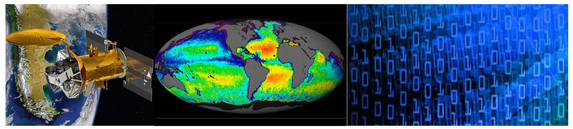
AGU Paper Number: IN53C-1579



Dataset Lifecycle Policy Development & Implementation at the PO.DAAC

V. Tsontos vardis.m.tsontos@jpl.nasa.gov
E. Tauer eric.p.tauer@jpl.nasa.gov
E. Armstrong edward.m.armstrong@jpl.nasa.gov
J. Hausman jessica.k.hausman@jpl.nasa.gov
D. Moroni david.f.moroni@jpl.nasa.gov

Jet Propulsion Laboratory/California Institute of Technology

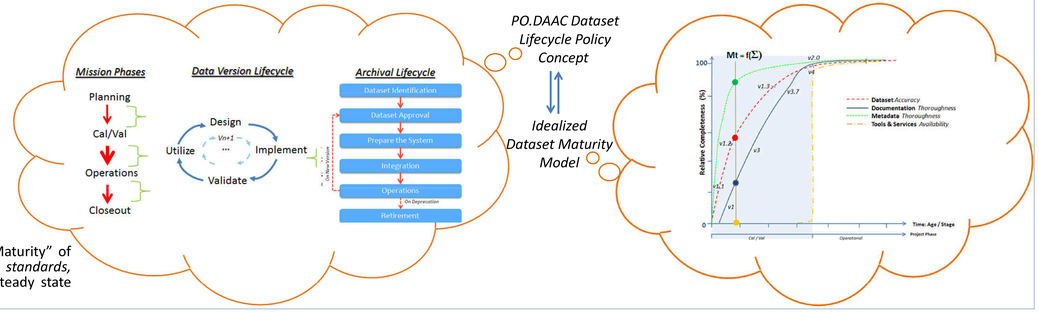
National Aeronautics and Space Administration

I. ABSTRACT

As part of on-going efforts to advance data stewardship standards and promote best practices in the field of earth science data management, NASA's Physical Oceanography Data archive (PO.DAAC) has developed and implemented a dataset lifecycle policy that governs the curation of our satellite data holdings. This lifecycle policy provides a framework, formalized process and requirements checklist useful for planning and managing datasets through the various phases of data archival at the PO.DAAC. Its purpose is to ensure quality and consistency of archived datasets, and a data management approach that is standardized, follows best practices and meets NASA, ESDIS, and DAAC requirements. Here we provide an overview of this framework including its motivation and the various lifecycle stages identified, from dataset identification/prioritization to archival planning, integration/testing, operations, ultimately through to retirement. Each phase is described in terms of its purpose, policy, associated artifacts and exit criteria, and the actors and roles involved. The generality of the framework to handle version control scenarios and the adaptability of the approach is reviewed. Usage of the dataset lifecycle policy as a tool for characterizing the status of PO.DAAC datasets and as metric that feeds into assessment of dataset maturity are also discussed.

II. Dataset Lifecycle & Policy Motivation

- Earth observation datasets exhibit complex life histories linked inherently to their maturity as successive versions are produced, validated, utilized, superseded & ultimately retired as satellite missions move between their planning, cal/val, operational, and closeout phases.
- Long-term stewardship of satellite data via the PO.DAAC system involves a sequence of distinct stages & associated processes designed to manage datasets as they transition through the various stages of their lifecycles.
- The PO.DAAC Lifecycle policy is an explicit acknowledgement of these complex relationships, providing a framework, well-characterized process, roles and requirements checklist useful for planning and managing datasets through the various phases of data archival.
- The Dataset Lifecycle and concept of Dataset Maturity are inexorably linked. "Maturity" of satellite datasets and related key attributes (eg. *data quality, metadata & file standards, documentation, interoperability*) increases with successive versions, reaching a steady state with less need for frequent revisions as the mission progresses to later stages.



III. PO.DAAC Dataset Lifecycle Policy

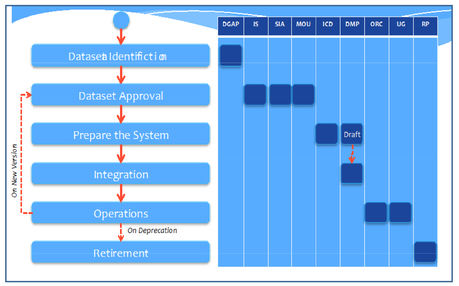
- **Governs** the end-to-end curation of PO.DAAC's satellite data holdings and related artifacts (*eg. documentation, software etc.*)
- **Purpose**: to ensure PO.DAAC data management approach …
  - Is consistent across our holdings
  - Follows best practices
  - Is adequate to meet our requirements
  - Promotes archival of highest quality data and advocacy for data standards
- ***"Document Driven"* & measured** by the artifacts required to manage the selection and integration of datasets into the operational PO.DAAC system.
- Uses **standardized document templates**
- **Well-Defined Processes & Roles** underlie the lifecycle policy and associated document completion

Document	Abbrev.	Source/Distribution
Dataset Gap Analysis & Prioritization	DGAP	PO.DAAC/Internal
Integrated Schedule	IS	PO.DAAC/Internal
System Impact Assessment	SIA	PO.DAAC/Internal
Dataset Submission Agreement	MOU	Provider – PO.DAAC
Interface Control Document	ICD	Provider – PO.DAAC
Data Management Plan	DMP	Provider – PO.DAAC – ESDIS
Operational Readiness Checklist	ORC	PO.DAAC/Internal
Data User's Guide	UG	PO.DAAC/Public
Retirement Plan	RP	Provider – PO.DAAC – ESDIS

