Advancing collaborative data stewardship with Python, Jupyter, Git

Examples from recent data releases at the **Physical Oceanography Distributed Active Archive Center (PO.DAAC)**

Jack McNelis (jmcnelis@jpl.nasa.gov), David Moroni, Ed Armstrong, Wen-Hao Li, Nga Chung, and the rest of the PO.DAAC
Jet Propulsion Laboratory, California Institute of Technology, Pasadena CA

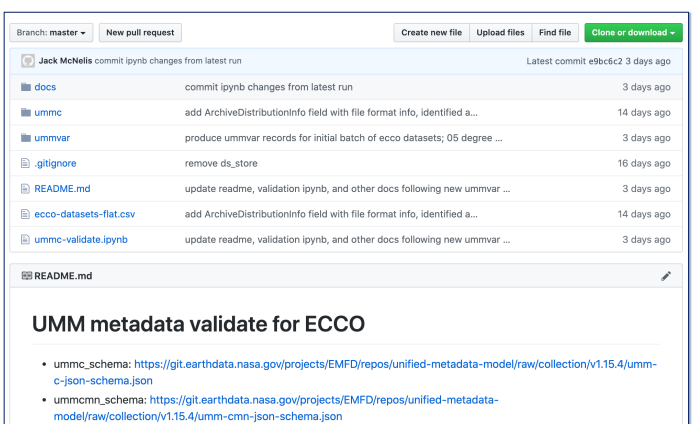
Abstract

Data stewardship depends on effective collaborations. At the **Physical Oceanography Distributed Active Archive Center (PO.DAAC)**, we collaborate with data providers to develop accurate and complete metadata about their datasets. We also collaborate with software developers who rely on consistent metadata to develop an effective user experience for our web portal and data services. We find that open source tools like Python, Jupyter, and Git are excellent for documenting data stewardship and workflows in a transparent, reproducible manner that fosters engaging collaboration with our colleagues. This poster discusses how PO.DAAC is leveraging these tools and environments for data stewardship using examples from our recent support of Earth science datasets.

In short, this poster cites examples of how the PO.DAAC is using open source tools for flexible data stewardship processes during the migration to the EOSDIS cloud.

Internal Collaboration

We develop flexible metadata maintenance procedures to fit the evolving requirements of cloud migration.

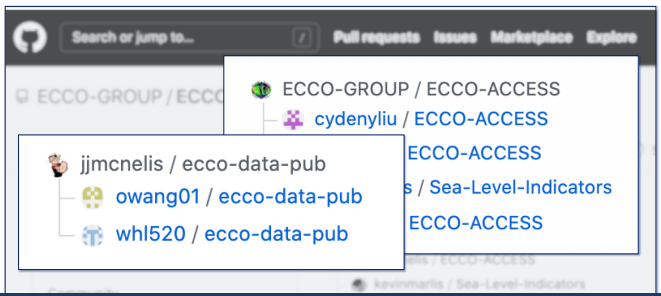


- Rapidly prototype & iterate metadata (Python)
- Version control each step of the process (Git)
- Produce reports about the status of metadata and the recent changes (Jupyter)
- Stage records in native format for use by other teams

We are iteratively revising metadata for Sentinel-6 MF and ECCO (see right) through this process.

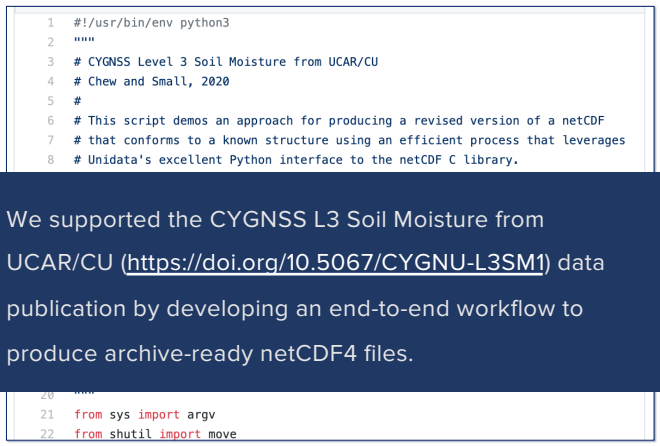
External Collaboration

We worked with data providers from Estimating the Circulation and Climate of the Ocean (ECCO) to develop & revise their data products to the ideal formats. ECCO version 4 release 4 will be released in early 2021.



- Developed routine to efficiently revise the netCDF attributes across the many diverse datasets
- Established & version controlled metadata with many stakeholders during the ingest process

We develop resources and one-off routines to streamline data transformations into optimal formats according to best practices and community standards like the Climate and Forecast (CF) Conventions and Attribute Convention for Data Discovery (ACDD).



We supported the CYGNSS L3 Soil Moisture from UCAR/CU (https://doi.org/10.5067/CYGNU-L3SM1) data publication by developing an end-to-end workflow to produce archive-ready netCDF4 files.

And we work with data providers to develop scripts and use case example for end users. Check our GitHub page for a recent example for GRACE: https://github.com/podaac

Acknowledgements

The PO.DAAC appreciates many collaborators throughout NASA, ESIP, and the wider earth science research community who share in the endeavor of data stewardship.

https://podaac.jpl.nasa.gov