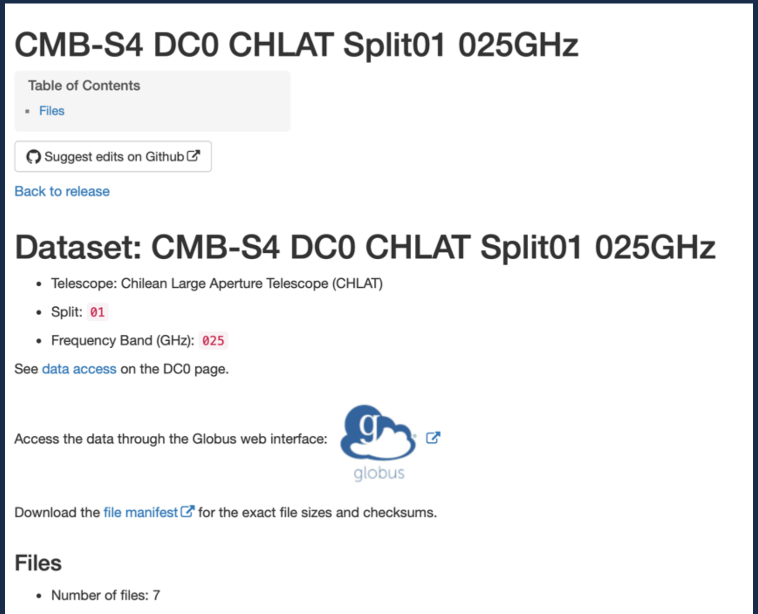
```
---
title: "CMB-S4 DC0 CHLAT Split01 025GHz"
author: "CMB-S4 Collaboration"
description: "CMB-S4 DC0 CHLAT Split01 025GHz"
date_created: "2023-12-07"
seo:
  type: Dataset
---

[Back to release](./dc0.html#datasets)

# Dataset: CMB-S4 DC0 CHLAT Split01 025GHz

- Telescope: Chilean Large Aperture Telescope (CHLAT)
- Split: `01`
- Frequency Band (GHz): `025`
```





CMB-S4 DC0 CHLAT Split01 025GHz

Table of Contents

- Files

Suggest edits on Github

Back to release

Dataset: CMB-S4 DC0 CHLAT Split01 025GHz

- Telescope: Chilean Large Aperture Telescope (CHLAT)
- Split: 01
- Frequency Band (GHz): 025

