



CEOS Interoperability Initiatives

WGDisasters-14
September 2020

Dr. Robert Woodcock
CEOS WGISS Chair
CSIRO



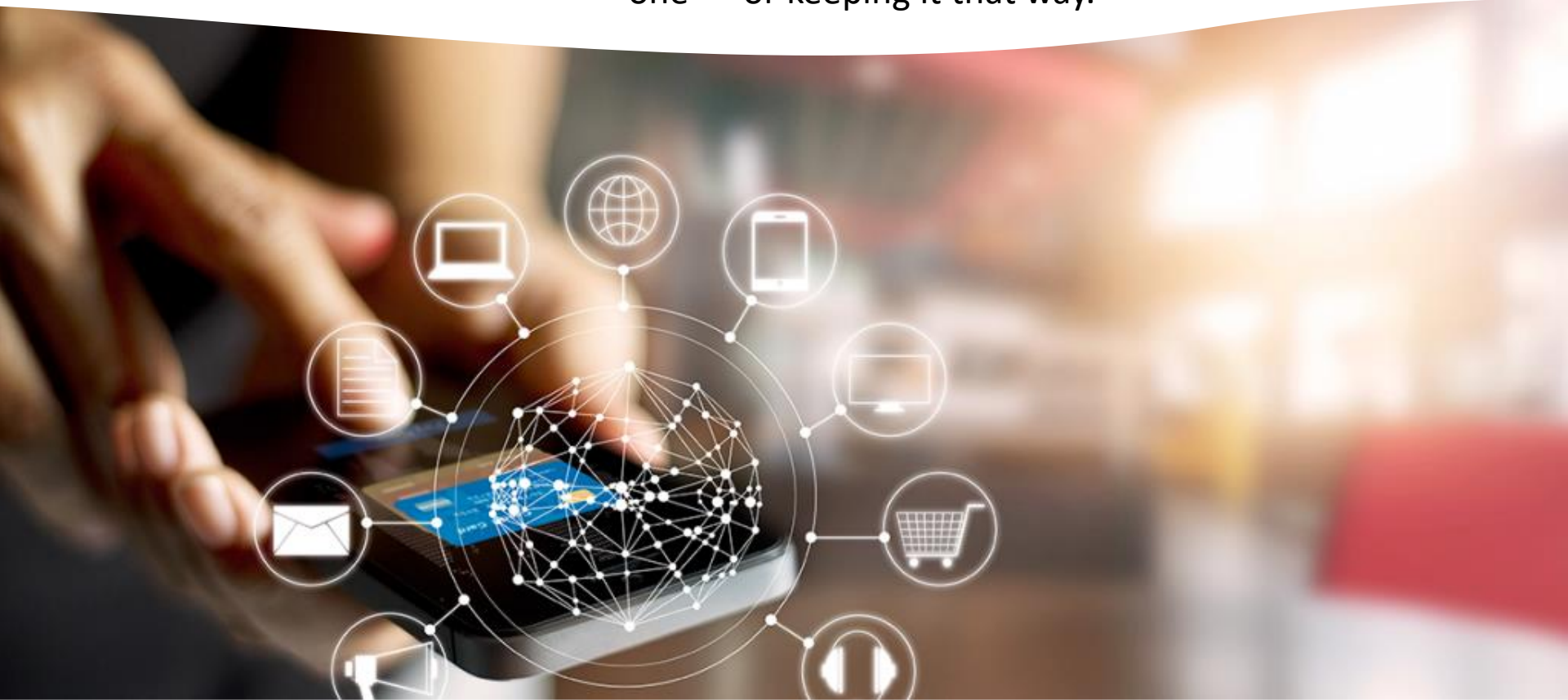
WGISS Interoperability Initiatives WIP

- CEOS Interoperability Terminology Report
- CEOS Earth Analytics Interoperability Lab
- Best Practices in:
 - Jupyter Notebooks for EO Analytics
 - Cloud data formats
 - CEOS Data Discovery and Access and Cloud



What is Interoperability?

