
# Data Management Plan by the Physical Oceanography Distributed Active Archive Center





This document was created at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018F0848)

JPL document clearance: CL#20-1986

# Change Log

| Revision | Date | Description | Author |
|---|---|---|---|
| Version 0 | Feb 10, 2020 | Outline of DMP and population in google document | Ed Armstrong |
| 0.1 | Mar 18, 2020 | Revision and edits | Ed Armstrong<br>Jessica Hausman |
| 1.0 | Apr 21, 2020 | Revisions, edits, table of contents | Ed Armstrong |
| 1.1 | Apr 28, 2020 | Further minor edits after review by PO.DAAC management team. Export to MS Word/PDF for Docushare archive (https://bravo-lib.jpl.nasa.gov/docushare/dsweb/View/Collection-28250) | Ed Armstrong |
| 1.1 | May 6, 2020 | JPL document clearance number assigned (CL#20-1986) | Ed Armstrong |

# Executive Summary

The NASA Physical Oceanography Distributed Active Archive Center (PO.DAAC) is one of twelve DAACs that are part of the NASA Earth Observing System Data and Information System (EOSDIS) project. The PO.DAAC, located at the NASA Jet Propulsion Laboratory in Pasadena, California is responsible for the distribution, archiving, and overall data management of Earth science data, specifically physical oceanographic and other related earth science observations from satellite, airborne, and in situ platforms, including model output. This Data Management Plan (DMP) provides an overview of data management provided by the PO.DAAC representing the spectrum of services and capabilities available to the dataset provider, and the Earth science research and applications community of data users.

# Table of Contents

# 1. Introduction

The NASA Physical Oceanography Distributed Active Archive Center (PO.DAAC) is one of twelve DAACs that are part of the NASA Earth Observing System Data and Information System (EOSDIS) project. The PO.DAAC, located at the NASA Jet Propulsion Laboratory in Pasadena, California is responsible for the distribution, archiving, and overall data management of Earth science data, specifically physical oceanographic and related earth science observations from satellite, airborne, and in situ platforms, including model output. The mission of the PO.DAAC is to preserve NASA's ocean and climate data and make these universally accessible and meaningful.

## 1.1 Purpose and Scope

The purpose of this DMP is to describe the PO.DAAC's capabilities and plans for archiving, distribution and provision of user services for Earth science datasets assigned to the PO.DAAC for Earth science related measurements.

## 1.2 DMP Development, Maintenance and Management Responsibility

The PO.DAAC Data Engineering (PDE) group is responsible for the development, maintenance, and management of the DMP. The PO.DAAC Manager, has overall responsibility for implementation of the plan, and has specific responsibility for final approval of any changes to the plan. All changes to the DMP will be controlled and documented in the change log.

# 2. Data Accession and Deaccession Overview

The majority of datasets are assigned to the PO.DAAC by NASA HQ through instructions to the ESDIS Project. These datasets may be generated by individual instrument science teams from satellite missions or aircraft investigations, field experiments, or Principal Investigator (PI) led projects (e.g.. MEaSUREs). Such assignments of datasets are documented through the ESDIS Project's dataset accession process (Appendix A) and the Planning, Programming, Budgeting, and Execution (PPBE) process. Some datasets are submitted by data producers with requests for archiving at the PO.DAAC. Such datasets are accepted for archiving using an accession process, essentially a written petition process by the PO.DAAC to ESDIS management at the

request of the PI. When a dataset is to be deaccessioned or purged, the DAAC will follow procedures outlined in CEOS Data Purge Alert Procedure [1], which provide for orderly removal of datasets from archives giving opportunities for other archives to take on the responsibility for continued archival if deemed appropriate. (Retirement of older versions of datasets that are superseded by newer versions at the direction of the PI/science teams is not considered deaccession).

