



Koninklijk Nederlands
Meteorologisch Instituut
Ministerie van Infrastructuur en Milieu

SWIRRL API for provenance-aware and reproducible workspaces

The EPOS and IS-ENES approach

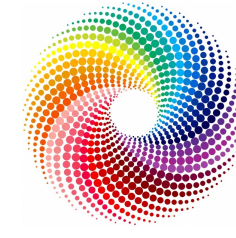
Alessandro Spinuso, Ian van der Neut

Hans Verhoef, Friedrich Striewski, Mats Veldhuizen

Tor Langeland, Christian Page, and Daniele Bailo

R&D Data Technology and Observations

SWIRRL in a nutshell



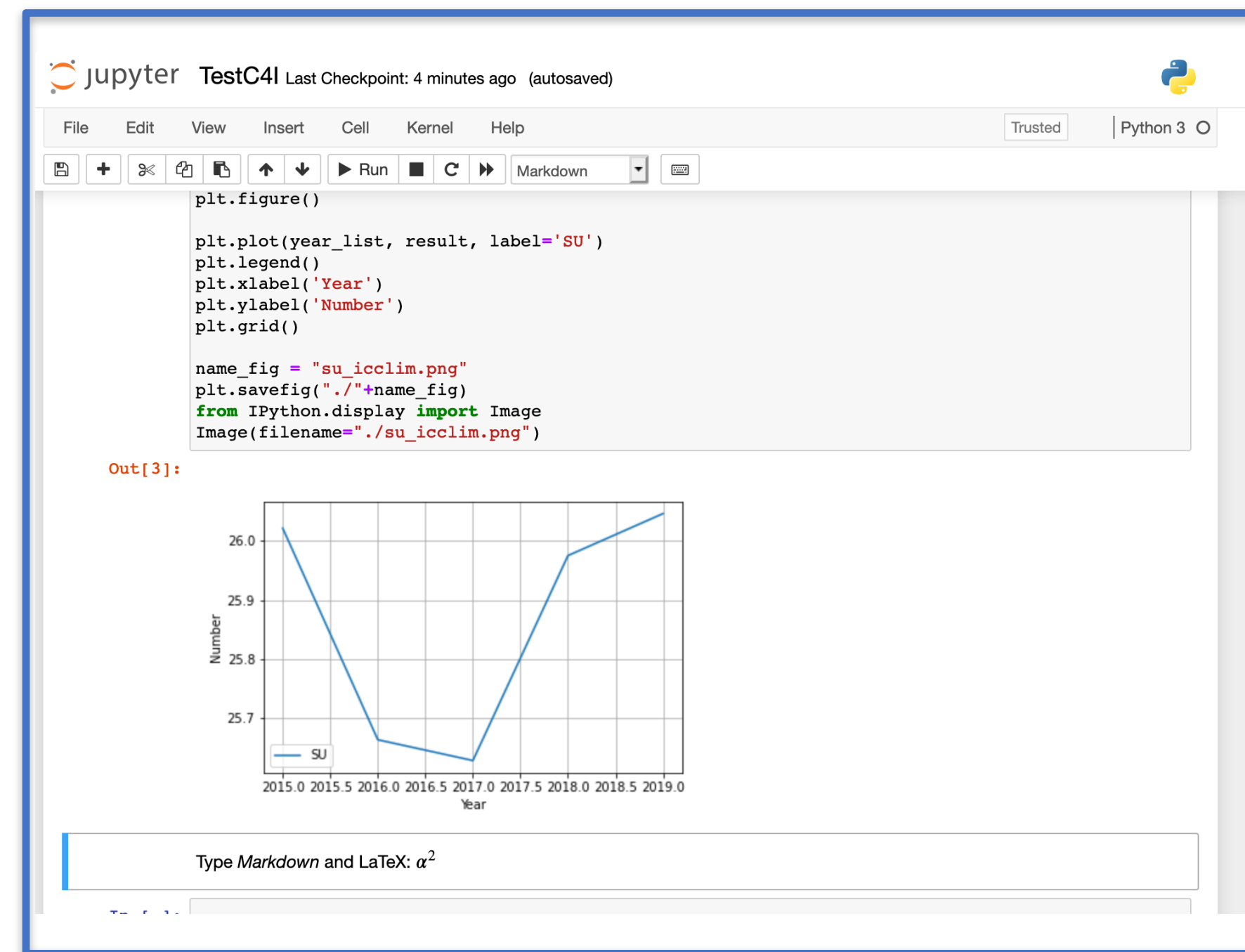
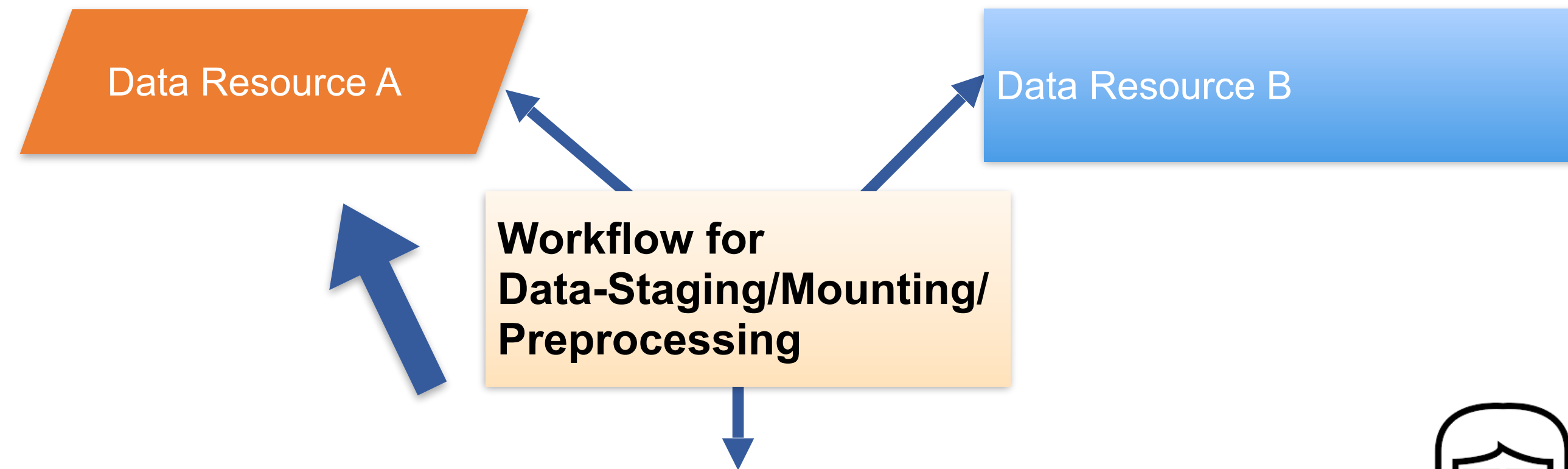
SoftWare for Interactive & Reproducible Research Labs

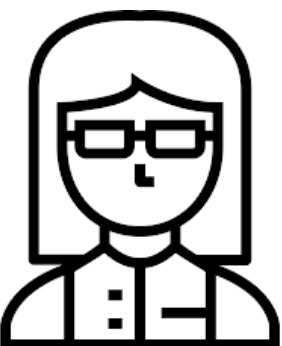
A Web API (high-level piece of infrastructure)

