



---

# The Next Generation Earth System Grid Federation



Jitendra (Jitu) Kumar, Oak Ridge National Laboratory

**ESIP Discovery Cluster**

*August 21, 2025*



U.S. DEPARTMENT OF  
**ENERGY**

Office of  
Science



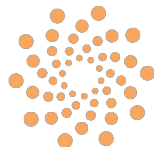
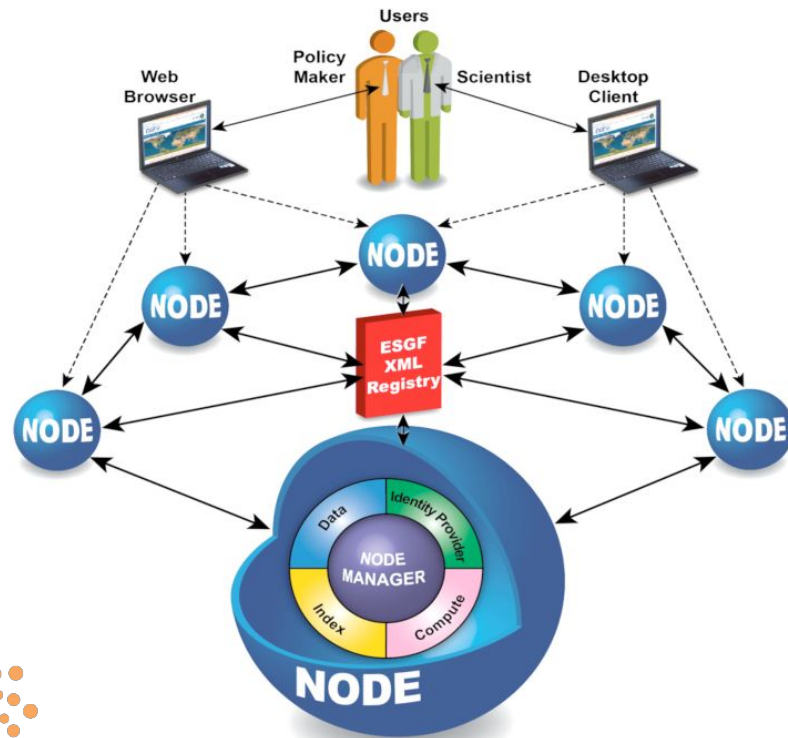


- Overview of ESGF
- Architecture of the next generation ESGF
  - Data index/search technology
  - Data search/discovery -- Web UI; Python library and CLI;
  - Data citations/DOI