## 3. Data Sources Overview

PO.DAAC, including its prior historical incarnations as NODS and PODS, has been a data management organization active in supporting oceanographic science since the launch of NASA's first ocean-observing satellite, Seasat in 1978. Since then PO.DAAC has evolved to become a premier data center for measurements focused on ocean surface topography (OST), sea surface temperature (SST), ocean winds, sea surface salinity (SSS), gravity, ocean circulation, sea ice, and hydrology. Most of PO.DAAC's datasets are satellite based although its catalog is increasingly diversified with additions of *in situ* and airborne observations, and model data.

A list of NASA and partnership missions can be found at the following site (https://podaac.jpl.nasa.gov/missions) and are listed individually below:

- ADEOS-II
- AQUA
- AQUARIUS
- CYGNSS
- GEOS-3
- GHRSST
- GRACE
- GRACE-FO
- ISS-RapidScat
- JASON 1
- JASON 3
- MEaSUREs
- NSCAT
- OMG
- OSTM-JASON 2
- QuikSCAT
- Saildrone
- Seasat
- SMAP
- S-NPP
- SPURS
- SWOT
- TERRA
- TOPEX-POSEIDON

Details of the missions including datasets, processing levels, and documentation can be found by visiting each mission page. Datasets associated with a mission have a unique "landing or information page" that exposes the dataset metadata and documentation, tools and services support, and citation references.

# 4. Interfaces

PO.DAAC maintains internet-based linkages to data providers via push/pull mechanisms. Data flows are either pushed by a data provider to a PO.DAAC endpoint via sftp with provider specific login credentials or pulled from a provider repository endpoint via a PO.DAAC https/ftp crawler. In each case data files are ingested into the PO.DAAC systems via dataset specific “data handlers” that check file integrity after transmission, ingest file metadata into an internal database, and deliver the file to the PO.DAAC data store for public distribution and services consumption.

For each data provider the PO.DAAC maintains a list of points of contact (POCs) to resolve any operational issues that arise. These issues include data outages due to safe holds and instrument and ground data processing anomalies.

The PO.DAAC also maintains linkages to the NASA Common Metadata Repository (CMR) and regularly transmits both collection (dataset) level and granule (file) level metadata to the CMR so its data can be discovered via NASA services such via the Earthdata search and the CMR API.

