Removing Barriers to Data Access (Publishing Non-Conforming Data Products into ESGF)

2013 ESGF F2F

Eric Nienhouse, Douglas Schuster, Kevin Manross, Don Middleton, Steve Worley (NSF/NCAR)
Quick Outline

Expanding access to climate and related data products

• Introduction to operational systems and user communities.
• Use metrics and projections.
• Requirements for integrating operational systems with ESGF.
• Status of NCAR ESGF node evaluation.
• Suggestions for ESGF
NSF NCAR Data Archives

We build data archives and curate data for diverse communities.

20K annual users, 300 data providers, 10K collections, 3.5PB, 2PB yearly downloads.

ACADIS
Advanced Collaborative Arctic Data Information Service

ESG-NCAR
Earth System Grid at NCAR

RDA
Research Data Archive

Reanalysis + obs products
Subset and re-format svcs
ECMWF, ICOADS, JRA-55
Actively curated

NSF Arctic projects
Self publishing tools
Many disciplines
Highly varied data
Long term preservation

Climate models (CESM)
CMIP5 products
Large data volume
Heavily accessed
Community Use and Access

Data products are growing in popularity

- Over 2000 users monthly
- Diverse and growing user base
- OAI services for search aggregators
- 10X download volume by 2017
- Data reduction increasingly utilized
- We want more ways to access data
Removing Barriers to Discovery and Access

Problem
• Data discovery is tedious and requires many archive visits.
• Scientifically related products are hard to retrieve and use.

Solution
• Search and discovery aggregators facilitate scientific research. (Provided the metadata is rich enough)
• Standards based data services make retrieval and use easier. (Provided security complexity doesn’t get in the way)

ESGF and Ana4MIPS conventions is viable option.
Key Requirements

Integrate ESGF and maintain continuity of operational systems

- Accurate use metrics which are easy to aggregate with existing systems.
- Support for access control and product license agreements.
- Minimal effort publication of existing non-conforming data products.
- Customizable UI for facets and navigation.
- HPSS based data services.
- Low friction node upgrades.
- Remote data node publication service (ORNL, NERSC, ANL)
- Staging system for testing and release promotion.
A Brief Word on Metrics...

Metrics enable cost minimization and efficient use of resources.

• Funders require metrics and are a “must do”.
• Metrics are critical to the “Build, Measure, Learn” loop.
• Adding measurement to ESGF can yield faster, better development.
Key Requirements (Again)

Integrate ESGF and maintain continuity of operational systems

1. Accurate Metrics
2. Easily Publish Existing Catalogs
3. License Agreements and Access Control
4. HPSS Data Access
5. Test Before Production
6. Extendable UI
ESGF at NCAR

Evaluate ESGF for sharing diverse collections from NCAR archives

General Goals

- Assess costs of operating an ESGF Node at NCAR for sharing data collections.
- Publish high value RDA managed reanalysis data collections into ESGF system.
- Maintain continuity of data access to ESG-NCAR and RDA data products.

Status

- Team assembled from ESG-NCAR, RDA and Enterprise Operations.
- Targeting ESGF Version 1.6 (Metrics, Publishing API)
- Virtual machine systems prepared (CentOS)
- Candidate datasets proposed
Deploy In a Phased Approach

Start simple and learn, providing high value first.

- Publish current CMIP5 holdings to dedicated v 1.6 data node.
- Use REST Publishing API to publish existing (or enhanced) catalogs.
- Augment UI with additional facets and navigation (CoG?)
- Simplify auth-n and auth-z integration with existing systems.
- Expose additional data access services (OpenDAP, NcSS, NcISO)
Customer Suggestions

An ESGF customer perspective

- Visible and practical 1 year+ product roadmap
- Establish a Governing body with an operational focus
- Organize & prioritize service API development
- Release more frequently
- Reduce security configuration surprises
- Support certificate-less scripted download
- Pursue identity bridging and IdP based Attribute Service
- Consider opt-in notifications
- Use well known CAs for SSL certificates
- Create a test federation to support installation
Thank you!

Questions?

ejn@ucar.edu
schuster@ucar.edu
Questions and Challenges

Where does the ESGF platform end and customization begin?

Tape?
- Tape access is unsupported, is this a core capability of the ESGF?

Search Facets?
- Policy for sharing custom of project based search facets?

Site Customization?
- Ease of upgrading customized node components (view/UI)
Extra Stuff
CMIP5 Data at ESG-NCAR
As a national center NSF-NCAR manages over 5PB of climate and related data products. Integration of these products with distribution systems such as ESGF removes barriers to scientific data discovery and access. In this talk we will present key requirements for including these valuable data products in ESGF:

- Publication of existing data product catalogs
- Access control and licensing workflows
- Aggregation of use metrics from distributed systems
- Remote publication of geographically distributed products
- Integration with legacy and tape storage systems