# ESGF<sup>2</sup> US What is the Earth System Grid Federation?

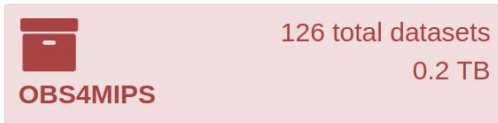
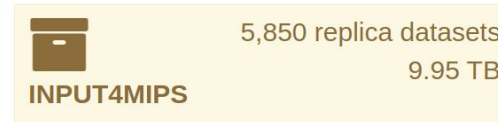
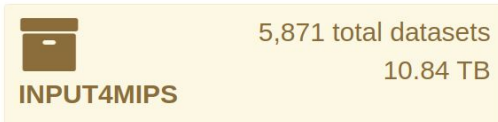
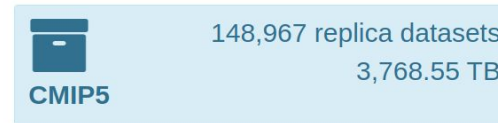
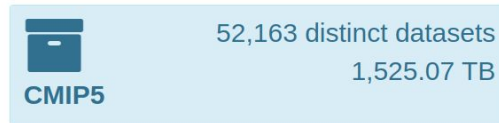
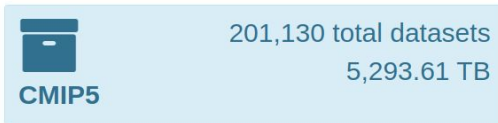
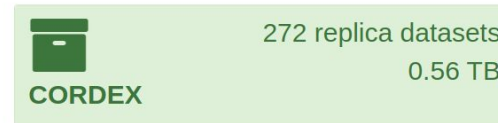
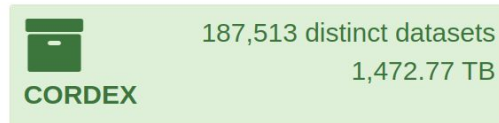
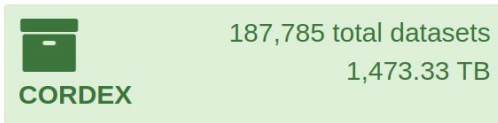
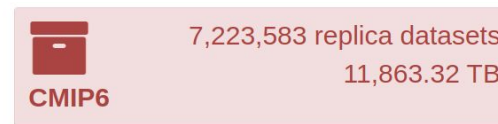
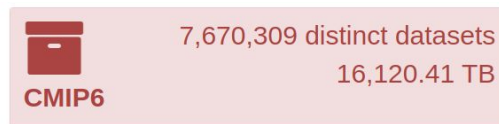
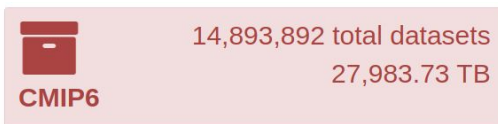
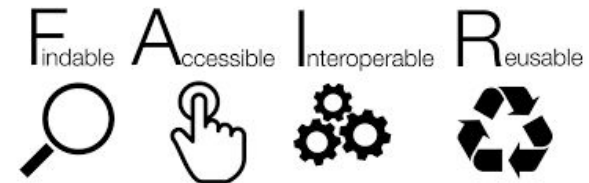
- **Earth System Grid Federation (ESGF)** is an *international consortium* and a *globally distributed peer-to-peer network of data servers* using a common set of protocols & interfaces to archive and distribute Earth system model output and related input, observational, and reanalysis data
- **Open Science data** are used by scientists all over the world to investigate Earth system variability and feedbacks and to inform research and assessments

## ESGF Conceptual Diagram



# ESGF US 2 ESGF Holdings are Open and Large

- CMIP5 totals >1.5 PB (>5 PB including replicas)
- CMIP6 totals >16.1 PB (>27 PB including replicas)
- CMIP7 is expected to have more experiments, high resolution output, and ensembles, totaling ~100 PB



As of August 21, 2025



# Planned ESGF-NG Architecture



**Data Access** via https & Globus Transfer (US)  
Zarr/aggregation data supported

via Kerchunk (limited nodes)  
**Direct Data Access** via User Computing (limited nodes)



**Identity & Access Management**  
Globus Auth (US)  
EGI Check-in (EU)

**Metagrid** –  
Data Discovery  
User Interface

**intake-ESGF** –  
Programmatic Search  
& Access

**esgpull** –  
Data Search & Bulk  
Data Transfer

Cloud-based Global Synchronized Catalogs (Can be expanded)



**West Catalog**  
STAC on  
Globus Search

DOE (US)

Data Search and Query  
Catalog Update & Synchronization

**East Catalog**  
STAC on  
ElasticSearch



CEDA (UK)

**Publisher** –  
Data Publication,  
Retraction, Replication

**Message Queue**



**Services**  
Data Movement, Value-Added Products, Rapid Evaluation Framework, Community Projects

High Capacity Storage



**Tier 1 Data Node**

User Computing / JupyterHub  
Server-side Computing / WPS & Globus Compute



**Tier 1 Data Node**



**Tier 1 Data Node**



**Tier 2 Data Node**



# Planned ESGF-NG Architecture



**Data Access** via https & Globus Transfer (US)  
Zarr/aggregation data supported

via Kerchunk (limited nodes)  
**Direct Data Access** via User Computing (limited nodes)



**Identity & Access Management**  
Globus Auth (US)  
EGI Check-in (EU)

**Metagrid -**  
Data Discovery  
User Interface

**intake-ESGF -**  
Programmatic Search  
& Access

**esgpull -**  
Data Search & Bulk  
Data Transfer

Cloud-based Global Synchronized Catalogs (Can be expanded)