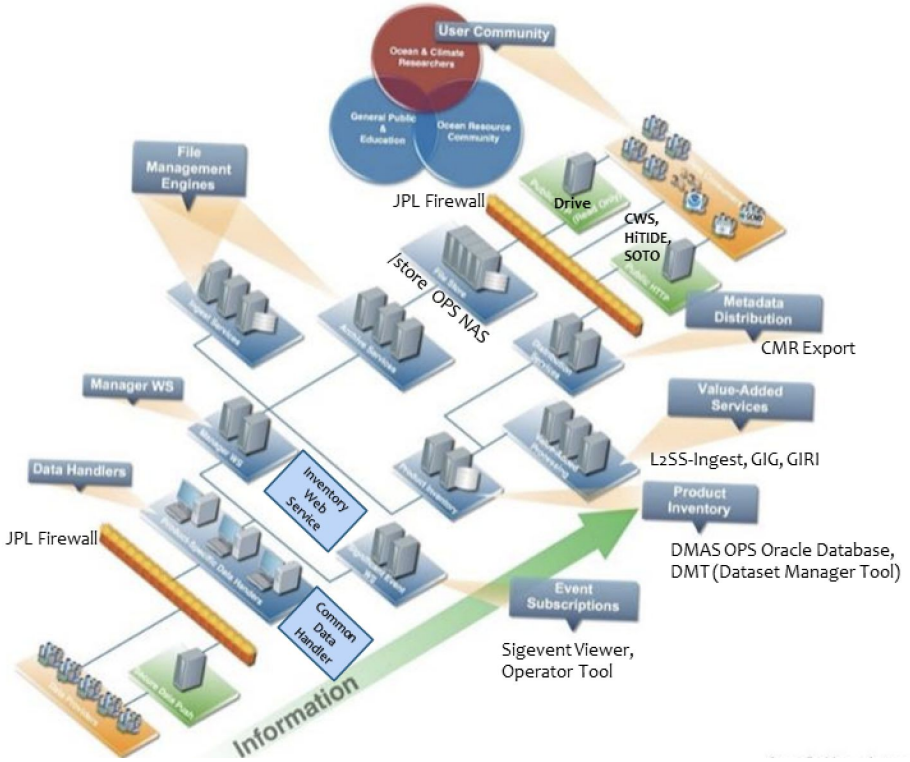

Figure 1. Architecture schematic of the PO.DAAC capturing the external interfaces for data ingestion and monitoring, internal data and metadata storage, and public consumption of data and services.

## 6. Data Processing

PO.DAAC performs data processing in support of MODIS Level 2P sea surface temperature (SST) datasets in collaboration with the NASA Ocean Biology Processing Center (OBPG). The PO.DAAC converts and reformats the OBPG MODIS Aqua/Terra and VIIRS L2P SST granules to be in compliance with the GHRSST format. The PO.DAAC has performed this task since 2006 in collaboration with the OBPG and the University of Miami to provide GHRSST compliant MODIS (and later VIIRS) L2P granules to the SST community within the 6 hour latency requirement.

PO.DAAC also produces global and regional browse imagery for all its datasets that are compliant with the requirements of the Global Imagery Browse Services (GIBS) and available via the NASA Worldview image browser.

PO.DAAC is also maintaining the operation of the SSHA MeASUREs production stream.

## 7. Data File Formats, Metadata and Quality

PO.DAAC requires data producers to produce datasets that adhere to a list of known and self-describing data formats. Self-describing refers to data and metadata structures contained within every file that make it both machine and human understandable without specialized software. A list of these formats can be found on the NASA Earth Science Standards Office website [2]. The self describing format requirement also fosters adherence to the FAIR data and metadata principles of findability, accessibility, interoperability and reusability [3].

As an exception, the PO.DAAC does distribute data that do not meet these requirements in cases where a legacy format known to the user community is preferred. In general, these legacy formats are duplicated with a self-describing format.

Vector formatted data in Environmental Systems Research Institute (Esri) Shapefile are supported.

## 7.1 Data containers and data structures

The primary recommended formats for PO.DAAC data are the netCDF4/HDF5 data containers (https://podaac.jpl.nasa.gov/PO.DAAC_DataManagementPractices#File%20Format). These are well known and NASA community endorsed data formats that are widely supported with software (libraries and package managers), visualization and data manipulation tools, and web services. Other formats recommended by the PO.DAAC include YAML ASCII encodings, ESRI Shapefiles, and GeoTIFF depending on the application, NASA mission or user community.

The data model of these formats prescribes the structure of variables and their types and dimensions, and spatial and temporal coordinate systems, and variable groupings. The accomplany metadata model of these formats is closely coupled to the metadata conventions of the Climate and Forecast (CF) Metadata Conventions and Attribute Conventions for Data Discovery (ACDD). These file metadata attributes both at the global and variable level are used to properly describe the file contents to promote data understanding and usability by both and machines.

PO.DAAC maintains a “best practice” set of CF/ACDD/ISO metadata attributes that data providers can leverage for their products. These can be found at https://podaac.jpl.nasa.gov/PO.DAAC_DataManagementPractices#Metadata%20Conventions. PO.DAAC also maintains a metadata compliance checker to determine file level metadata completeness and accuracy (https://podaac-tools.jpl.nasa.gov/mcc//).

Currently this recommended metadata model set is being updated with additional attributes and data structure types to improve support *in situ* and airborne datasets.

## 7.2 Data quality for files

PO.DAAC encourages data producers to document data files with variable quality indicators or quality flags that will allow users (or software) to determine overall observation accuracy or usefulness. These include flags for determining missing or bad data, algorithm degradation or signal contamination, or any other instrument, processing or environmental anomalies or states that assist a user in determining the suitability of the observation of interest.