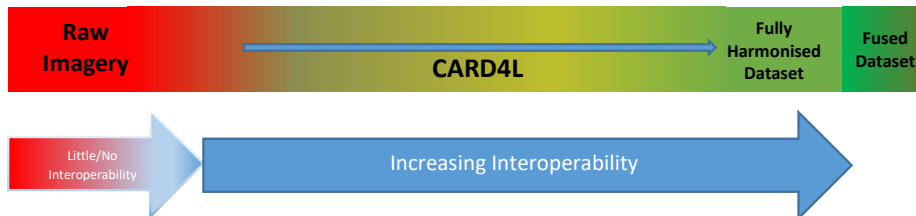
- CEOS Interoperability Terminology Report (Draft):
 - ...the terms Analysis Ready Data (ARD), interoperability, and harmonization ***are often used and, to a large extent, used inconsistently***
 - Interoperability represents a continuum of compatibility for products, services, algorithms.
- “Making an unmanageable situation a manageable one” – or keeping it that way.



CEOS Interoperability Terminology Report

Five ARD terms are defined in this document:

1. Analysis Ready Data (ARD)
2. CEOS ARD for Land (CARD4L) Products
3. Interoperable Products
4. Harmonised Products
5. Fused Products



CEOS Interoperability Terminology Report

Three Cloud data format terms for ARD:

1. Cloud-friendly
2. Cloud-native
3. Cloud Data Access API

Three types of Analysis Interoperability:

1. Executable code (black box package)
2. Source code
3. Algorithm

Why do we need an Interoperability Lab?

- A significant number of CEOS activities are now engaged in the CEOS ARD and FDA strategies and in Integrated Earth observation data analysis
- ***All recognize the broader implications*** – ARD relates to *Discovery and Access, Analytics & Cloud Data Formats*
- **Many are looking for advice or implementations.** It can be difficult for a single project to cover all viewpoints and technologies.
- **Validating interoperability between multiple CEOS organizations and working groups is complex.**



The Opportunity

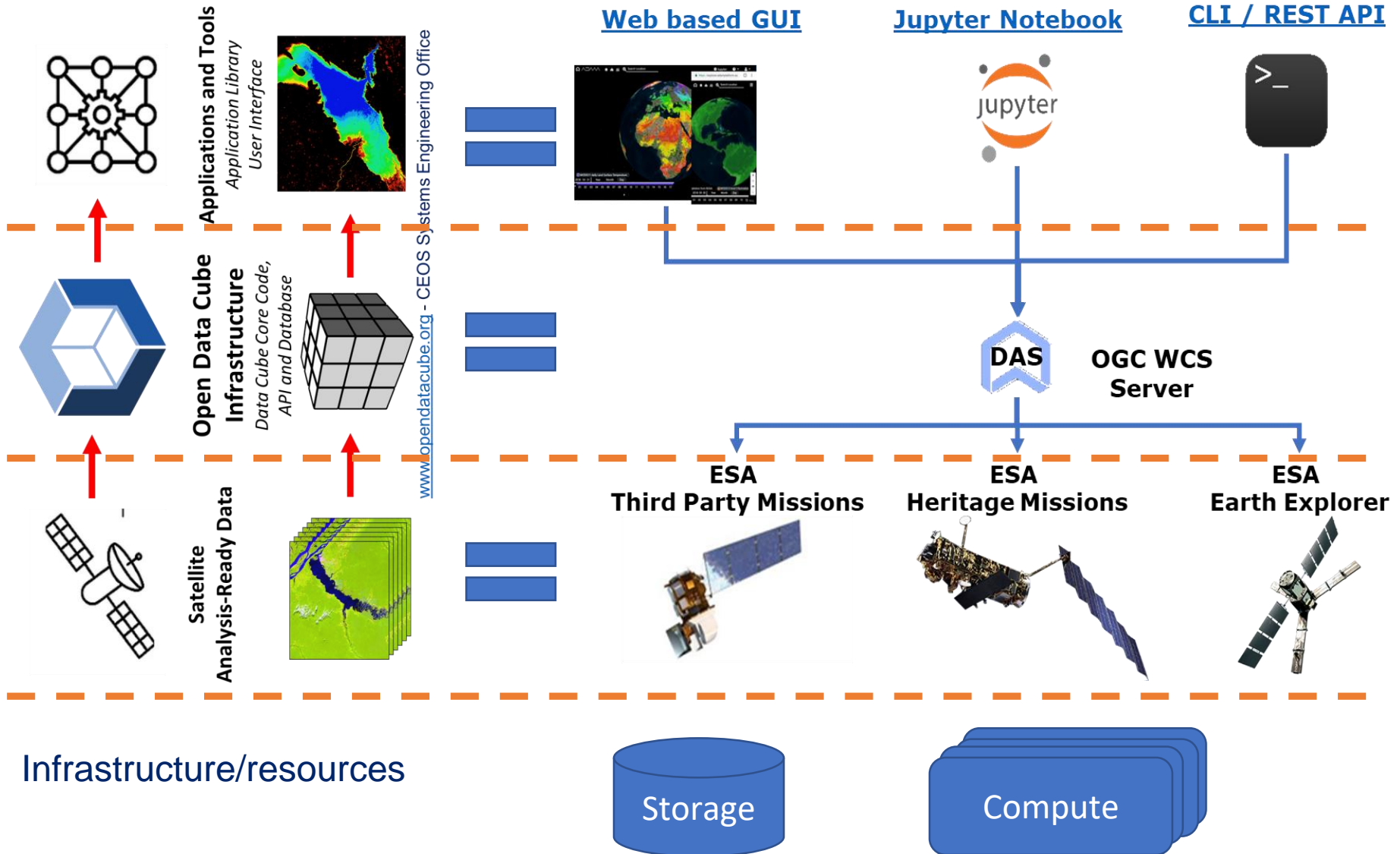
- The first Lab node is hosted by CSIRO and jointly operated by WGISS and SEO to provide:
 - Examples of FDA components in active use
 - A shared candidate ARD storage and access capability
 - Connection to existing and emergent services for Discoverability and Access
 - Collaboration on Analytics tools for integrated analysis using Jupyter Notebooks
 - Jointly develop CEOS Best Practices via interop experiments
- In doing this, **validate interoperability approaches**