**West Catalog**  
STAC on Globus Search  
DOE (US)

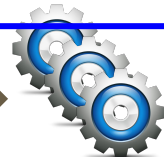
Data Search and Query

**East Catalog**  
STAC on ElasticSearch  
CEDA (UK)

Catalog Update & Synchronization

**Publisher -**  
Data Publication,  
Retraction, Replication

**Message Queue**



**Services**  
Data Movement, Value-Added  
Products, Rapid Evaluation  
Framework, Community Projects

High Capacity Storage

User Computing / JupyterHub

Server-side Computing / WPS & Globus Compute

High Capacity Storage

User Computing / JupyterHub

High Capacity Storage

Server-side Computing / WPS & Globus Compute

High Capacity Storage

Tier 1 Data Node

Tier 1 Data Node

Tier 1 Data Node

Tier 2 Data Node



# ESGF Data Catalog/Index

- In past, ESGF has used Apache Solr for storing and searching metadata for all datasets.
- Apache Solr Index nodes were operated and maintained by individual ESGF nodes [5+2 Solr nodes currently in operation].
- These index nodes were federated and Web UI conducted distributed search across these federated index nodes.



# ESGF Data Catalog/Index

- In past, ESGF has used Apache Solr for storing and searching metadata for all datasets.
- Apache Solr Index nodes were operated and maintained by individual ESGF nodes.
- These index nodes were federated and Web UI conducted distributed search across these federated index nodes.
- **Challenges/Limitations:**
  - However, maintaining these Solr indices have been difficult and resource consuming
  - Security vulnerabilities
  - Resulted in varying search experiences depending on whether nodes were online/offline



# ESGF Data Catalog/Index

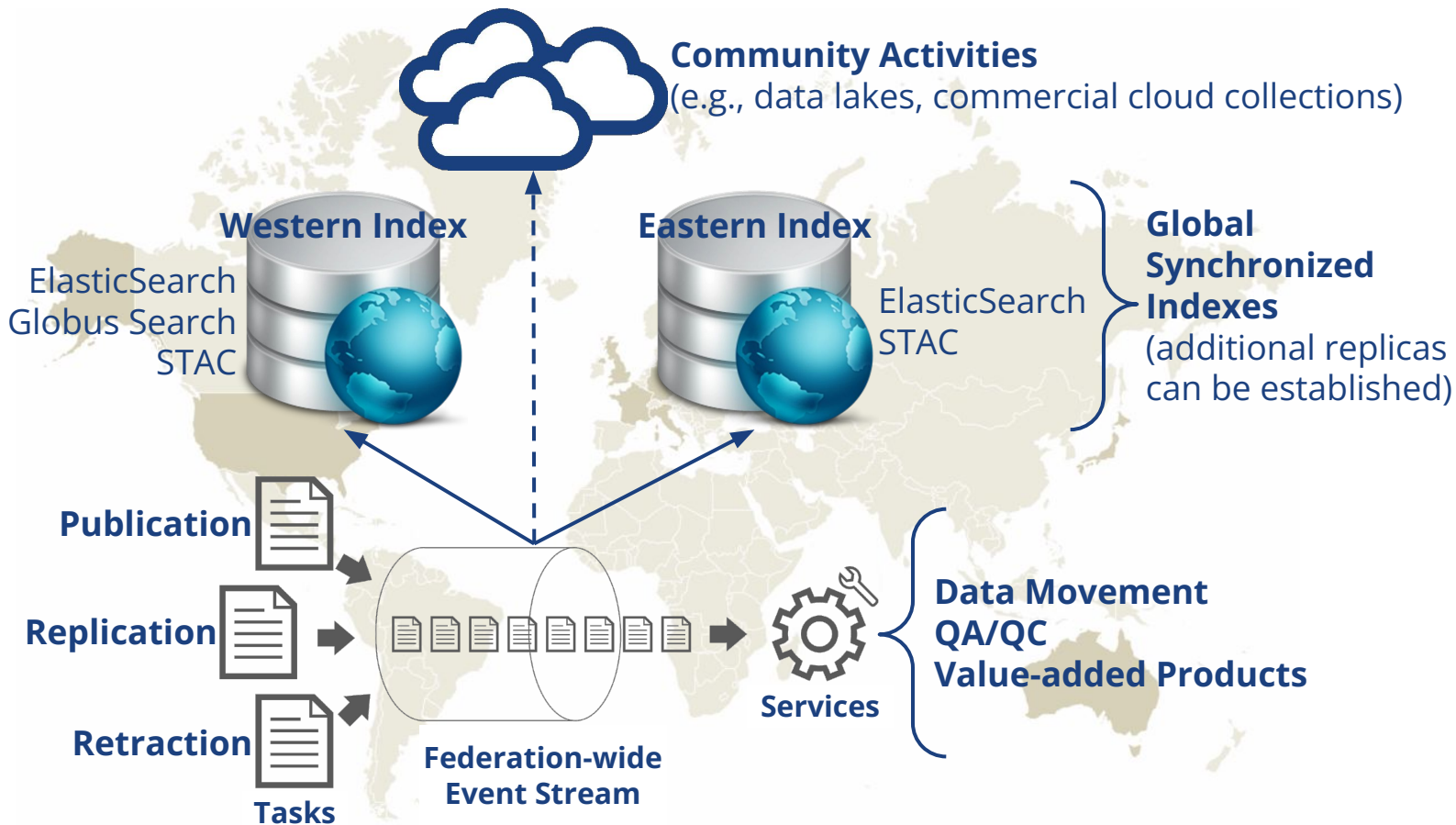
- The next generation of ESGF, currently in development, will use Spatio Temporal Asset Catalog (STAC) for storing and searching ESGF metadata
  - Benefits from range of community developed tools to interact with STAC
  - Enables integration with earth observation data from other providers
- Two cloud based index nodes to enable a seamless and consistent search experience for the users
- Solr nodes will be retired and technology no longer supported

