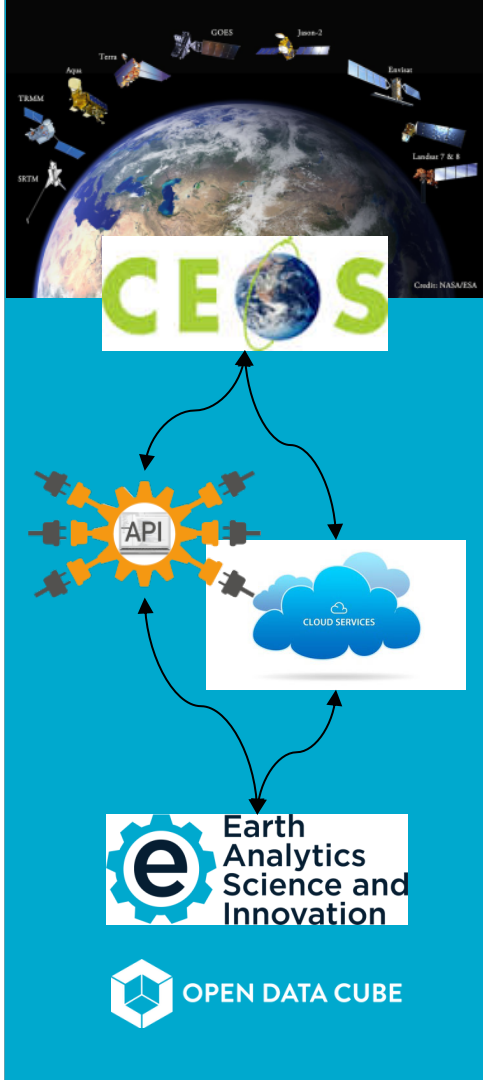
# CSIRO EASI Hub data-pipelines

Matt Paget, Jonathan Hodge, Peter Wang, Robert Woodcock

CEOS WGISS-50, 22-24 September 2020

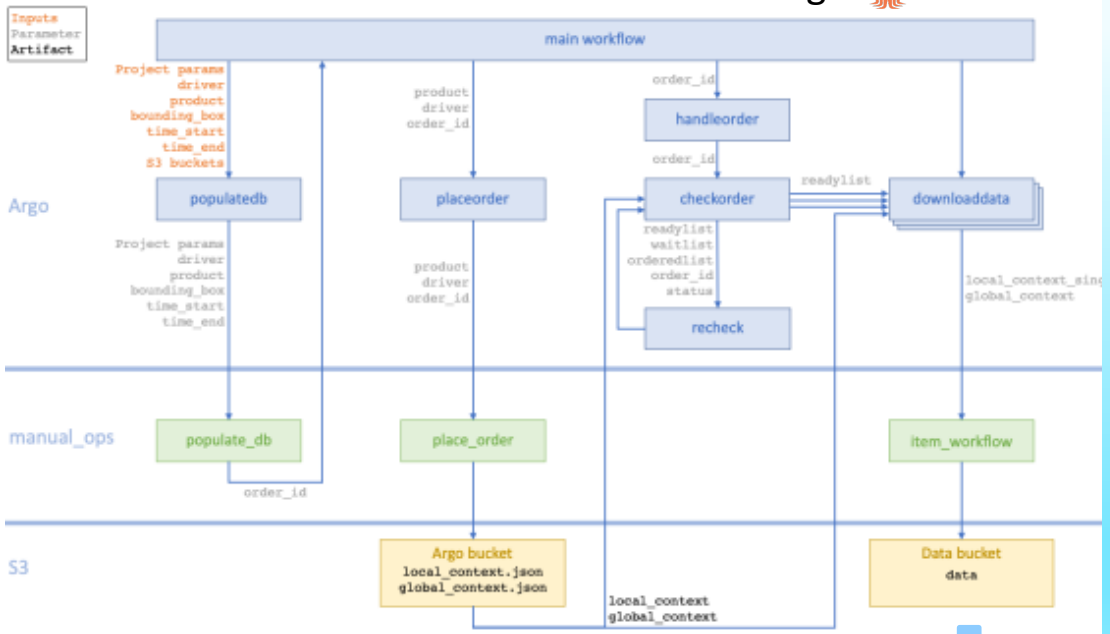
# Overview and intent

- Consume CEOS data in a hybrid world of APIs and Clouds
- Simplify the task of download, pre-process and prepare for a data cube instance
- Modular, scalable and schedulable workflows
- Manage caches of data for current use, rather than build local archives