**Manage data-driven & reproducible Working Sessions,
combining Notebooks, Workflows and Visualisation tools**

<https://gitlab.com/KNMI-OSS/swirrl>

Workspace Use Cases




**a researcher
wants**



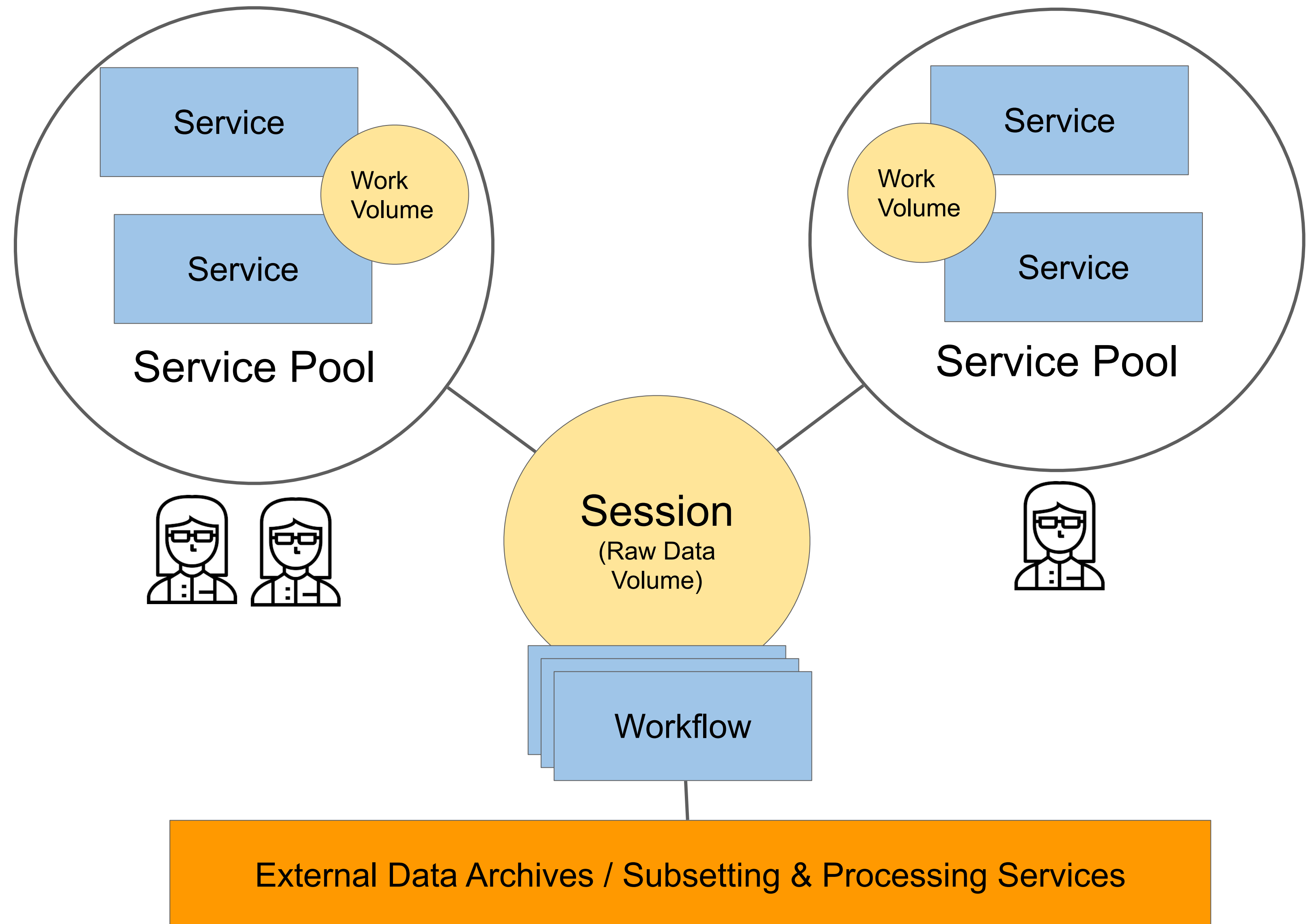
- **access distributed raw data**
- **develop, document and reuse** methods for processing and visualisation.
- **update/extend** raw data and software
- **Track changes and rollback** (Traceability/Recovery)
- **keep old versions of the data** after updates (Reproducibility)
- **snapshot, share, restore** the state of a workspace (Reproducibility)

Terminology and Orchestration

Session, Workflow, Service, Service Pools



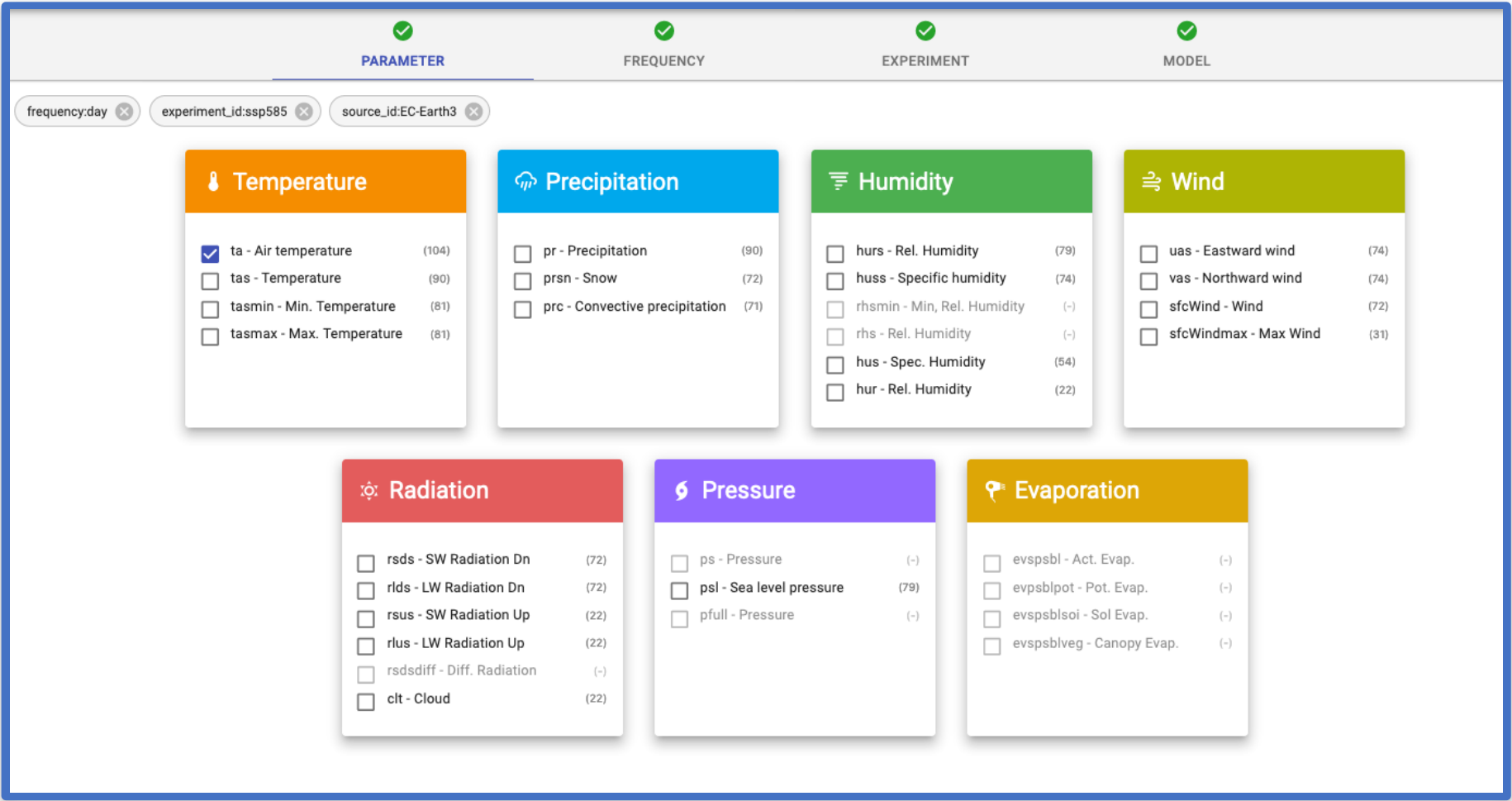
- **A Session is characterised by**
a Raw Data Volume
- **A Workflow is a batch Job**
Stage & preprocess data onto the
Raw Data Volume
- **A Service is an interactive system**
It is attached to the Session (read-
only). It has a Work Volume (r/w)
- **A Service Pool is a collection of**
Services that share the same Work
Volume. The pool is accessed by
users involved in the same research
campaign.