Additional nodes?

CEOS SEO Node

ESA Node (TBC)



Jupyter Labs – exploratory data analytics

WGISS Survey on use of Jupyter in CEOS agencies

Spawner Options

EASI Open Data Cube (CSIRO variant) environment
The EASI Open Data Cube with CSIRO variants. ODC 1.7
EASI 1.0.5

Scipy environment
To avoid too much bells and whistles: Python, r...

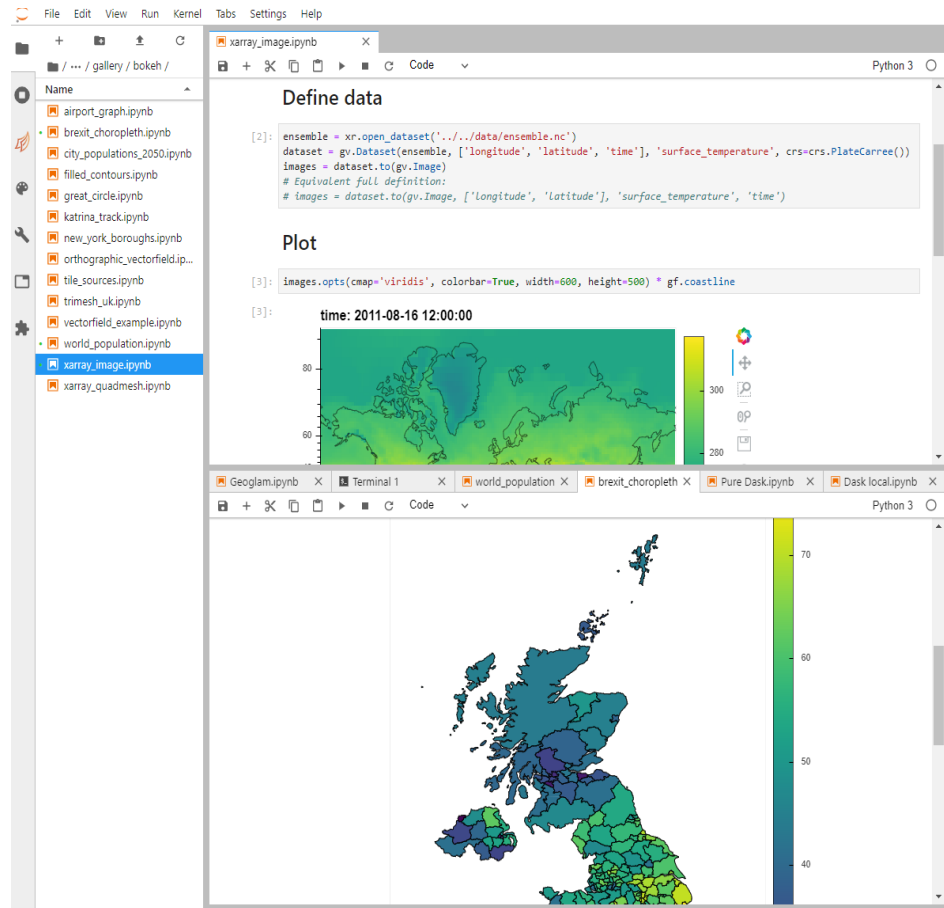
Datascience environment
Includes libraries for data analysis and visualization. Julia, Python, and R communities.