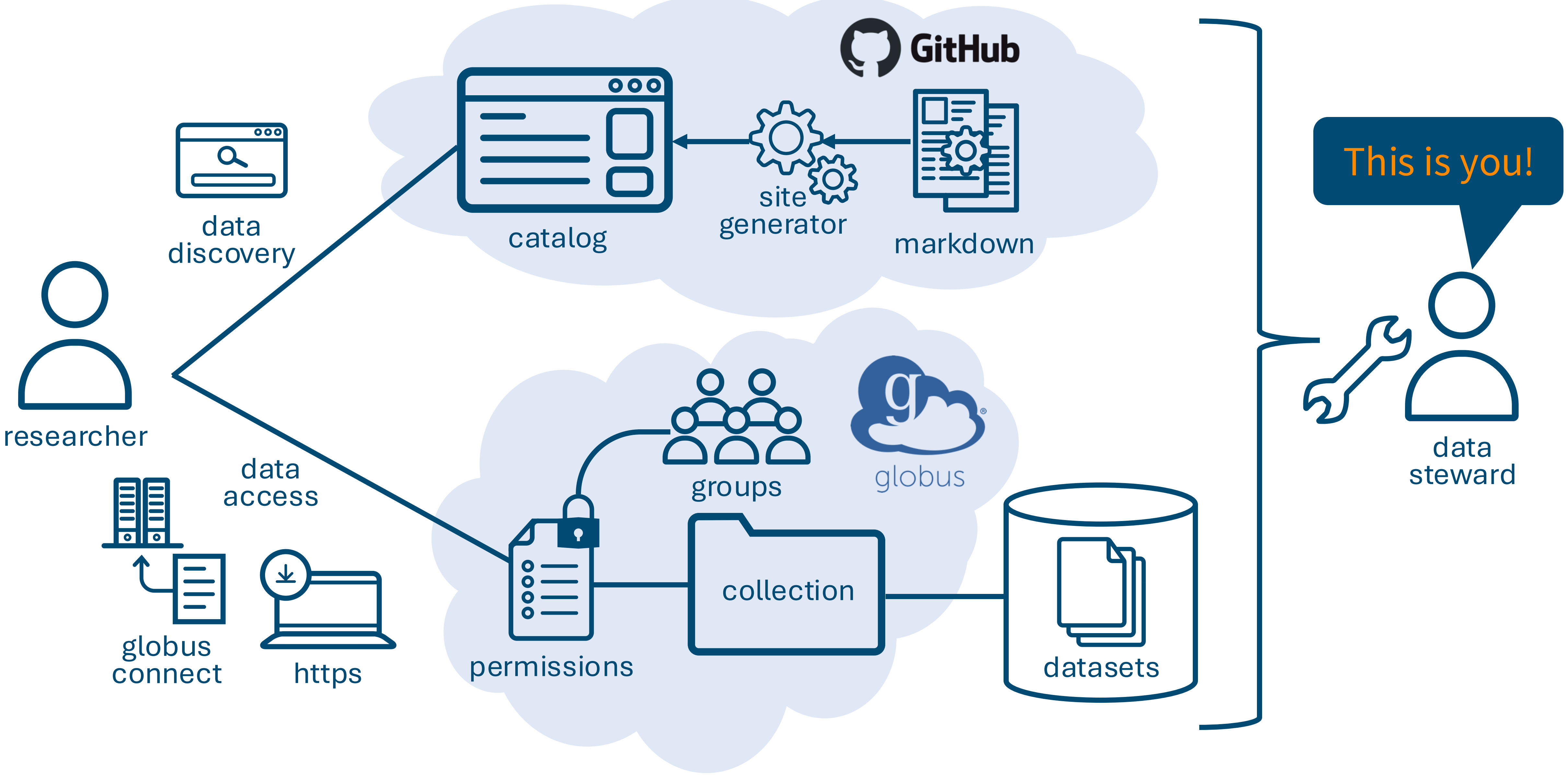
See data access on the DC0 page.

Access the data through the Globus web interface: globus

Download the file manifest for the exact file sizes and checksums.

Files

- Number of files: 7


GitHub
catalog
site generator
markdown
data discovery
researcher
data access
globus connect
https
groups
globus
permissions
collection
datasets
This is you!
data steward

Defining “Cheap”

Minimizing project costs

- ”Serverless”--no systems maintenance
- Catalog hosted on GitHub pages for free
- Dataset storage allocated or cost scale by usage
- Globus Guest Collection covered by institutional subscription
    - Over 200 institutional subscribers
    - 70% of R1 universities
- Model can be applied to other systems beyond Globus

Being FAIR

Findable

- Human & machine-readable metadata
- Datasets listed in catalog
- Unique dataset identifiers

Accessible

- Data & landing pages via HTTPS
- Access control using OAuth/OIDC
- Persistent landing pages