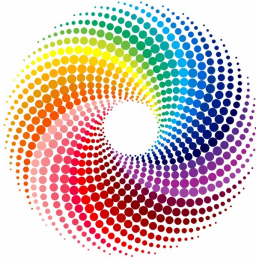
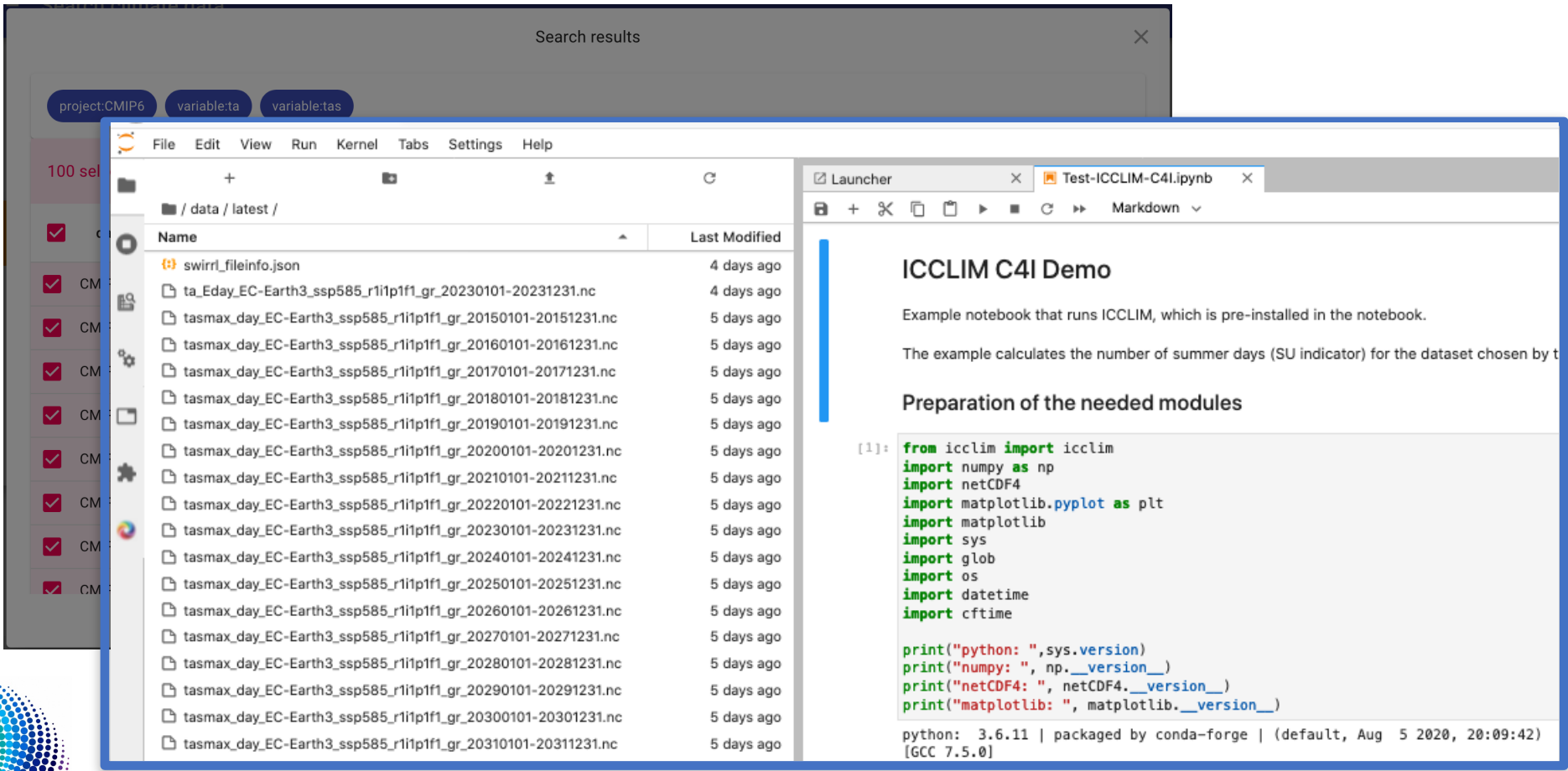
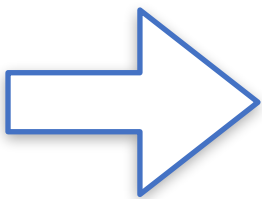


Objective: Extend Community Portals with focussed & reproducible workspaces

Climate4Impact Search for CMIP5/6 Cordex Data
(Distributed Data)



Incremental data staging/subsetting onto customisable and Reproducible
Notebooks (extensible to other tools..)



SWIRRL-API

- Trace Changes to Software and Data
- Restore Environments



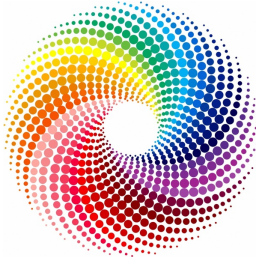
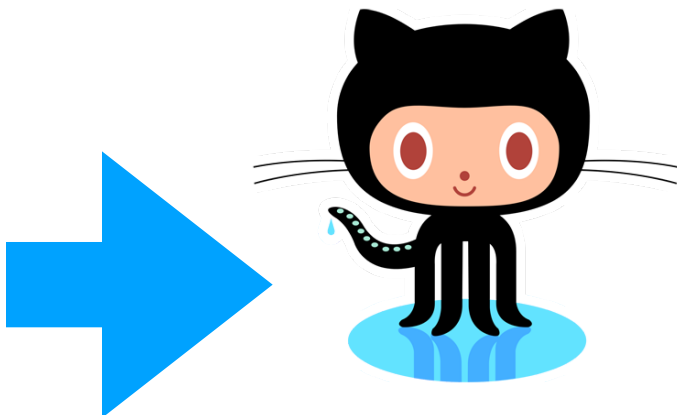
Objective: Extend Community Portals with focussed & reproducible workspaces

Climate4Impact Search for CMIP5/6 Cordex Data
(Distributed Data)

The screenshot shows a web interface for searching climate data. It has tabs for PARAMETER, FREQUENCY, EXPERIMENT, and MODEL. Under the PARAMETER tab, there are several categories of climate variables, each with a list of specific parameters and their counts. The categories are: Temperature (104), Precipitation (90), Humidity (79), Wind (74), Radiation (72), Pressure (79), and Evaporation (72). Each category has a list of parameters with checkboxes and counts.

Incremental data staging/subsetting onto customisable and Reproducible Notebooks (extensible to other tools..)

The screenshot shows a Jupyter Notebook interface. On the left, there is a search results table with columns for Name and Last Modified. The table lists various climate data files. On the right, there is a code cell titled 'ICCLIM C4I Demo' containing Python code for importing modules and running a simulation. The code includes imports for icclim, numpy, netCDF4, matplotlib, sys, glob, os, datetime, and cftime. It also includes print statements for the versions of these modules.



