ESS-DIVE Webinar: Model Data Archiving

Maegen Simmonds, William Riley, Charuleka Varadharajan
November 22, 2019
Webinar Goals

1. Review & discuss project objectives & proposed plan
2. Background on other model data repositories and approaches
3. Discuss & improve the feedback form, and discuss answers
ESS-DIVE: Current Functionality for Model Data Archiving

CREATE SINGLE OR MULTIPLE DATA PACKAGES
GOALS & PROPOSED PLAN

ESS-DIVE: Current Functionality for Model Data Archiving

First model dataset on ESS-DIVE
ESS-DIVE: Current Functionality for Model Data Archiving

Opportunities for improvement!
Problem Statement

Model data storage is limited
• currently only archiving a limited set of small-sized model outputs
• infrastructure limitations on data size:
  • upload limits 2GB/file on portal and 1GB/file on API
  • architecture limits how much data ESS-DIVE can store and serve
• web interface not the best tool for uploading/downloading large datasets
• API helps but there are still physical limitations

No community consensus yet on what to archive, standards, storage space needed, etc.
Main objectives of this project

Assess:
• What model data should be archived, purposes of storage, storage capacity needed
• Best approach to store data

Deliverable:
• White paper describing data storage capabilities ESS modeling projects need based on community feedback and a few use-cases, and potential storage options
Timeline

Feb 2019
CI meeting on model data

Nov 2019
ESS-DIVE webinar kick-off of model data white paper project

Dec-Apr 2020
Deeper dive-one-on-one conversations

Feb/Mar 2020
ESS-DIVE Virtual Model Data Workshop

May 2020
Finalize white paper

GOALS & PROPOSED PLAN

Model Data Archiving Feedback Form
Examples of cloud-based web storage for model & observational data

- Earth System Grid Federation (ESGF)
- NASA's Earth Observing System Data and Information System (EODIS)
- NCAR's Earth Observatory Laboratory (EOL) Data Archive
- NCAR's Research Data Archive (RDA)
- CYVERSE
ESGF

- Federated system for storing and serving data from multiple locations and sources
- Developed for sharing climate model (CMIP) data
- Currently stores some ESS modeling data
- NOT guaranteed to be a long-term host of data, to follow digital library standards, or to assign DOIs to data
EOSDIS

Distributed Active Archive Centers (DAACs)

Alaska Satellite Facility DAAC
- SAR Products, Sea Ice, Polar Processes, Geophysics

Land Processes DAAC
- Land Cover, Surface Reflectance, Radiative Transfer, Vegetation Indices

Goddard Earth Sciences Data and Information Services Center
- Global Precipitation, Solar Irradiance, Atmospheric Composition, and Dynamics, Global Modeling

Socioeconomic Data and Applications Center
- Human Interactions, Land Use, Environmental Sustainability, Geospatial Data

Physical Oceanography DAAC
- Gravity, Sea Surface Temperature, Ocean Winds, Topography, Circulation & Currents

National Snow and Ice Data Center DAAC
- Frozen Ground, Glaciers, Ice Sheets, Sea Ice, Snow, Sea Moisture

Crustal Dynamics Data Information System
- Space Geodesy, Solid Earth

Oak Ridge National Laboratory DAAC
- Biogeochemical Dynamics, Ecosystem Data, Environmental Processes

Ocean Biology DAAC
- Ocean Biology, Sea Surface Temperature

Global Hydrology Resource Center DAAC
- Hazardous Weather, Lightning, Tropical Cyclones and Storm-induced Hazards

Level 1 and Atmosphere Archive and Distribution System (LAADS) DAAC
- MODIS/LAND, Atmosphere Data Products

Atmospheric Science Data Center
- Radiation Budget, Clouds, Aerosols, Thermospheric Chemistry

EOSDIS Core System (ECS) Sites

FIGURE 1. HISTORICAL AND PROJECTED CUMULATIVE ARCHIVE VOLUME IN EOSDIS. (YEARS RUN FROM OCTOBER TO SEPTEMBER.)
Earth Observatory Lab (EOL) Data Archive

EOL Metadata Database and Cyberinfrastructure (EMDAC)

Web access to archive
- ZINC (new CODAC-style UI)
- Mass Store (MSS) retrieval
- Master Lists
- Community Data Portal

EOL Metadata Database
- Metadata catalog export
  - THREDs
  - ISO19115
  - NASA GCMD
  - RSS feeds
  - Other/XML

Internal tools
- Metrics
- Web Services
- Dataset Tracking
- Maintenance

Browsing and visualization
- Field Catalog
- Mapserver
- ILV
- VOC
- Other visualization tools