If known by the data producer, variables documenting the observation uncertainty should be included. These could include observation bias, standard deviation or

RMS error, standard error, or similar statistics. How these uncertainties are calculated should be documented in a peer reviewed publication, ATBD or similar reference.

As part of the dataset accuracy assessment data producers should also provide information in the user guide to assist a user in understanding the fitness of use or purpose of the particular dataset.

# 8. Data Stewardship

The following sections describe the data stewardship and data lifecycle processes that the PO.DAAC employs for each dataset. The dataset lifecycle policy is documented here: https://podaac.jpl.nasa.gov/PO.DAAC_DataManagementPractices#Dataset%20Lifecycle

## 8.1 Ingest and Levels of Service Planning

The PO.DAAC ingest, archive, and distribution system is built to adhere to the ESDIS Archiving, Distribution, and User Services Requirements Document (ADURD) document [https://earthdata.nasa.gov/esdis/esdis-policy/adurd]. The ADURD covers the general requirements for the DAAC. PO.DAAC's services also comply with the Open Archival Information System (OAIS, see https://en.wikipedia.org/wiki/Open_Archival_Information_System) requirements and design philosophy including employing submission information packages (SIPs) and archival information packages (AIPs) for each granule ingested. Distribution information packages (DIPs) are embedded in the operational metrics collection workflows that the PO.DAAC maintains as part of the ESDIS metrics capturing activity.

While PO.DAAC staff work a typical business day schedule, its systems are operational 24x7 within an on-premise data center with backup power and active systems monitoring.

PO.DAAC supports automatic user order fulfillment for all public datasets, and posts notices of its scheduled maintenance at least two business days prior to any planned maintenance on the PO.DAAC Web site and also broadcasts the announcement via email to the user community.

A data provider specific Interface Control Document (ICD), part of the PO.DAAC data lifecycle policy, determines dataset delivery and

machine-to-machine protocols, points of contact and communication mechanisms that are agreed upon between the two parties.

As part of the communication policy, the PO.DAAC works with data providers to inform users of data outages or interruptions, or upcoming reprocessing campaigns and post operational issues.

## 8.2 Prevention of loss of data

PO.DAAC maintains on-site disk backups of all of its archived data using network attached storage (NAS), in addition to remote off-site backups of our entire historical data archive. The PO.DAAC team annually reviews the data preservation posture by measuring each of our archival strategies against risk of data loss and user impact to primary data system outage. The PO.DAAC uses the NASA "Archiving, Distribution and User Services Requirements Document" (ADURD) and the "ESDIS Data Backup Analysis Report" as a guide to this review and adjusts our archival strategies as needed [4]

Some specifics on the current logistics of PO.DAAC backup implementation include using an on-premise virtual tape library. The library stores all operational data on-premise in a secure location. Data backups are rolled off after three months. If a data restore is needed, the PO.DAAC first looks to the local virtual tape library copy before a request is made to an offsite location.

PO.DAAC maintains an offsite backup at a commercial data center located hundreds of kilometers from the on-premise data center. This facility contains one full backup plus nearly 3 years of incremental snapshots. In this scenario, any data needing restore beyond three months can be remotely delivered to PO.DAAC. In addition, the remote data center can also be switched to function as the operational NAS in case there is a disaster in the local area.

The PO.DAAC follows the preservation specification document provided by NASA [5].

## 8.3 Dataset Metadata

Dataset metadata is curated in an internal PO.DAAC database that follows the schema of the PO.DAAC metadata model. This metadata model contains a complete spectrum of metadata attributes to fully describe a dataset including its internal and public identifiers, space/time/region bounds, descriptive keywords, archive and access endpoints, available tools, software, and

services, provenance, version and citation details and many other descriptive metadata. Web services provided by the PO.DAAC can convert these contents to other formats including ECHO-10, DIF, ISO-19115 and FGDC. As noted previously, a service known as the metadata compliance checker [6] is used internally to check, iterate, and validate the file level metadata.