The screenshot shows the STAC catalog interface for CMIP6 data. The main heading is "CMIP6" in CEDA STAC API. Below this is a description of the WCRP Coupled Model Intercomparison Project, Phase 6 (CMIP6), and information about the archive's management and access. A "Description" section provides details about the project's coordination by PCMDI and its role in providing input for the IPCC 6th Assessment Report. A "License" section indicates "other" and a "Temporal Extent" section shows "1850-01-01 00:00:00 UTC - 4114-12-16 12:00:00 UTC". A map view shows a world map with a red bounding box over the North Atlantic region. The "Asset" section at the bottom shows a thumbnail of the selected asset. On the right, an "Items" section displays a list of search results, each with a title, ID, and temporal extent. The results include items like "CMIP6.CMIP.EC-Earth-Consortium.EC-Earth3-Veg.piControl.r11p1f1f1.Amon.clivi.gv.v20210419" and "CMIP6.ScenarioMIP.CSIRO.ACCESS-ESM1-5.ssp126.r11p1f1f1.day.us.gn.v20210318".

STAC catalog in operation @CEDA



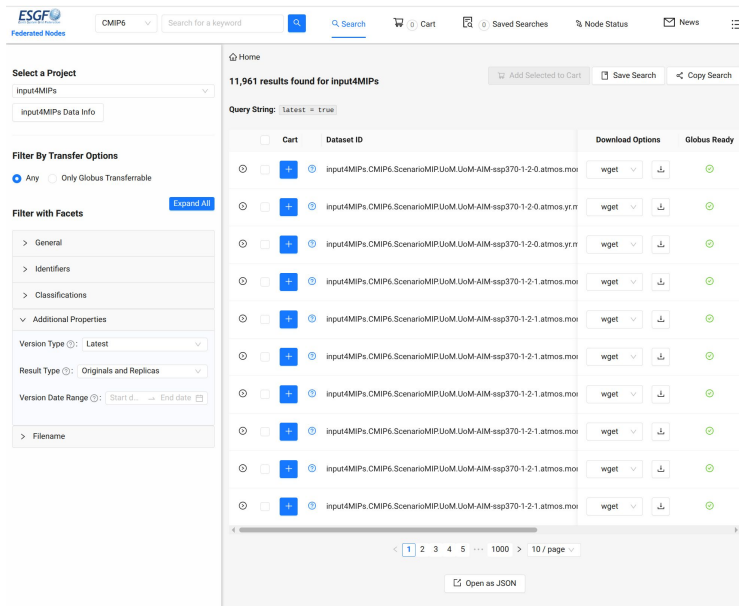
# New Redundant Index Strategy





# Metagrid Enhances ESGF Search

- New **Metagrid faceted search user interface**, developed at LLNL on popular React Javascript framework, deployed at ORNL, LLNL and ANL
- Offers new features, including a **shopping cart**, ability to **save and share searches**, integration with **Globus authentication & transfer** and a search page **tour & support dialog**
- User experience enhancements make it **faster and easier** to discover published data
- **Globus integration** offers faster and more reliable data access



The Metagrid Web Interface for ESGF search is a completely redesigned interface from CoG. It features a familiar faceted search and a new capability to save searches.