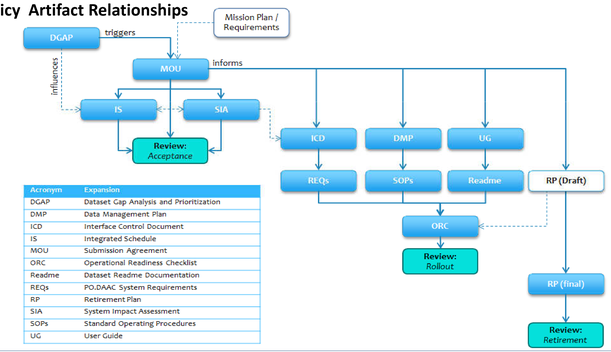
IV. Lifecycle Phases & Policy Documents

- Policy document(s) and applicable management process(es) relative to data archival phase over the dataset lifecycle are shown adjacent.
- Movement of a dataset between phases requires that all previous assessments have been undertaken, information captured in associated policy documents, and stage exit criteria met.

Eg: Dataset identification & prioritization for PO.DAAC archival of non-mandated/mission dataset occurs via DGAP, and approval of a dataset for archival first requires the completion of an Submission Agreement (MOU) with the data provider and due consideration of scheduling (IS) and assessment of system impacts (SIA).



V. Lifecycle Policy Artifact Relationships



VI. Lifecycle Policy Stages, Artifacts & Roles

1. Identification & Prioritization

(applicable to non-mandated NASA mission data)

Approver: PO.DAAC User Working Group (UWG)
Responsible Actor: Science Team Lead
Exit Criteria: Generation of Dataset Gap Analysis & Prioritization (DGAP)

Description: This is the entry point into the Lifecycle of a dataset. It includes the periodic evaluation of datasets collected from traditional sources, or those that have been suggested by the community within or external to the PO.DAAC project. (Note that NASA mission data are mandated for archival, so this step applies principally to PI-provided datasets). This phase controls the identification of quality/high value datasets and their promotion as candidate data for archival based upon its scientific significance, as captured in the DGAP spreadsheet.

2. Approve the Dataset for Archival

Approver: Project Manager
Responsible Actor: Assigned Data Steward
Project System Engineer
Exit Criteria: Generation of MOU, IS, SIA

Description: In this phase, a new dataset is evaluated to determine its impact upon the data management system, the NASA New Products process, and the PO.DAAC Dataset Lifecycle Policy are reviewed to identify compliance and any areas that will pose a significant challenge. The Cost/Benefit analysis is completed and the goals of PO.DAAC and the user community measured against the resources required to bring the dataset into the PO.DAAC holdings. Also, high-level, dataset-related metadata, contacts, information, etc. regarding this dataset are captured in the Dataset Submission Agreement.

3. Prepare the System

Approver: Project System Engineer
Responsible Actor: Assigned Data Steward
Exit Criteria: Generation of ICD & DMP*(draft)*

Description: The method to obtain data granules of this dataset is determined and verified, and the means to capture items related to the dataset's scientific validation are identified and verified as do-able. In this phase, the Data Stewards and Deployment team determine and verify how to process the data and how to extract the required metadata. Data reader routines are collected or written and tested, and detailed guides for interpreting the data are drafted. The initial validation strategy, including PI or provider-side validation, is also determined.

4. Integration

Approver: Project System Engineer
Responsible Actor: Deployment Lead
Exit Criteria: Generation of DMP *(final)*

Description: Completion of all work needed to enroll this dataset into the catalogs, FTP sites, etc. The dataset is integrated, as applicable, into search engines, visualization tools, added to accountability and metric collection, and dataset information pages are produced. The DMP is finalized, refining the draft based on knowledge gained from the activities in this phase, including how operations will be carried out, and how the dataset's granules will be added to data access services (eg. THREDDS, OPENDAP, FTP) and tools (eg. HiTIDE, LAS, SOTO). In this phase, the metadata capture and mapping per the plan and data dictionary is also verified.

5. Rollout (*Gate*)

Approver: Project Manager
Responsible Actor: All Project Leads
Exit Criteria: Generation of ORC

Description: Not a phase per se, Roll-Out is an opportunity to review the adherence of the dataset to the Lifecycle Policy and the completion of the Operational Readiness Checklist (ORC). This phase also requires the review of liens and waivers, which represent respectively required deferred ORC deliverable and ORC action items that are not required for rollout of that specific dataset. On satisfactory review, the release of the dataset to operations is approved and it is made available to the target audience. (Note: This is a highly critical readiness review prior to public dataset release. It is a gate where all Operational documents and artifacts are due and reviewed. If all is satisfactory, this step ends with the dataset being cleared for distribution.)

6. Operations

Approver: Project Manager
Responsible Actor: Operations Lead
Exit Criteria: Authorization to retire dataset.

Description: This phase includes any actions needed to maintain the dataset in-house over the projected lifespan, including standard operational procedures and activities, as well as any deviations identified in the DMP. This will likely represent the largest amount of time, but the minimal amount of direct activity. It includes ongoing audits and reviews of Ops and Data Stewardship leads, and includes off-nominal cases such as re-processing, superseded datasets, and versioning. Any modifications to the dataset or its disposition within the system requires review of the full lifecycle.

7. Dataset Retirement

Approver: Project Manager
Responsible Actor: Assigned Data Steward
Driving Artifact: RP

Description: This phase is an exception to the default long-term storage of a dataset, but is a valid use case within the PO.DAAC system. This phase treats dataset retirement, obsolescence, corrupted data, and de-activation. For PI-provided datasets, the PO.DAAC User Working Group and Project Science Team Lead function as the primary approvers. Directed projects/missions dataset retirement is governed by NASA/ESDIS mission closeout policy guidelines.

URS CL# 240544