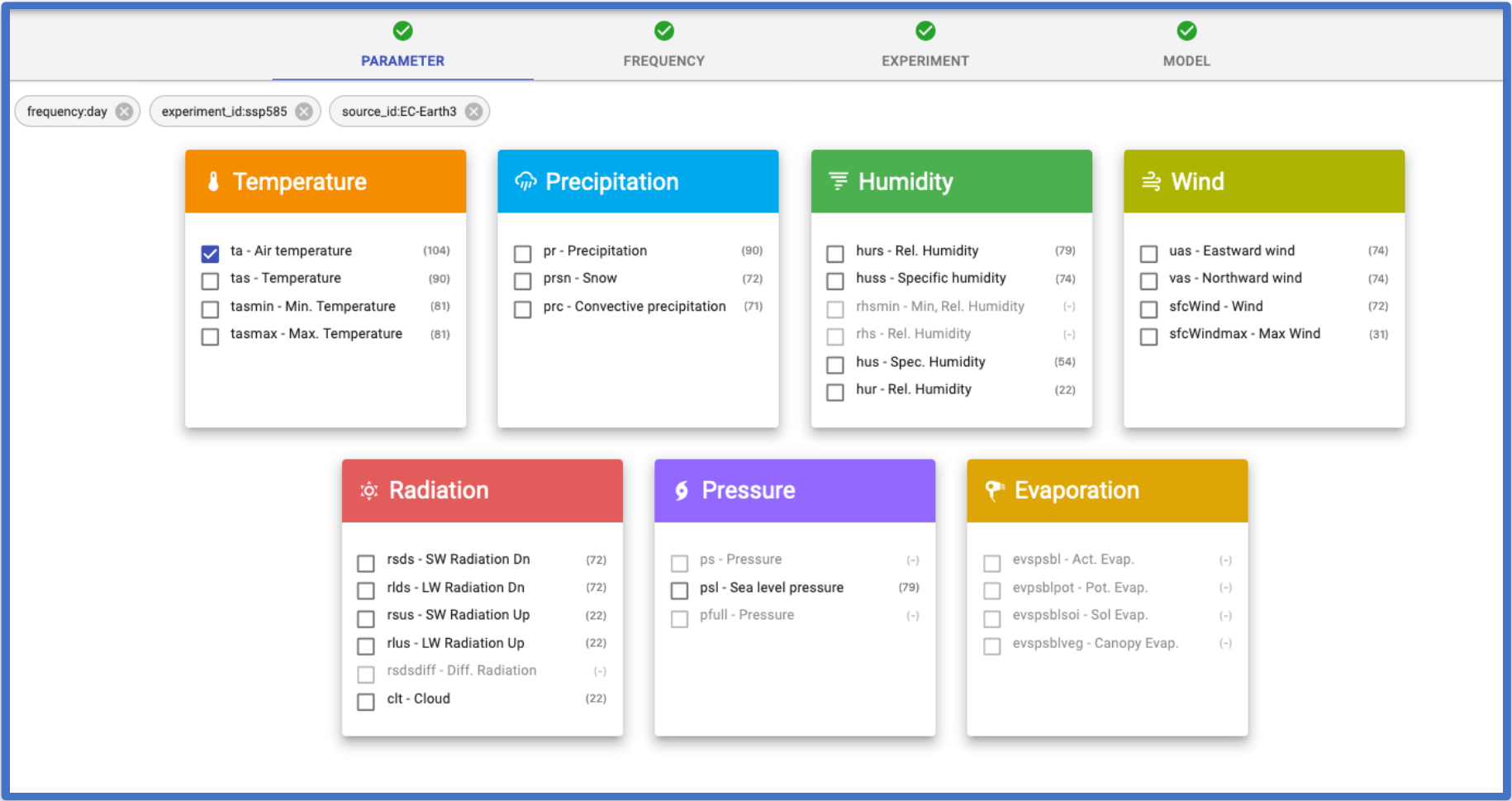
SWIRRL-API

- Trace Changes to Software and Data
- Restore Environments

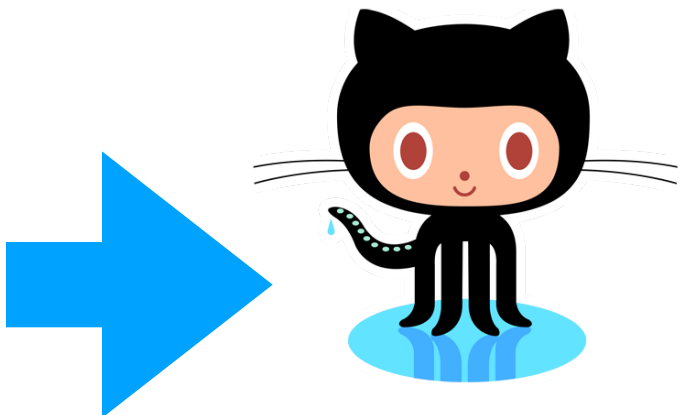
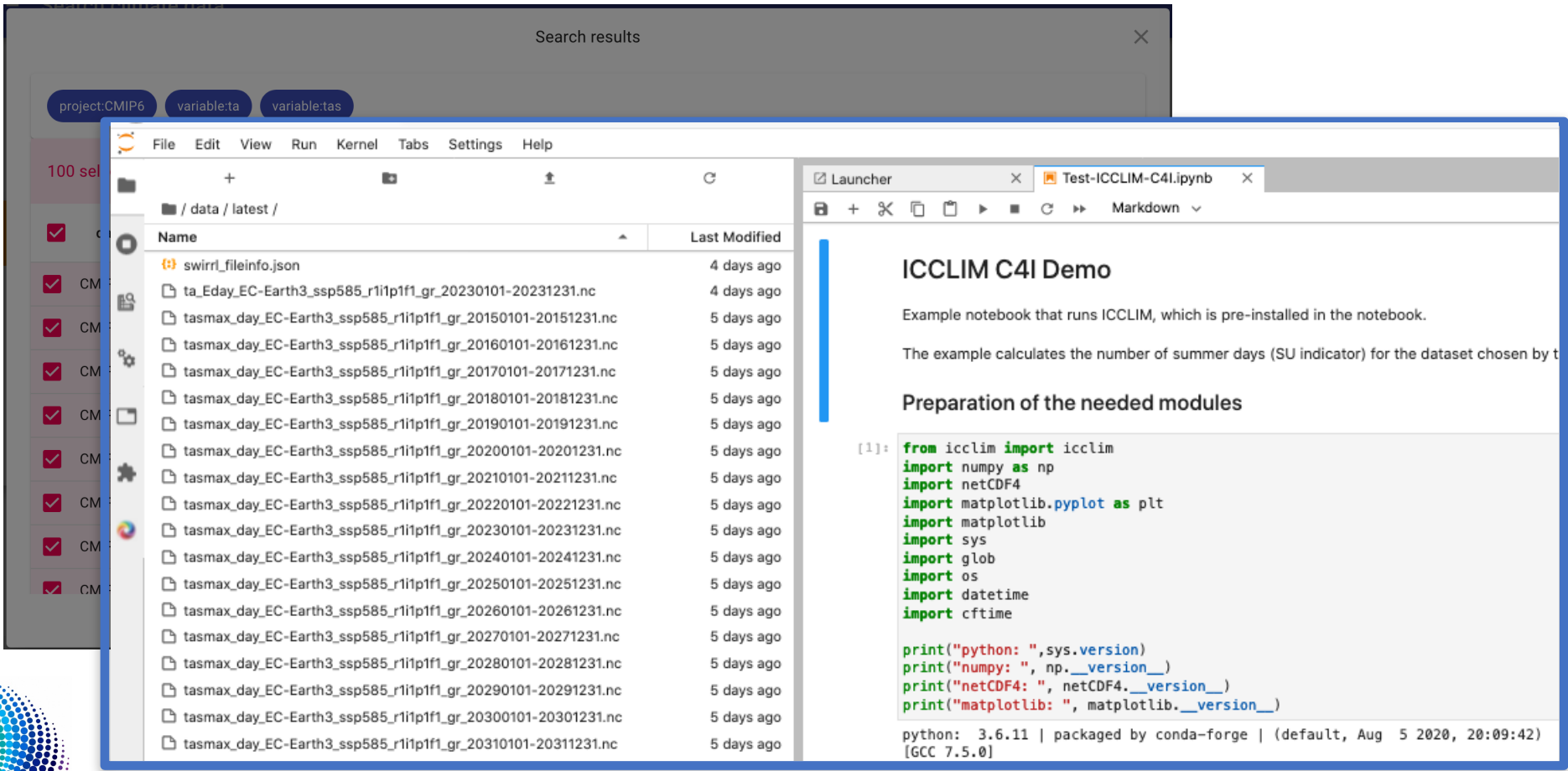


Objective: Extend Community Portals with focussed & reproducible workspaces

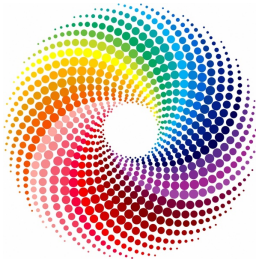
Climate4Impact Search for CMIP5/6 Cordex Data
(Distributed Data)



Incremental data staging/subsetting onto customisable and Reproducible
Notebooks (extensible to other tools..)