# Design

Cloud-native workflows with Argo




Datacube indexing is a separate step

- Common interface between order request for platform + product + space + time and agency APIs
- Optional ‘tasks’ after download for data management, pre-processing (ARD), reformatting (COG, Zarr) and preparing (data cube)
- State managed through dict/JSON/YAML (flexible, idempotent)
- Python code with Argo Workflows for large scale orchestration in Kubernetes and EASI Hub

# Ecosystem

- data-pipelines is not a unique idea (its just our implementation)
- API python libraries, or at least programmable interfaces, exist for most CEOS archives
  - data-pipelines will use these, or borrow from them

- Current implementations for:
  - USGS ESPA
  - USGS AppEEARS
  - Copernicus Open Access Hub
  - GA AWS public
  - NCI Aust CopHub

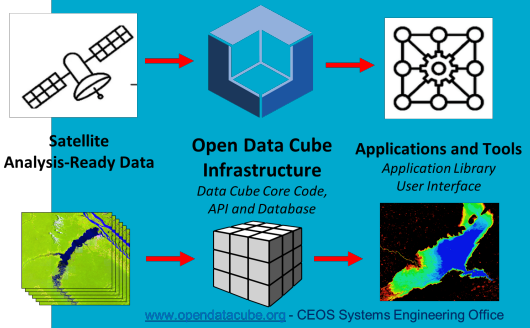
In development:

- NovaSAR
- Himawari
- CEOS OpenSearch
- MODIS/VIIRS L2
- Landsat C2
- ARD chooser
- GEDI
- Prisma
- …

# Trajectory

- CEOS agencies are moving more data resources into the cloud
  - Plus mature API interfaces
  - Plus global ARD (“agency standard”)

- No need for consumers to build their own data archives, provided connection to cloud is 'fast enough’ to work



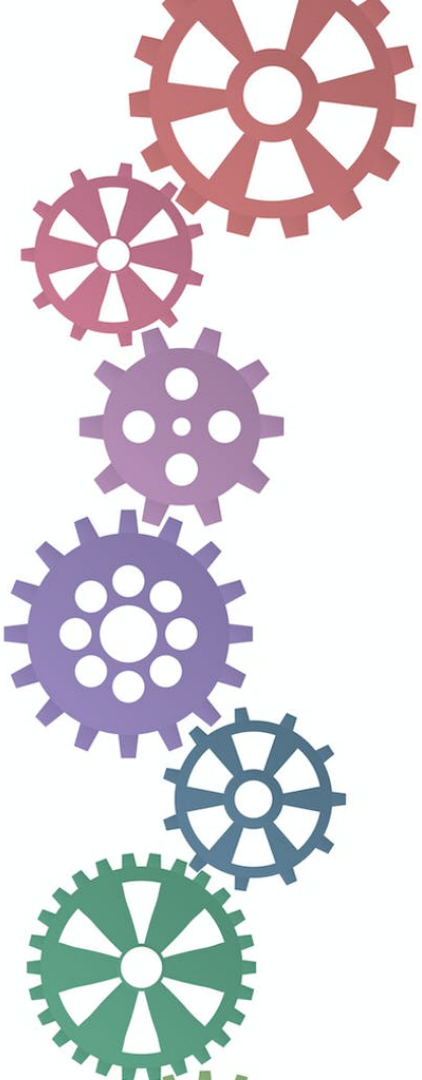

*CEOS Goal?*

- *Fewer common interfaces access more data sources*

# Towards more interoperability

## *.. and data users*

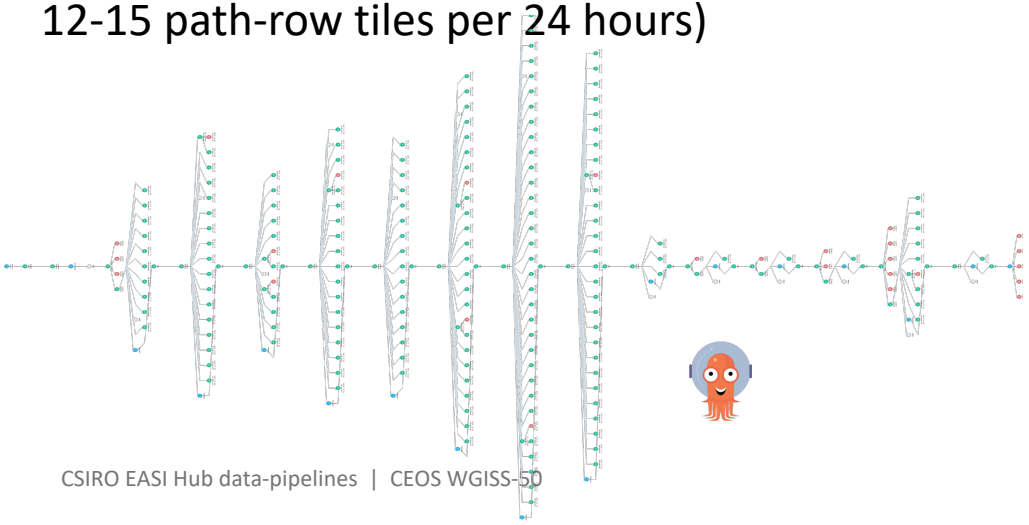
- ARD processing history / provenance
  - We often hear “we do ARD” but which ARD?
    - algorithm, version, ancillary data
  - Empower or encourage consumers to pass on the ARD specifics
- CEOS OpenSearch is great
  - Except when granules are not available
- Same data, different clouds
  - Multiple access is great but how can consumers confirm, reconcile or choose when there are differences?
- Some CEOS data in the cloud are not readily consumable or optimised (packaging, format)
  - Limits the advantages of using the cloud
  - Cloud regions and user-pays buckets: a concern for consumers