PO.DAAC delivers dataset and file metadata to populate the NASA ESDIS Common Metadata Repository (CMR). This DAAC-centric metadata repository contains metadata for all NASA earth science data records. Dataset metadata is known as collection level metadata in the CMR terminology. Granule metadata (the finest aggregation of data) typically corresponds to file level metadata.

Although the process of internal metadata curation and production has worked well in the past, the PO.DAAC has longterm plans to migrate its entire metadata curation lifecycle to the NASA CMR. This includes using CMR interface tools like the Metadata Management Tool (MMT) to enter and maintain dataset metadata in the repository as well as using its search API and the Earthdata Search client to provide for public dataset and granule discovery.

## 8.4 DOIs and Dataset Information Pages

All the DAACs in the NASA EOSDIS Network encourage data citations as indicated at Data Citations and Acknowledgements. Assignment and registration of DOIs to the datasets archived and distributed by the PO.DAAC are handled by the ESDIS project. Of the datasets held by PO.DAAC, 95% have been assigned DOIs, and it is expected that all datasets will eventually have DOIs. Each dataset with a DOI also has a landing page that provides the recommended citation format. For example, see https://podaac.jpl.nasa.gov/dataset/QSCAT_LEVEL_2B_OWV_COMP_12 and click on the "Citation" tab.

Each PO.DAAC dataset DOI references a unique dataset landing page containing description metadata (abstract, time/space resolutions, GCMD and other discovery keywords, etc.), data download endpoints and other tools and and services, read software, and citation details. The metadata behind each landing page is curated with entries and maintenance of an PO.DAAC database with a schema following the internal PO.DAAC project metadata model. Web based services implemented by the PO.DAAC can transform these

database contents into a number of output metadata formats including ISO-19115, FGDC, and others.

## 8.5 Collection of Associated Preservation Content

Most of the data and associated items held at the PO.DAAC are governed by NASA's Earth Science Data and Information Policy [7]. The few exceptions (e.g., documents governed by International Trade and Arms Regulation – ITAR) requiring clearance from
NASA for distribution are handled on a case-by-case basis.

As part of following ESDIS guidelines for dataset preservation the PO.DAAC dataset lifecycle has an explicit retirement checklist and policy including preserving project documentation, communications, and software if necessary. The dataset retirement (through a multi step checklist) is initiated on datasets which are deprecated by newer versions and have been removed from public visibility with reduced levels of service as determined by the dataset provider. Aside from version deprecation, dataset retirement may also result from a situation in which a proliferation of forward stream erroneous data no longer makes the data suitable for scientific use.

## 8.6 Dataset Documentation and Preparation

PO.DAAC collaborates with data providers to develop or improve dataset user guides before dataset release or curation which contain sections specific to data quality and integrity as reported by the data producer or documented through community feedback or scientific literature. The user guides also include a References and Related Publications section specific to peer-reviewed publications written about the data set. In addition, PO.DAAC maintains a listing of peer-reviewed publications from the earth science community describing their research and the use of particular datasets that can be found at the dataset specific landing page.

In preparation for the public release of a dataset, PO.DAAC performs data quality checks as a part of its data stewardship process. These quality checks include:

- Verification of file sizes, checksums and number of files
- Dataset and its contents are clearly described

- Geospatial and temporal information are complete and described
- Variables and units follow standards
- Publication or a user guide describing the data is provided
- Methodology, calibrations, and algorithms are documented
- Known issues/limitations are clearly described
- Data Quality information is clearly described
- Statements are properly referenced