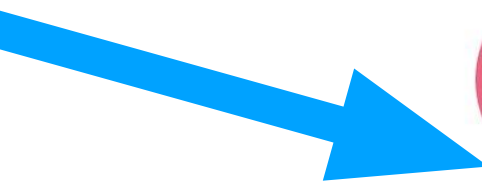


Binder



SWIRRL-API

- Trace Changes to Software and Data
- Restore Environments

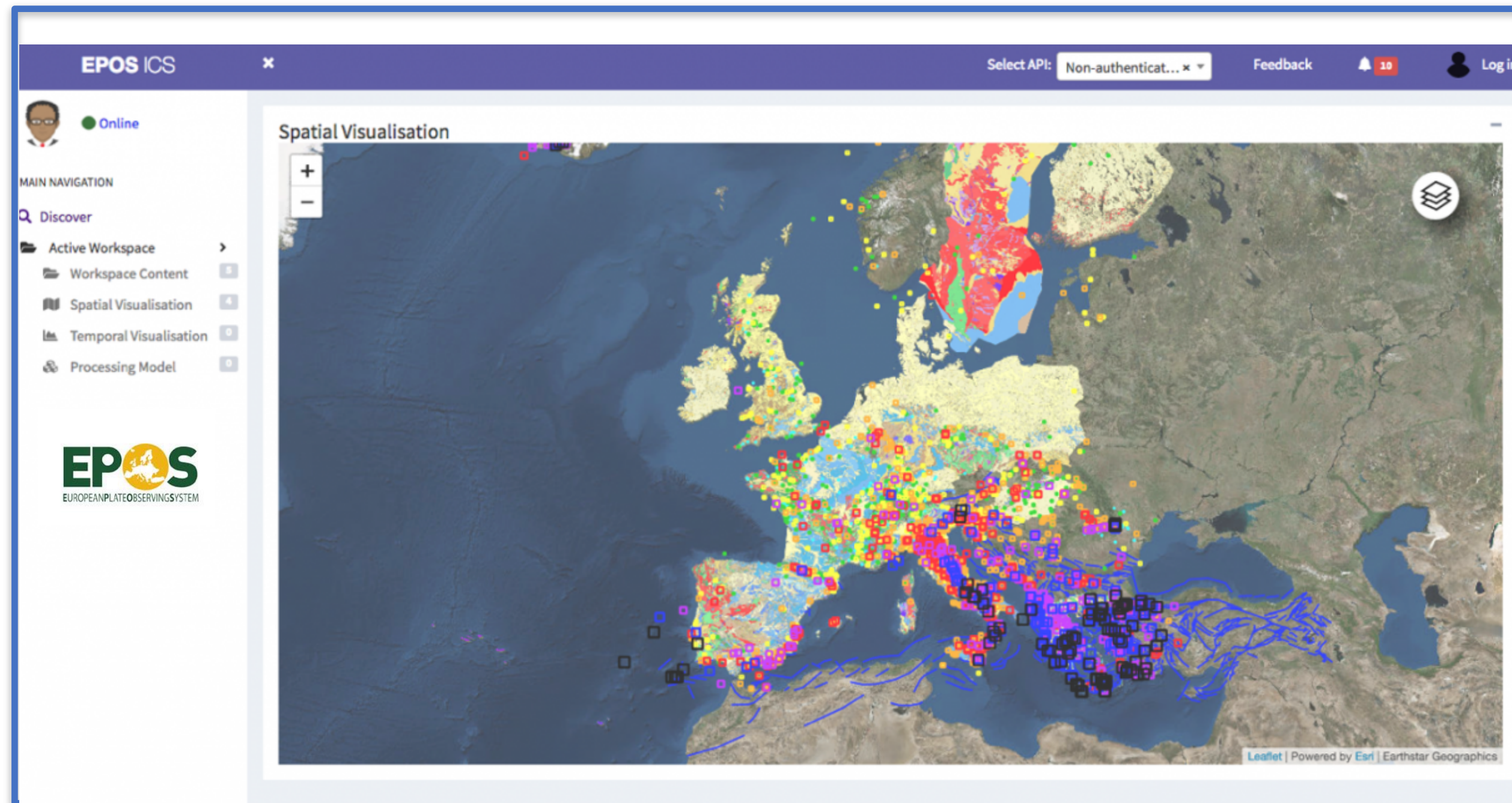


Data



Objective: Extend Community Portals with focussed & reproducible workspaces

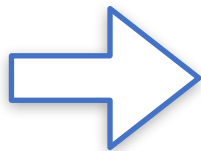
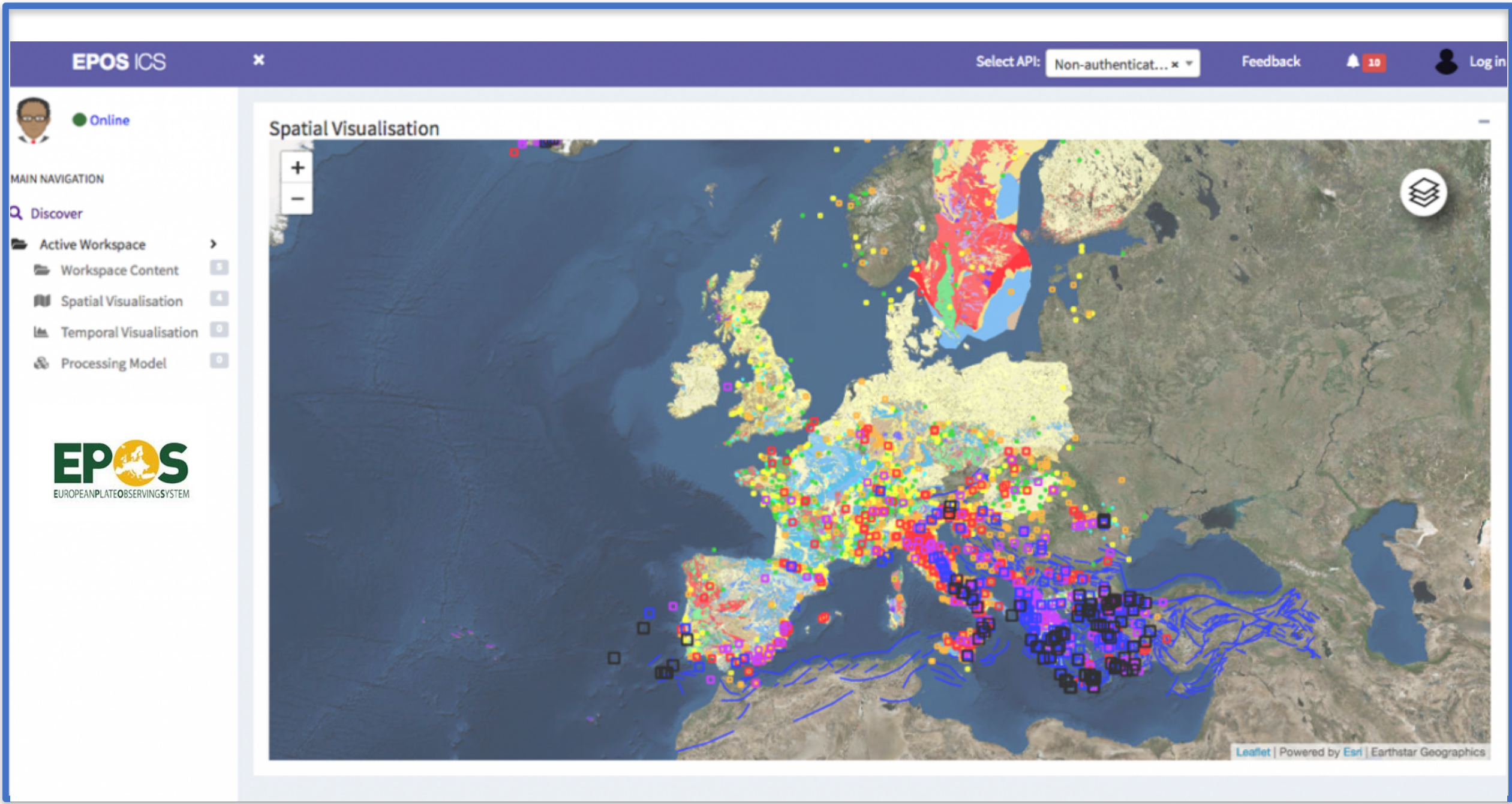
EPOS Distributed Solid Earth Data