Interoperable

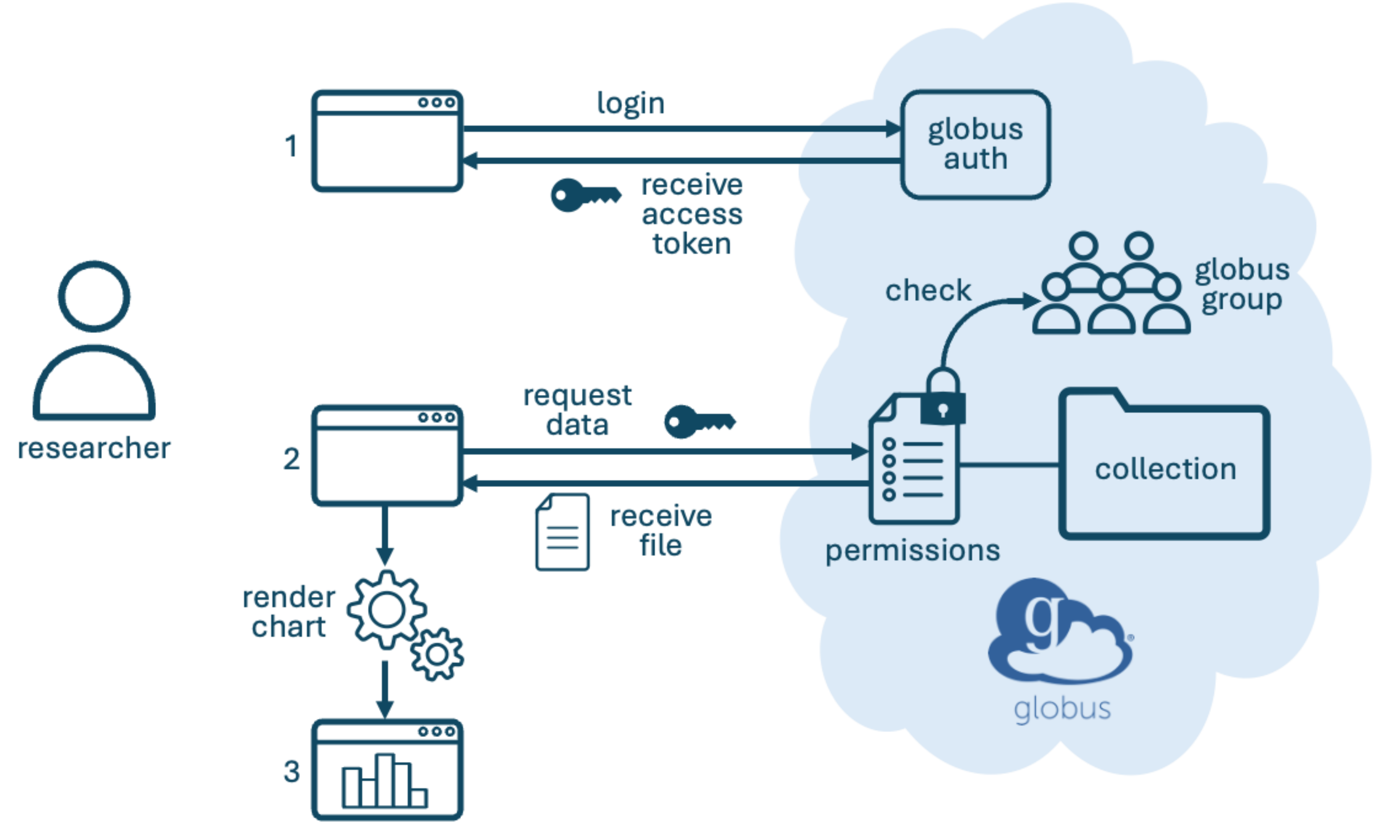
- Metadata based on Schema.org & DataCite
- Metadata in JSON-LD