## 8.7 Dataset Levels of Service

PO.DAAC strives to expose consistent levels of service (LOS) across the spectrum of its public datasets which can vary by format (e.g., netCDF, HDF, ASCII etc.), metadata content (e.g., CF, ACDD, etc.) and processing level (i.e, Level 1, 2, 3 or 4). Datasets that meet PO.DAAC recommendations for data formats and metadata content are assured a high level of compliance and access from its tools and services. The following table presents some expectations for processing Level 2 through 4. The guidelines follow the general principles and guidelines of the recently published ESDIS Level of Services [8]. PO.DAAC LOS capabilities will align directly with the recommendations and requirements of this document in the future.

| | **Dataset Processing Level** | | |
|---|---|---|---|
| **Service or Capability** | 2 | 3 | 4 |
| Data set information page (with DOI) | Default (nominal) | Default (nominal) | Default (nominal) |
| GIBS imagery | Default (nominal) | Default (nominal) | Default (nominal) |
| CMR population and Earthdata search | Default (nominal) | Default (nominal) | Default (nominal) |
| PO.DAAC Web services (dataset/granule discovery) | Default (nominal) | Default (nominal) | Default (nominal) |
| PO.DAAC Web services (granule | Requires extra integration | Not Available | Not Available |

| | | | |
|---|---|---|---|
| extraction) | | | |
| HTTPS access | Default (nominal) | Default (nominal) | Default (nominal) |
| OPeNDAP | Default (nominal) | Default (nominal) | Default (nominal) |
| THREDDS | Not Available | Default (nominal) | Default (nominal) |
| LAS | Not Available | Default (nominal) | Default (nominal) |
| HiTIDE | Requires extra integration | Not Available | Not Available |
| SOTO | Not Available | Requires extra integration | Requires extra integration |
| ERDDAP | Coming soon | Coming soon | Coming soon |

## 9. Data Discovery and Access

All the DAACs within the NASA EOSDIS Network provide the metadata for their respective archives to the Common Metadata Repository (CMR), managed by the ESDIS Project. The CMR is a high-performance, high-quality, continuously evolving metadata system that catalogs all data and service metadata records for the EOSDIS system and will be the authoritative management system for all EOSDIS metadata. These metadata records are registered, modified, discovered, and accessed through programmatic interfaces leveraging standard protocols and APIs. (See https://cmr.earthdata.nasa.gov/search).

To facilitate users' search for data, the DAACs offer several methods. At the EOSDIS Network level, the Earthdata Search provides search, preview, and access end-points for all the DAAC holdings. It also serves as a platform to feature planned EOSDIS services as they become available. (See https://search.earthdata.nasa.gov/search). In addition, the search and order tools listed at https://earthdata.nasa.gov/earth-observation-data/tools, many of which are DAAC-specific, are available for users with various specialized capabilities. Also, all the DAACs provide landing pages for datasets which have been

assigned Digital Object Identifiers (DOIs). The specific search mechanisms supported by PO.DAAC are listed below:

- Free text keyword search. Each dataset contains a curated list of associated keywords that are indexed in a PO.DAAC search service
- Faceted search. Faceted metadata attributes are presented to the user in a PO.DAAC discovery web site. This allows users to navigate and set "filters" such as science keywords, missions and satellite instruments to constrain the search request.
- Dataset landing page keyword lists. These curated keywords improve discoverability by commercial search services.
- Markup of PO.DAAC dataset landing pages to schema.org recommendations. This implementation improves discoverability by commercial search services.

## 10. User Services

PO.DAAC User Services is responsible for public communications and community engagement. It maintains a ESDIS sponsored online help desk (website) where users can ask conventional and expert level questions on PO.DAAC datasets and services. It redirects questions to appropriate PO.DAAC subject matter experts or other NASA resources as needed, and responses are categorically tracked. The PO.DAAC aspires to respond with the proper level of technical expertise to user inquiries with no more than a 24 hour latency. For significant public communications the User Services group maintains a Twitter and Facebook feed, a private email list of PO.DAAC data users, and a PO.DAAC website sub-section for data and operational announcements, dataset highlights, data-in-action stories, and PO.DAAC conference participation. User Services also maintains the PO.DAAC user online web forum where the community can post detailed inquiries or respond to existing posts.