Python deep learning environment
Includes a SciPy environment plus tensorflow and keras learning libraries.

EXPERIMENTAL - EASI Open Data Cube (CSIRO variant) environment
The EASI Open Data Cube with CSIRO variants. ODC 1.7
EASI Latest

Spawn

Customise per CEOS activity as necessary



Interactive visualisation

Dashboards (from notebook)

The screenshot displays a JupyterLab environment with the following components:

- File Explorer:** Shows three open notebooks: `xarray_image.ipynb`, `Pure Dask.ipynb`, and `Dask local.ipynb`.
- Code Editor:** Contains Python code for data selection and NDVI calculation. A tooltip provides instructions on plotting data from cloud storage.
- Output View:** Currently empty.
- Dask Workers Dashboard:** A table showing the status of various Dask workers.
- Dask Progress Dashboard:** A summary of the overall Dask task progress.

```
4.9e+06 -4.9e+06
* time      (time) datetime64[ns] 2018-06-19 2018-07-05 ... 20
19-09-10
Data variables:
  B4      (time, y, x) float64 dask.array<shape=(24, 7901, 8
041), chunksize=(1, 2048, 2048)>
  B5      (time, y, x) float64 dask.array<shape=(24, 7901, 8
041), chunksize=(1, 2048, 2048)>

[ ]: # Here is the syntax to plot the same image as before
# Again, the actually image data is downloaded from cloud storage
da = DS.sel(time='2019-06-22')['B4']
print('Image size (Gb): ', da.nbytes/1e9)
da.plot.imshow()

[23]: NDVI = (DS['B5'] - DS['B4']) / (DS['B5'] + DS['B4'])
```

Worker	Address	Free	Used	Busy	Idle	Memory	Local	Progress
tcp://10.0.40.16	tcp://10.0.40.16	2	0.0 %	415 MiB	7 GiB	5.8 %		
tcp://10.0.42.83	tcp://10.0.42.83	2	2.0 %	520 MiB	7 GiB	7.3 %		
tcp://10.0.51.15	tcp://10.0.51.15	2	2.0 %	43 MiB	7 GiB	0.6 %		
tcp://10.0.58.25	tcp://10.0.58.25	2	0.0 %	368 MiB	7 GiB	5.1 %		
tcp://10.0.73.65	tcp://10.0.73.65	2	2.0 %	348 MiB	7 GiB	4.9 %		
tcp://10.0.79.15	tcp://10.0.79.15	2	2.0 %	421 MiB	7 GiB	5.9 %		
tcp://10.0.81.12	tcp://10.0.81.12	2	2.0 %	44 MiB	7 GiB	0.6 %		
tcp://10.0.95.15	tcp://10.0.95.15	2	0.0 %	43 MiB	7 GiB	0.6 %		
tcp://10.0.98.20	tcp://10.0.98.20	2	0.0 %	43 MiB	7 GiB	0.6 %		

Progress -- total: 0, in-memory: 0, processing: 0, waiting: 0, erred: 0

Cloud data formats – Best Practices

Direct data access

- Compatibility (ARD, Geotiff)