# Integrating with intake-esgf

- Improve the APIs to access data; simplify searching for data programmatically across the federation
- Provide STAC-based index query in addition to the existing Solr and Globus indices [backwards compatibility]
- Extend the interface to provide capability for data streaming (OPeN-DAP, Kerchunk, Virtual Zarr) as available
- Integrate the errata service provided by es-doc into intake-esgf catalogs

The screenshot shows the documentation page for 'intake-esgf'. The page has a light blue header with the 'ESGF US 2' logo. Below the logo is a search bar and a navigation menu with sections: USER GUIDE (Beginner's Guide to ESGF, Quickstart), FEATURES (Automatic Cell Measures, Simplifying Search with Model Groups, Configuring the ESGFCatalog, Reproducibility, Output Dictionary Key Format, Logging), and EXPERIMENTAL (Globus Transfers). The main content area has the 'ESGF US 2' logo at the top right, followed by the title 'Documentation for intake-esgf' and a brief description: 'intake-esgf is an intake and intake-esm inspired package under development in ESGF2. The data catalog is populated by querying a number of index nodes and puts together a global view of where the datasets may be found. If you are familiar with the interface for intake-esm, then using this package should be straightforward.' Below this is the 'Installing' section, which states 'intake-esgf can be installed using pip:' and provides the command 'pip install intake-esgf'. It also says 'or through conda-forge' and provides the command 'conda install -c conda-forge intake-esgf'. The version 'v: latest' is shown at the bottom left of the page.

- Intelligently determines the quickest way to access data (download, Globus Transfer, stream, load locally)
- **Provides method to package compute + flows**

Repository: <https://github.com/esgf2-us/intake-esgf>

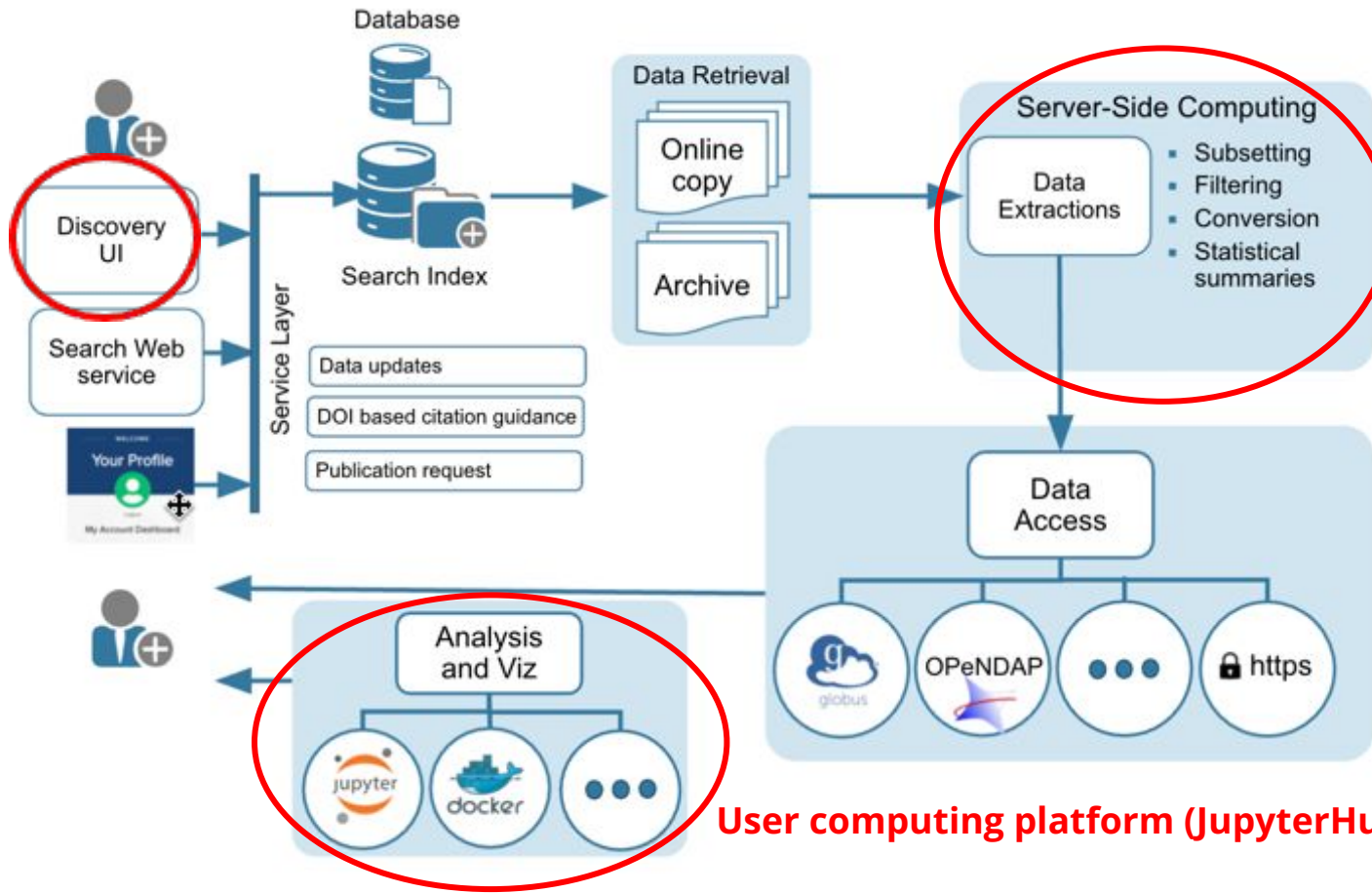
Documentation: <https://intake-esgf.readthedocs.io/>