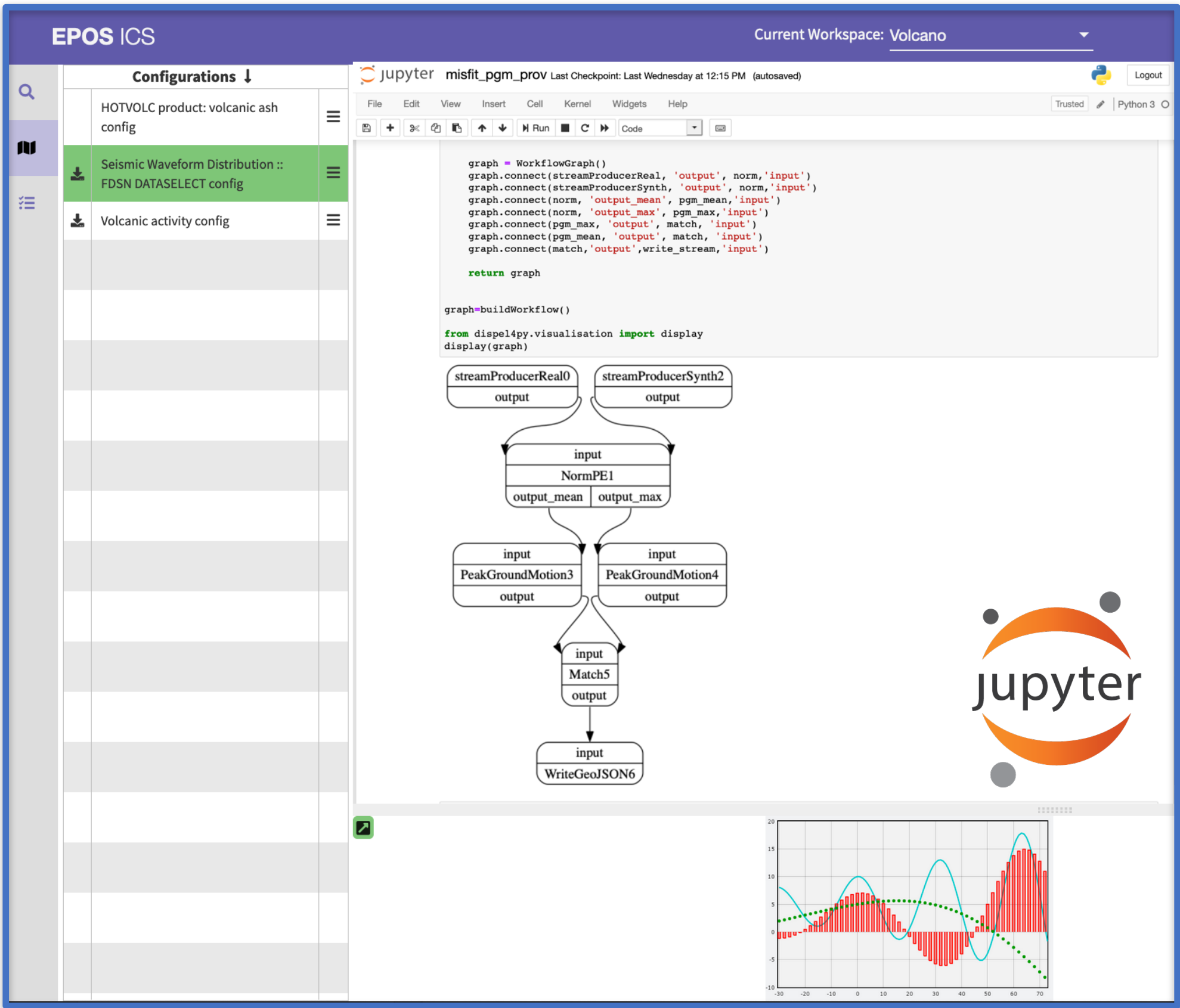
Objective: Extend Community Portals with focussed & reproducible workspaces

EPOS Distributed Solid Earth Data



Combined Services around Datasets

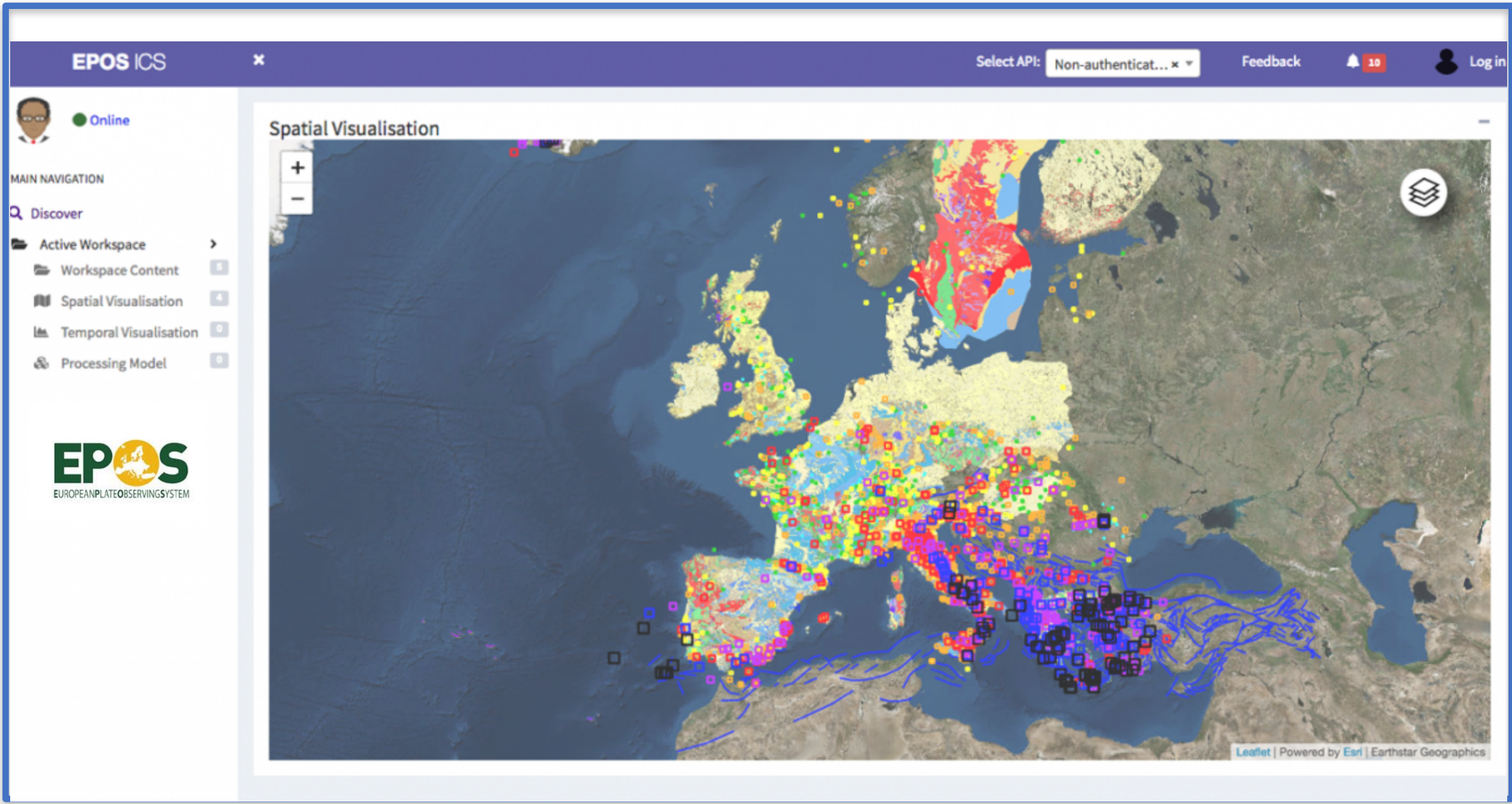
Documented Analysis, Workflows, Visualisation





