GHRC DAAC Dataset Life Cycle Document

Introduction

Established in 1991, the Global Hydrology Resource Center (GHRC) is one of twelve NASA Distributed Active Archive Centers (DAACs), and is managed jointly by the Earth Science Department at NASA's Marshall Space Flight Center and the University of Alabama in Huntsville's Information Technology and Systems Center. GHRC is a full-service science data center, performing data ingest, processing, archive, cataloging, documentation, distribution and user services. GHRC is a member of national and international data organizations including the Federation of Earth Science Information Partners (ESIP), NASA's Earth Science Data and Information System (ESDIS), and the International Council for Science (ICSU) World Data System (WDS).

The GHRC ingests, processes, archives and distributes data from several thematic collections including global lightning data from NASA and DoD satellites as well as ground-based lightning detection networks, long-term climate data records from numerous passive microwave instruments, and airborne and ground-based observations from Hurricane Science field campaigns and Global Precipitation Measurement (GPM) Ground Validation experiments.

GHRC Mission

The mission of the GHRC is to provide a comprehensive archive of both data and knowledge augmentation services with a focus on atmospheric phenomena, their governing dynamical and physical processes, and associated environmental applications. Unique capabilities include:

- Providing data stewardship services to both NASA's Research and Applied Science missions;
- Focusing on hazards and disasters to help accelerate operational use of NASA data, with a goal of serving as the interface to other agencies wanting NASA data for specific hazards, past and/or present;
- Providing a state-of–the-art infrastructure to support all stages of Field Campaigns from mission planning and coordination to data archive and publication.

Funding Support and Guidance

GHRC is funded and overseen by NASA's ESDIS Project. Additional guidance is provided by the GHRC User Working Group, who serve as representatives of the GHRC user community. The principal purposes of the UWG are to firmly establish science user input in the planning, development, and operations of the GHRC, and to represent the science user communities in reviewing and guiding the GHRC activities and development. Membership is drawn from the science community appropriate to the GHRC's discipline areas.

DAAC Requirements

NASA DAAC requirements are detailed in NASA's "Requirements for Archiving, Distribution and User Services Document (423-10-69)." Key requirements include:

- Archive and distribute the data for a given mission for as long as the data are being actively used by NASA funded investigators.
- Ensure that the data are transitioned to the appropriate Long-Term Archive when NASA notifies the DAAC that it is appropriate to do so.
- Have a data search and order/request/access function that 1) provides users with information about the available data products (e.g., guide documents), 2) gives users the capability to identify and select their desired information and data products before ordering/requesting/accessing, and 3) delivers the requested products to users.
- Provide users with functionalities to manipulate certain data products prior to ordering (e.g., spatial and/or spectral subsetting).
- Provide a User Services function to assist users with questions, for example, regarding data formats, data usage, system access, etc.

Additional requirements and guidelines include those detailed in these documents:

- NASA Earth Science Data Preservation Content Specification
- Digital Object Identifiers (DOIs) for EOSDIS

Key Data Holdings

Current

Data Collection	Description
Lightning observations from space	Daytime and nighttime observations of cloud-to-ground and intra-cloud lightning from NASA's Lightning Imaging Sensor (TRMM) and Optical Transient Detector (Microlab1)
DISCOVER MEaSUREs	Atmospheric and ocean products derived from satellite microwave radiometer measurements and delivered as a package of inter-related geophysical parameters, including: • sea surface wind speeds • atmospheric water vapor • cloud liquid water • rain rates
Hurricane Science field campaigns	Airborne and ground-based observations of tropical cyclones and cyclogenesis collected during a series of campaigns from 1998 to the present
GPM Ground Validation experiments	Precipitation datasets from ground and airborne instruments supporting physical validation of satellite-based precipitation retrieval algorithms, from targeted field observations in different precipitation regimes (2010-2015).

Future

Dataset	Description
Lightning observations from space	LIS on International Space Station: a second LIS instrument to be deployed on ISS in 2016

Data Stewardship and Levels of Service provided by GHRC

GHRC ranks between Level 3 and Level 4 on the five point data stewardship maturity matrix proposed by Peng et al., which covers functional areas including accessibility, usability, integrity, etc. While Peng's data stewardship maturity measures provide an overview of the general level of service the GHRC provides across various data management functions, the specific services provided for different categories of data will vary. Levels of service for each function, as applied to different data categories, are discussed in a separate Levels of Service document.

