Automated Data Submission: From the Scientist to the Archive

Daine M. Wright (wrightdm@ornl.gov), Tammy Beaty, Robert Cook, Ranjeet Devarakonda, Pete Eby,
Les Hook, Ben McMurry, Harold Shanafield, Dave Sill, Suresh K. S. Vannan, Yaxing Wei
Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC, http://daac.ornl.gov)

Introduction

The ORNL DAAC archives and publishes data and information relevant to biogeochemical, ecological, and environmental processes. The ORNL DAAC primarily archives data produced by NASA's Terrestrial Ecology Program; however, any data that are pertinent to the biogeochemical and ecological community are of interest.

The data set submission process at the ORNL DAAC has been recently updated and semi-automated to provide a consistent data provider experience and to create a uniform data product. The data archived at the ORNL DAAC must be well formatted, self-descriptive, and documented, as well as referenced in a peer-reviewed publication. If the ORNL DAAC is the appropriate archive for a data set, the data provider will be sent an email with several URL links to guide them through the submission process.

Motivation

Formalized work flow makes data set submission easier for data providers and faster for the ORNL DAAC staff.

Benefits for the data provider

- Data provider starts metadata record with answers to data provider questions reducing incorrect interpretations
  - Form should only take about 20 minutes to complete
- Data upload standardized, straight forward, and fast
- Confirmation email is receipt of data submission

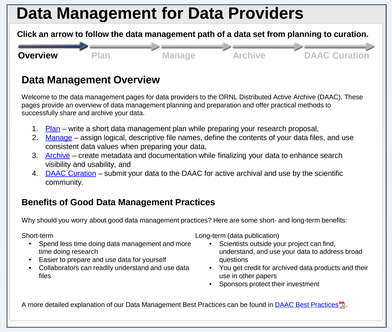
Benefits for the archive

- Data uploaded to single secure upload area instead of several points of entry
- Pending data sets for archival recorded in single place
  - Status of submissions in one table
  - Reporting of data set submission status becomes simple
- Data products and metadata are uniform and maintainable

Find out more

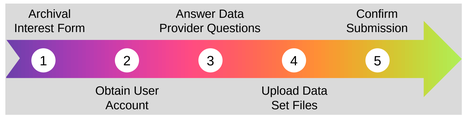
Visit http://daac.ornl.gov

- if you are interested in archiving your data at the ORNL DAAC
- if you are interested in our data set submission work flow
- to find our best practices for data management

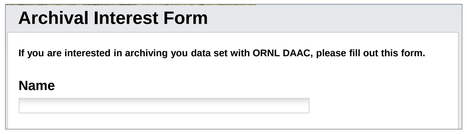


From the Scientist …

Quick and easy data submission allows data providers to archive data for a finished project and move on or maintain a data repository for an ongoing project.

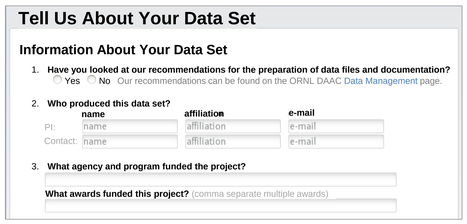


1. Archival interest form initiates submission request



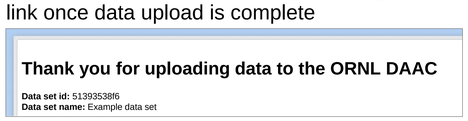
2. Obtain user account

3. Answer data provider questions to provide preliminary metadata for data set
   - Easy to answer
     - Answers should be readily available
     - Should take about 20 minutes
   - Answers are the basis for metadata record



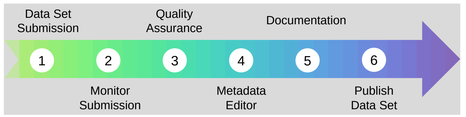
4. Upload data set files to secure data upload area
   - Multiple users can upload data

5. Confirm data set submission by following provided link once data upload is complete

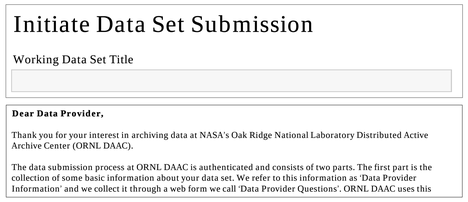


… to the Archive

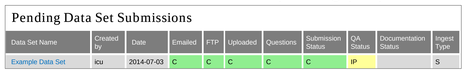
Formalized work flow makes data set submission faster for the ORNL DAAC staff and provides uniform data products and metadata.



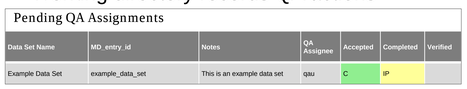
1. Data set submitted by data provider
   - Initial email to data provider includes links to user registration, data providers questions, secure upload area, and data upload confirmation



2. Monitor submissions to quickly assess status
   - For data set submission reporting purposes



3. Quality Assurance assigned to staff member/team
   - Raw data stored in read-only directory
   - Working directory records QA actions



4. Metadata collected for data set and granules
   - Answers from data provider questions form basis of data set level metadata
   - Additional granule level metadata collected
   - Automation development in progress

5. Documentation
   - Data set documentation html and xml generated directly from metadata data base
   - Automation development in progress

6. Publish data set
   - Data becomes available for download at http://daac.ornl.gov

Archival functions

There are five major functions the ORNL DAAC performs during the process of archiving data:

1. **Evaluation**
   - Check to ensure that data files and documentation agree and that data values are reasonable
   - Generate additional metadata to complete the metadata record, if needed
   - Convert any proprietary data files to archival formats (e.g., .csv, GeoTIFF, KML)
2. **Metadata and documentation**
   - Prepare searchable metadata record and documentation
   - Include a permanent URL to data set in the documentation
3. **Publication**
   - Generate data set citation and DOI (digital object identifier)
   - Publish data product (data files, metadata, documentation) and distribute metadata to NASA
   - Advertise data through email, DAAC website, and related metadata repositories
   - Provide tools to explore, access, and extract data (e.g., Mercury, SDAT, MODIS Subsets, THREDDS)
4. **Post-Project Data Support**
   - Provide long-term, secure archival
   - Address user questions, and serve as a buffer between data users and data contributors
   - Provide usage statistics and data citation statistics
5. **Stewardship**
   - Provide security, backups, and disaster recovery
   - Migrate data products to new computer systems (servers and file storage) every ~3 years

How to submit your data set

If you are interested in archiving your data set with the ORNL DAAC, email our User's Services Office: uso@daac.ornl.gov.

Consider best practices for data management (see lower left panel) while preparing your data set for archival.

Future work

The entire archival process will be semi-automated and formalized into a standardized work flow. Efforts are underway to automate the work flow for:

- Metadata editor
- Data set documentation

Acknowledgements

This effort was supported by the NASA's Earth Observing System Data and Information System under grant number NNG09HP121.

ORNL is managed by the University of Tennessee-Battelle LLC under contract DE-AC05-00OR22725 with the U.S. Department of Energy.