## 11. Data Rights

PO.DAAC public data (and services) are generally provided to the user community without restriction. As part of NASA's Open Data Policy, data are not copyrighted; however, when a user publishes with PO.DAAC data or

derives results from these data, the PO.DAAC requests that an acknowledgment be placed within the text of the publication and reference list (see https://podaac.jpl.nasa.gov/CitingPODAAC).

For some data that are undergoing an active review by a science calibration and validation team or other review team, the data may be embargoed (not publically available for a specified time) as expert evaluation and fine tuning of the data processing occurs.

A small minority of PO.DAAC data are International Traffic and Arms Regulations (ITAR) controlled which are not suitable for use outside of a specific mission team. These data remain restricted.

## 12. Data Provider Data Management Plan

Data providers are encouraged to develop their own internal project DMPs to manage data production, provenance, quality and delivery. The ESDIS project provides a template and guidelines for developing a data quality and science requirement section of such a DMP (error assessment, algorithm performance criteria, high value science and mission requirements etc.) [9].

## 13. References


[1] CEOS WGISS 2016, Data Purge Alert Procedure, Version 1.0

http://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/Recommendations/WGISS_DSIG_Data%20Purge%20Alert_WP.pdf

[2] ESO Document list

https://earthdata.nasa.gov/esdis/eso/standards-and-references/eso-document-list

[3] FAIR Metadata Principals

https://en.wikipedia.org/wiki/FAIR_data

[4] Archiving, Distribution and User Services Requirements Document

https://earthdata.nasa.gov/about/esdis-project/esdis-policy/adurd

[5] NASA Earth Science Data Preservation Content Specification
https://earthdata.nasa.gov/esdis/eso/standards-and-references/preservation-content-spec

[6] PO.DAAC Metadata Compliance Checker

https://podaac-tools.jpl.nasa.gov/mcc/

[7] NASA Earth Science Data and Information Policy

https://earthdata.nasa.gov/collaborate/open-data-services-and-software/data-information-policy

[8] Level-of-Service Model, Earth Science Data Systems Program, v1.0 (February 2020)

[9] ESDIS Data Management Plan Guidance

https://earthdata.nasa.gov/collaborate/new-missions/data-management-plan-guidance

## 13. Acronyms

| Abbreviation | Description |
|---|---|
| ACDD | Attribute Convention for Data Discovery |
| ATBD | Algorithm Theoretical Basis Document |
| CEOS | Committee on Earth Observation Satellites |
| CF | Climate and Forecast |
| CMR | Common Metadata Repository |
| DAAC | Distributed Active Archive Center |
| DMAS | Data Management and Archive System |
| DMP | Data Management Plan |
| DOI | Digital Object Identifier |
| EOSDIS | Earth Observing System Data and Information System |
| ESDIS | Earth Science Data and Information System |
| Esri | Environmental Systems Research Institute |
| HITIDE | High-level Tool for Interactive Data Extraction |
| ISO | International Standards Organization |

| | |
|---|---|
| ITAR | International Traffic in Arms Regulations |
| LAS | Live Access Server |
| LoS | Level of Service |
| MEaSUREs | Making Earth Science Data Records for Use in Research Environments |
| NASA | National Aeronautics and Space Administration |
| NODS | NASA Ocean Data System |
| OGC | Open Geospatial Consortium |
| OPeNDAP | Open-source Project for a Network Data Access Protocol |
| OSTP | Office of Science and Technology Policy |
| PI | Principal Investigator |
| PO.DAAC | Physical Oceanography DAAC |
| PODS | Physical Oceanography Data System |
| SOTO | State Of The Ocean |
| SME | Subject Matter Expert |
| THREDDS | Thematic Real-time Environmental Distributed Data Services |
| UMM | Unified Metadata Model |

# 14. Appendix A

Key figure from the ESDIS Project's dataset accession process.