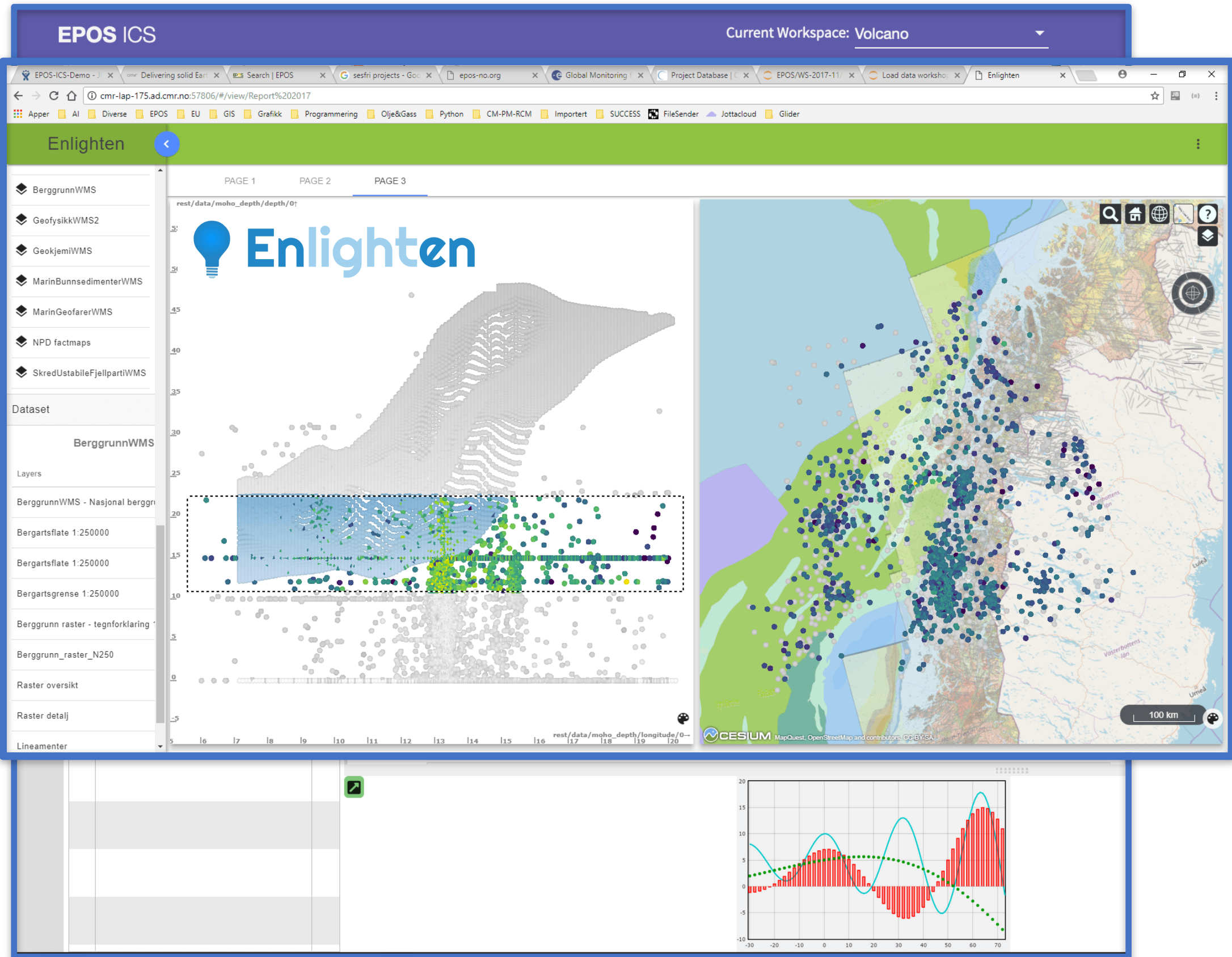
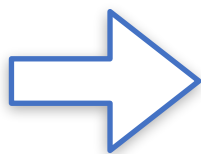
Objective: Extend Community Portals with focussed & reproducible workspaces

EPOS Distributed Solid Earth Data



Combined Services around Datasets

Documented Analysis, Workflows, Visualisation





notebook

provide/manage a service environment in a Session/ServicePool

▼

POST	/notebook	Deploys a notebook and creates a data and working directory.
POST	/notebook/{id}/snapshot	Creates a notebook snapshot
PUT	/notebook/{id}/restorelibs/{activityId}	Restore Notebook libraries from a previous update
PUT	/notebook/{id}	Update Notebook libraries
GET	/notebook/{id}	Retrieves the running notebook for the given notebookId.
DELETE	/notebook/{id}	Deletes notebook for the given notebookId.

Run/Monitor Workflows in a Session



notebook

provide/manage a service environment in a Session/ServicePool

▼

POST	/notebook	Deploys a notebook and creates a data and working directory.
POST	/notebook/{id}/snapshot	Creates a notebook snapshot
PUT	/notebook/{id}/restorelibs/{activityId}	Restore Notebook libraries from a previous update
PUT	/notebook/{id}	Update Notebook libraries
GET	/notebook/{id}	Retrieves the running notebook for the given notebookId.
DELETE	/notebook/{id}	Deletes notebook for the given notebookId.

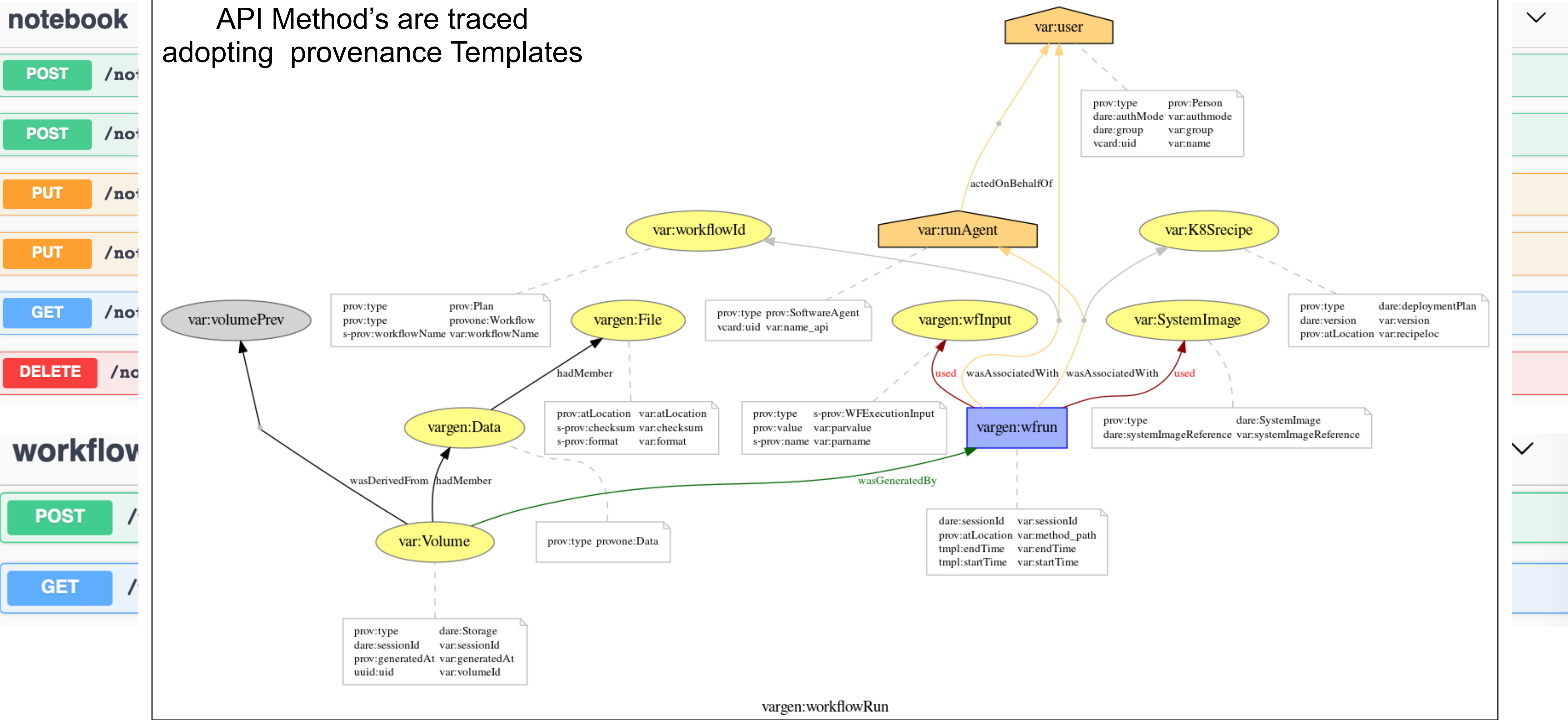
workflow

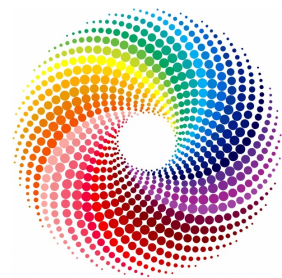
Run/Monitor Workflows in a Session

▼

POST	/workflow/{workflowId}/run/	Runs the specified workflow on a work space (notebook)
GET	/workflow/{workflowId}/run/{runId}/	Return the status of the workflow given by the given runId.







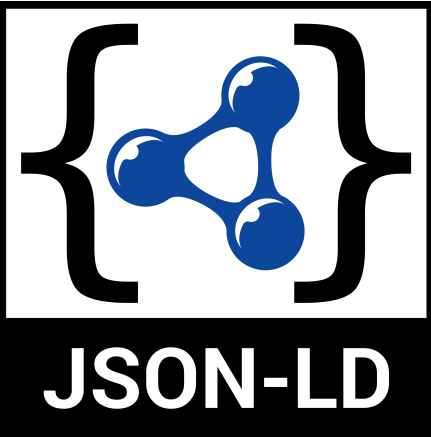
Provenance Storage, Acquire, Query, Exploit

provenance

POST /provenance Saves provided provenance template expansion.