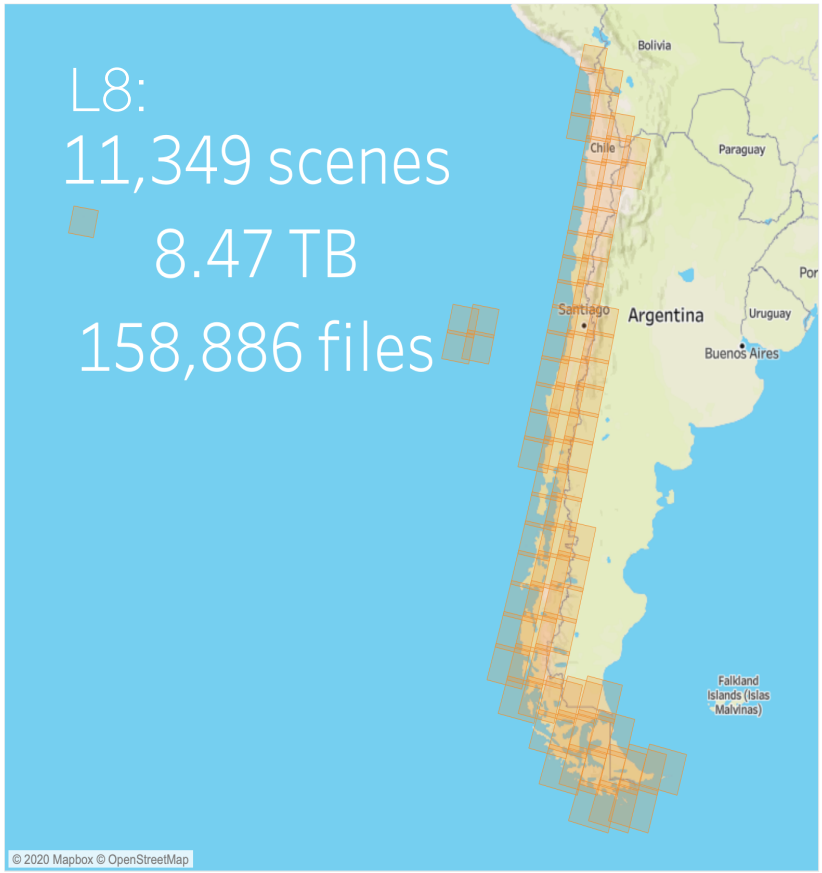


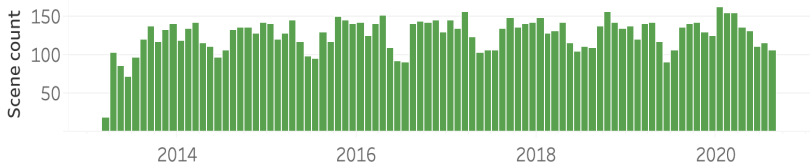
# Chile coverage

Data preparation for the Data Observatory initiative

- Landsat 5, 7, 8 SR and AR processed by USGS ESPA API
- Automated ordering, download and resilience by EASI data-pipelines software
- Chile coverage built over 1 week (L8 full series, 12-15 path-row tiles per 24 hours)





# Thank you


**Land & Water**
Matt Paget
matt.paget@csiro.au

**CSIRO Chile**
Jonathan Hodge
jonathan.hodge@csiro.au

**Centre for Earth Observation**
Robert Woodcock
robert.woodcock@csiro.au

**Data61**
Peter Wang
peter.wang@csiro.au


https://mikeroysoft.com/blog/pipeline-1