Installation: PyPI and Conda-forge



# Data Discovery, Access & Analysis Platform

**Friendly user interface**



**Server-side computing platform (Web Processing Service)**

**User computing platform (JupyterHub)**



# Proposed DOI Workflow

## DOI Pre-Registration System

Modeling Center registers for a DOI for every [Institution + Model + Experiment] entity on web interface (shares publishing auth)

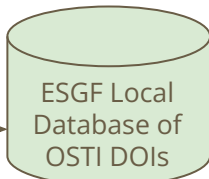
## Data Publication Process

Modeling Center rewrites model output, includes DOI in netCDF file (doi global attribute?), and runs QA/QC checks

Modeling Center runs Publisher workflow to update regional STAC catalog

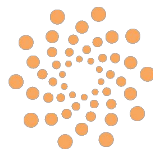
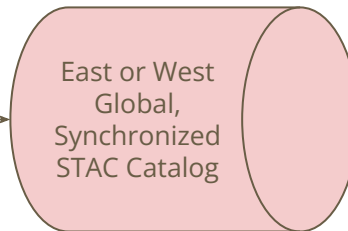
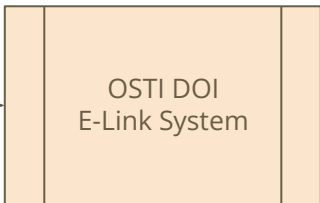
- Modeling center pre-registers for DOI through web interface
- They encode the DOI within every netCDF file
- The Publisher extracts and validates the DOI, and updates the local database, landing page, and OSTI registration
- DOI is included in STAC catalog along with metadata

System uses OSTI's API to mint DOI, saves registration to ESGF Local Database, and provides it to web interface user



Publisher validates registered DOI from ESGF Local Database, updates ESGF Local Database, creates/updates landing page, updates OSTI records, and writes record (including DOI) to the regional STAC catalog

[Institution + Model + Experiment] Land Page





- Data Granularity:

- DOIs will be issued for unique combination of *mip\_era* + *institution\_id* + *source\_id* + *experiment\_id*

- Issued DOIs will follow the format:

10.25981/ESGF.{**mip\_era**}.{**institution\_id**}.{**source\_id**}.{**experiment\_id**}/XXXXXX

10.25981/ESGF.**CMIP7.MIROC.MIROC-ES2L.ssp245**/99999999

- DOIs will be included in the global header of the NetCDF files

- ESGF-Publisher will perform QA/QC before data publication and DOI release

- DOI registration:

- We are leveraging DOE OSTI ELINK service to register data DOIs with DataCite

- ESGF DOI Service will auto-generate landing pages for DOIs

- Follow landing page best practices from DataCite

(<https://support.datacite.org/docs/landing-pages>)



Thank you for your attention!

Comments/feedback/suggestions are welcome on our ongoing developments.

Email: [kumarj@ornl.gov](mailto:kumarj@ornl.gov)