GET /provenance/session/{sessionId}/activities Returns the list of activities that are related to the session with their id's properties

GET /provenance/activity/{activityId} Returns the information about an activity based on the activityId



Provenance info
is searchable
accessible/
actionable

FAIRness of
Operations

```
{ "@graph": [
  { "swirrl:serviceId": "15c0e8a0-19f4-43ff-a619-b99845c01280",
    "swirrl:sessionId": "a99f05e9-7570-46a3-b046-3c442ab1e23d",
    "@type": [
      "Resource",
      "prov:Activity",
      "swirrl:CreateNotebook"],
    "prov:person": "johnp",
    "@id": "urn:uuid:375fe1aa-7957-4f19-81de-6549962fdbb4",
    "prov:endedAtTime": "2020-12-01T14:32:05.000Z",
    "prov:startedAtTime": "2020-12-01T14:31:21.337Z"},
  { "swirrl:serviceId": "15c0e8a0-19f4-43ff-a619-b99845c01280",
    "swirrl:sessionId": "a99f05e9-7570-46a3-b046-3c442ab1e23d",
    "@type": [
      "Resource",
      "prov:Activity",
      "swirrl:Update"],
    "prov:person": "johnp",
    "@id": "urn:uuid:daddea98-eb60-40d2-98e6-b2f026348da7",
    "prov:endedAtTime": "2020-12-01T14:32:46.137Z",
    "prov:startedAtTime": "2020-12-01T14:32:36.252Z"}
], "@context": {...}}
```

SWIRRL-API (Provenance)



provenance

POST

/provenance

GET

/provenance

GET

/provenance

Provenance info
is searchable
accessible/
actionable

FAIRness of
Operations

File Edit View Run Kernel Tabs Settings Help

SWIRRL

Notebook idle

Github

LOGIN

Please review your access using [this link](#) to revoke your access tokens.

Create Snapshot

Snapshot name

SNAPSHOT

Data staging

ROLL BACK

Activities

LOAD ACTIVITIES

Type	Created at	Action
Library Update	2021-03-18 15:43	RESTORE
Snapshot	2021-03-18 15:42	OPEN
Workflow	2021-03-17 08:49	
Workflow	2021-03-17 08:46	
Workflow	2021-03-17 08:44	

Launcher

swirrl_fileinfo.js

swirrl_fileinfo.js

Test-ICCLIM-C4I

huss_AFR-44_E

hurs_AFR-44_E

Python 3

ICCLIM C4I Demo

Example notebook that runs ICCLIM, which is pre-installed in the notebook.

The example calculates the number of summer days (SU indicator) for the dataset chosen by the user on C4I

Preparation of the needed modules

```
[ ]: from icclim import icclim
import numpy as np
import netCDF4
import matplotlib.pyplot as plt
import matplotlib
import sys
import glob
import os
import datetime
import cftime

print("python: ", sys.version)
print("numpy: ", np.__version__)
print("netCDF4: ", netCDF4.__version__)
print("matplotlib: ", matplotlib.__version__)
```

Specification of the parameters and period of interest

```
[ ]: # studied period
dt1 = datetime.datetime(2015,1,1)
dt2 = datetime.datetime(2019,12,31)

out_f = 'su_icclim.nc'
filenames = glob.glob('./data/latest/tasmax_day*.nc')

icclim.indice(indice_name='SU', in_files=filenames, var_name='tasmax', slice_mode='JJA', time_range=dt1, time_end=dt2)

matplotlib.use('agg')

nc = netCDF4.Dataset(out_f)

time = nc.variables['time']
var = nc.variables['SU']

time = cftime.num2date(time[:], time.units, time.calendar)
year_list = np.array([t.year for t in time])
```

Simple

1

Python 3 | Idle

Mode: Command

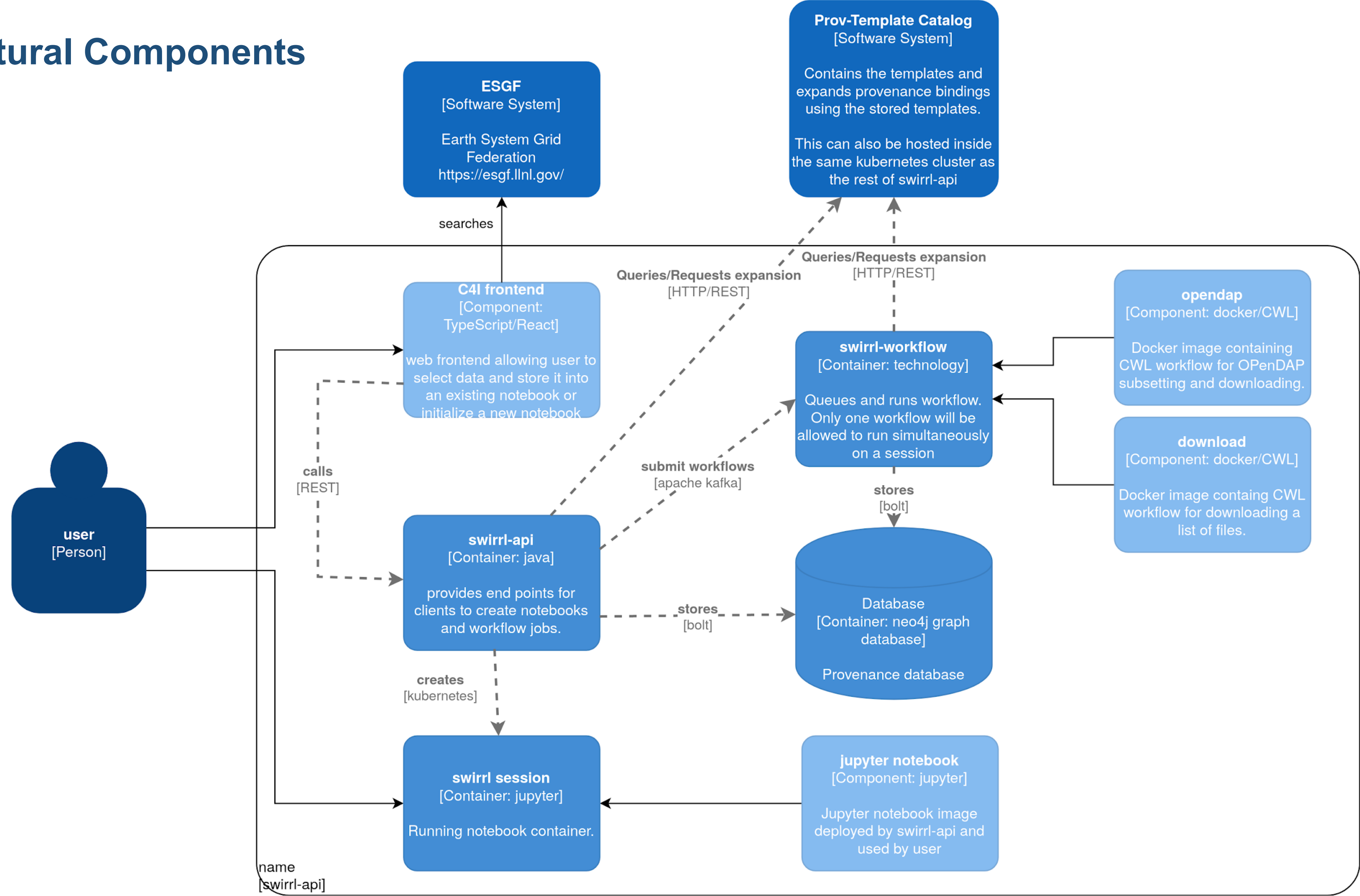
Ln 1, Col 1

Test-ICCLIM-C4I.ipynb

actionable
provenance



Architectural Components



Future Work



- **AAI in C4I/SWIRRL**
(Deployment Based - C4I/EPOS)
- **Restore Snapshot from Git onto SWIRRL**
- **Registration of new workflows**
- **Collaborations & Use Cases**
(remote subsetting workflows WPS, analysis and visual analytics of seismic data - Enlighten integration)