Reusable

- Data files in widely-used formats
- Usage guidelines defined

From Repository to Interactive Portal

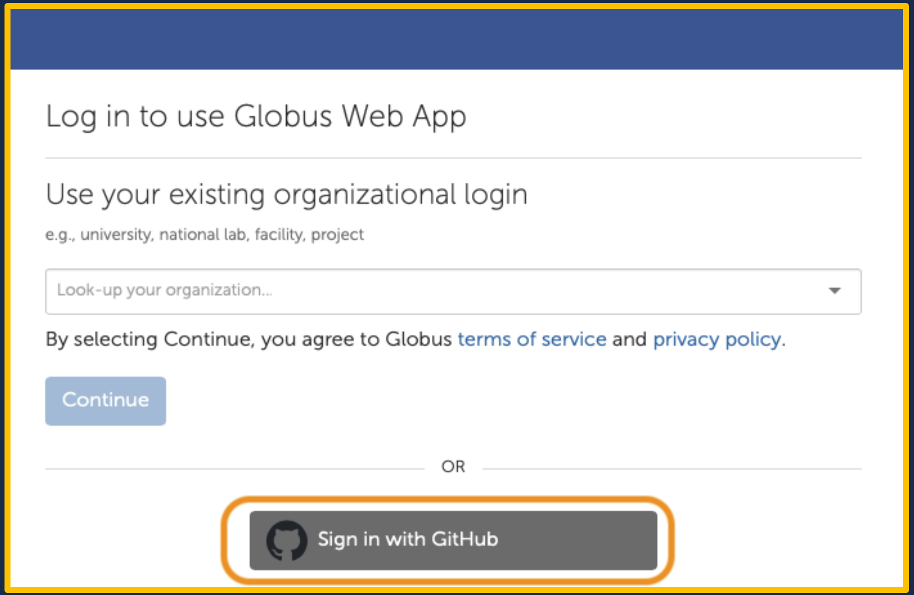
- HTTPS enables in-browser data access
- Interactive: plot, subset, analyze, etc.
- Both public and restricted data
- Can also call APIs to drive workflows



Did you Complete the Setup Steps?

Yes ✓

✗ *No*



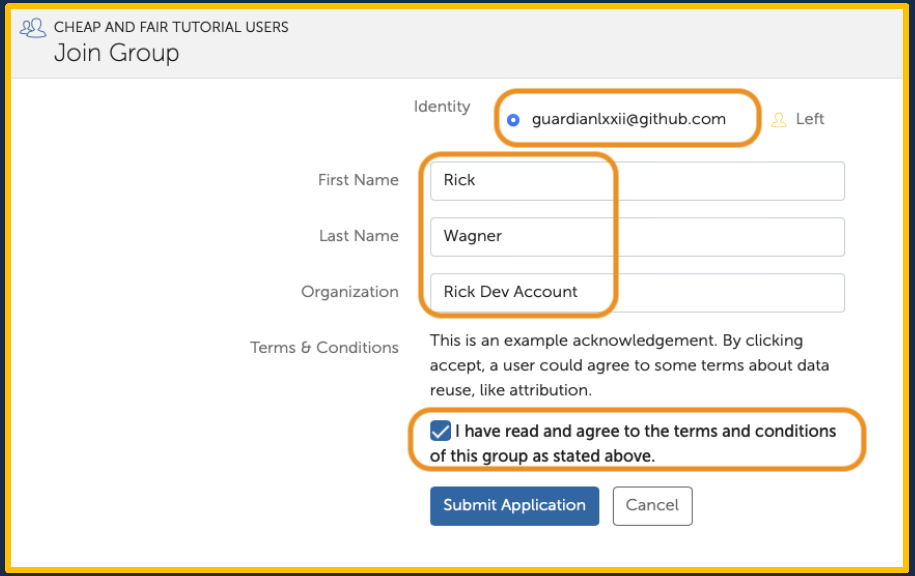
Log in to use Globus Web App

Use your existing organizational login

e.g., university, national lab, facility, project

Look-up your organization...

By selecting Continue, you agree to Globus terms of service and privacy policy.



CHEAP AND FAIR TUTORIAL USERS

Join Group

Identity: ● guardianlxxii@github.com Left

First Name: Rick

Last Name: Wagner

Organization: Rick Dev Account

Terms & Conditions: This is an example acknowledgement. By clicking accept, a user could agree to some terms about data reuse, like attribution.

☑ I have read and agree to the terms and conditions of this group as stated above.

Let’s Get Started!

- Link to the JupyterHub instance below
- Will clone the tutorial repository into your environment

Look in chat

https://bit.ly/srdr-24

Hit the yes when you’re logged in

✓

✗

Send the no if something goes wrong