Other Services

FIND DATA
- Projects
- Categories
- Platforms
- Instruments
- GCMD
  - 12 HOUR PRECIPITATION AMOUNT (2)
  - 24 HOUR MAXIMUM TEMPERATURE (13)
  - 24 HOUR MINIMUM TEMPERATURE (13)
  - 24 HOUR PRECIPITATION AMOUNT (7)
  - 3 AND 6 HOUR PRECIPITATION AMOUNT (7)
  - 6 HOUR MAXIMUM TEMPERATURE (1)
  - 6 HOUR MINIMUM TEMPERATURE (1)
  - ABSOLUTE HUMIDITY (10)
  - ACCUMULATIVE CONVECTIVE PRECIPITATION (1)
  - ACOUSTIC VELOCITY (2)

EOL data archive

The EOL data archive contains atmospheric, meteorological, and other geophysical datasets from operational sources and the scientific research programs and projects for which NCAR/EOL has provided data management support. The project list may be sorted by selecting the header keys and full project descriptions and dataset lists are available by selecting the project title. You may search for projects with the form below. You may also search for datasets by keyword or space and time.

Projects

502 projects (100 shown)

Name: Title / Summary
MethaneAIR: MethaneAIR
MethaneAIR is an airborne imaging spectrometer funded by the Environmental Defense Fund (EDF) and private philanthropy as part of the EDF International Methane project. The first scientific goal of the EDF program is to accurately measure emissions of...
Research Data Archive (RDA)

Browse the RDA

There are 696 public datasets in the CISL RDA. You can begin browsing the datasets by choosing one of the facets in the menu to the left. Facet descriptions are given below, along with the number (in parentheses) of datasets in each.

Variable / Parameter (996)
A variable or parameter is the quantity that is measured, derived, or computed – e.g. the data value.

Type of Data (996)
This refers to the type of data values – e.g. grid (interpolated or computed gridpoint data), platform observation (in-situ and remotely sensed measurements), etc.

Time Resolution (928)
This refers to the distance in time between discrete observation measurements, model product valid times, etc.

Platform (968)
The platform is the entity or type of entity that acquired or computed the data (e.g. aircraft, land station, reanalysis model).

Spatial Resolution (341)
This refers to the horizontal distance between discrete gridpoints of a model product, reporting stations in a network, measurements of a moving platform, etc.

Topic / Subtopic (998)
Topic and subtopic are high-level groupings of parameters – e.g. Atmosphere (topic), Clouds (subtopic of Atmosphere).

Project / Experiment (998)
This is the scientific project, field campaign, or experiment that acquired the data.

Supports Project (99)
This refers to data that were acquired to support a scientific project or experiment (e.g. GATE) or that can be used as ingest for a project (e.g. WRF).

Location (998)
This the name of the (usually geographic) location or region for which the data are valid.

Data Format (998)
This refers to the structure of the bitstream used to encapsulate the data values in a record or file – e.g. ASCII, netCDF, etc.

Faceted Dataset Search:

- **All Datasets**
- **Variable/Parameter**
- **Type of Data**
- **Time Resolution**
- **Spatial Resolution**
- **Topic/Subtopic**
- **Project/Experiment**
- **Supports Project**
- **Location**
- **Data Format**
- **Common Data Formats**
  - WMO BUFR
  - WMO GRIB
  - NetCDF

Get Help:
- Frequently Asked Questions
- Reset your password
- A-Z Site Index
- RDA Blog
- RDA video tutorials
- Email Us

From Our Blog:
- ERAS 0.75 degree monthly mean data available
- ERCR Interim updated through August 2019
- Changes to NCAR RDA
- seasoned ERAD
- More blog posts..
CYVERSE (originally iPlant)

- Data storage geared specifically towards data analysis
- Interactive, web-based analytical platform
- Cloud computing, analysis and storage
- Support services for scaling up computational algorithms & on how to use CI
Any thoughts so far?
(DRAFT) Feedback form for developing the ESS-DIVE model data repository

• Currently comprised of 20 questions to:
  – inventory models and assess their specific data storage needs
  – evaluate what's worth archiving and for how long
  – get recommendations for archiving protocols and storage options

• Let's review for completeness and start discussing answers!
ESS-DIVE: Vision

Diagram showing the interaction between users, data packages, databases, and tools within the ESS-DIVE ecosystem.
Next steps

1. Synthesize our discussion today: compile preliminary poll of responses to feedback form, and email link for additional comments on it from the ESS community.

2. Revise feedback form and distribute to everyone in ESS community.

3. Connect for follow-up discussions.

4. Email any questions or more ideas to Maegen (mbsimmonds@lbl.gov) and Bill (wjriley@lbl.gov).