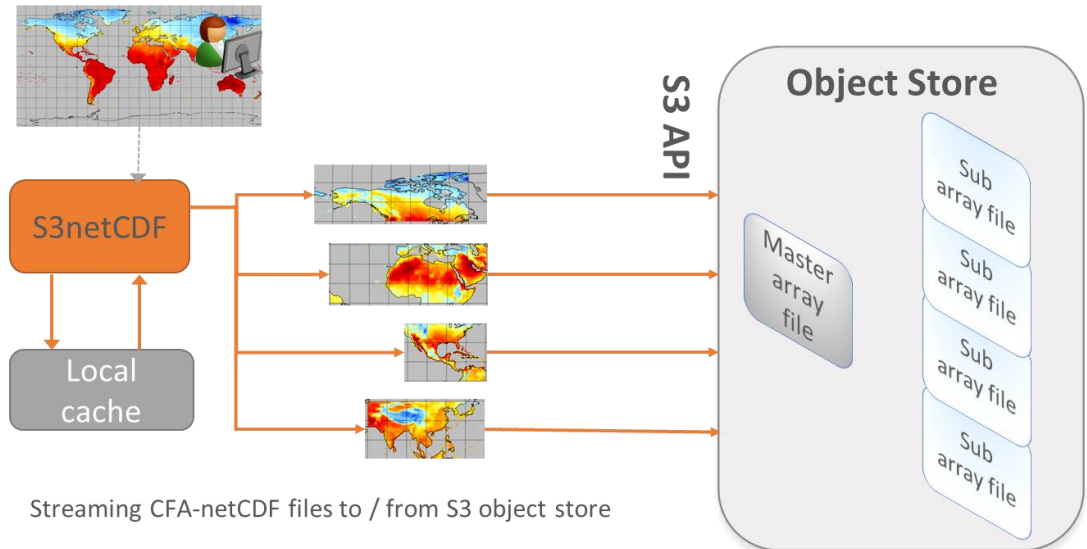
Analytics APIs

- Services like OGC WPS
- Direct compute access:
 - python Dask

• Cloud Compute Ready Data

- COGS – Cloud optimised geotiffs
- NetCDF in the Cloud
- Zarr – Cloud native array storage

Files split into CFA-netCDF sub-array files using the variable splitting algorithm



CEOS Data in Cloud – Developing Best Practices

Landsat 8

disaster response earth observability

Description

An ongoing collection of data provided by the Landsat 8 satellite

Update Frequency

New Landsat 8 scenes are available every 16 days

License

There are no restrictions on the use of this data. The data is provided by the Geological Survey's Earth Resources Observation and Science Center or NASA's Land Processes Distributed Active Archive Center (DAAC), unless expressly stated otherwise. More information on licensing is available from USGS.

Documentation

<https://docs.opendata.aws/landsat8/>

Managed By



Resources on AWS

Description

Scenes and metadata

Resource type

S3 Bucket

Amazon Resource Name (ARN)

`arn:aws:s3:::landsat-pds`

AWS Region

`us-west-2`

Description

New scene notifications

Resource type

SNS Topic

Amazon Resource Name (ARN)

`arn:aws:sns:us-west-2:274514004127`

AWS Region

`us-west-2`

Sentinel-2

disaster response earth observability

Description

The Sentinel-2 mission is a satellite mission that provides high resolution imagery with temporal continuity for the current Sentinel-2 mission. Sentinel-2 provides a global coverage of the Earth's land surface, making the data of great use for a wide range of applications. The data is available from June 2015 globally and from 2017 over wider Europe region.

Update Frequency

New Sentinel data are added to the bucket as they are available on Copernicus.

License

Access to Sentinel data is free of charge. The data is provided by the National, European and International Copernicus Conditions.

Documentation

Documentation is available at <https://sentinel2.copernicus.eu/>

Managed By



Resources on AWS

Description

Level 1C scenes and metadata in Requester Pays S3 bucket

Resource type

S3 Bucket Requester Pays

Amazon Resource Name (ARN)

`arn:aws:s3:::sentinel-s2-l1c`

AWS Region

`eu-central-1`

Description

S3 Inventory files for L1C (ORC and CSV)

Resource type

S3 Bucket

Amazon Resource Name (ARN)

`arn:aws:s3:::sentinel-inventory/sentinel-s2-l1c`

AWS Region

`eu-central-1`

Roadmap – Next steps

- Roadmap:
 - WGISS+SEO deploy the lab
 - CEOS COAST and others: inventory of data, analytics, etc
 - Define joint interop experiments for capabilities needed by these projects
 - Demonstrate and validate jointly in the Lab(s)
 - Cookbook of Jupyter notebook examples
 - Plenary endorse first edition of CEOS Interoperability Terminology and promote uptake
 - ARD terms already updated in WGISS Data Preservation Glossary