GHRC Data Life Cycle

Evaluation of Proposed New Dataset

New datasets may be assigned to the GHRC by ESDIS, or they may be added to the GHRC archive by request of the data provider or on the suggestion of the UWG. In any case, the GHRC will ask the science data provider or principal investigator (PI) for basic information about the proposed new dataset using the GHRC Data Acquisition Form. Currently, the Data Acquisition Form is a spreadsheet exchanged via email. Future plans call for an online form that the data provider will fill in. GHRC Data Management personnel will review the PI-provided information for completeness, and discuss data publication timeline and levels of service with the provider.

For those datasets not assigned by ESDIS, GHRC personnel will prepare a data assessment package including a description of the proposed dataset and its science value, along with an estimated cost for acquiring and publishing the data if this can't be accomplished within existing resources. The GHRC User Working Group will evaluate this package, and depending on UWG recommendation, GHRC will forward the data assessment package to ESDIS for evaluation (Figure 1).

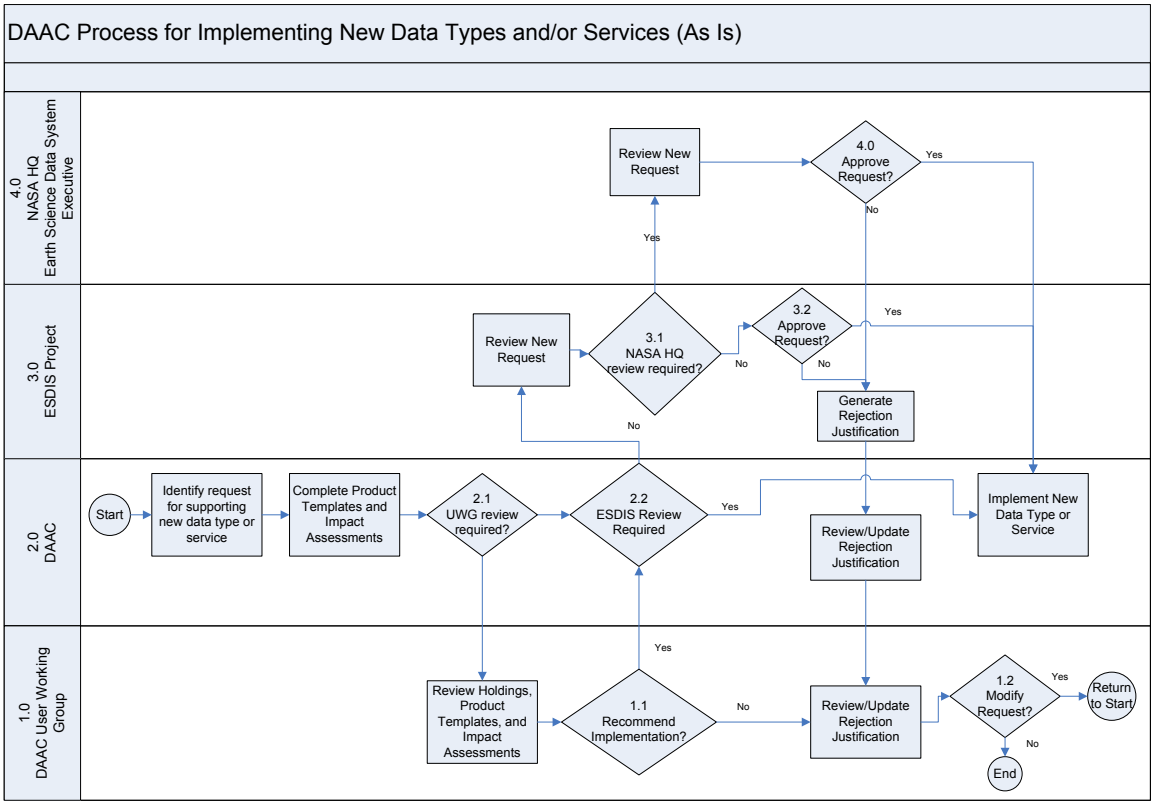


Figure 1. New dataset evaluation process

If the proposed dataset is not approved by either the UWG or ESDIS, GHRC will communicate the reason for rejection to the data provider. If the data provider chooses to update the data assessment package with additional information about the data and its scientific value, GHRC may initiate an additional review cycle with UWG and ESDIS.

Acquisition of New Dataset

After a new dataset, collection of datasets, or mission is approved by ESDIS, GHRC staff will begin the acquisition process, shown in Figure 2. The major steps in this process are outlined in the following sub-sections.

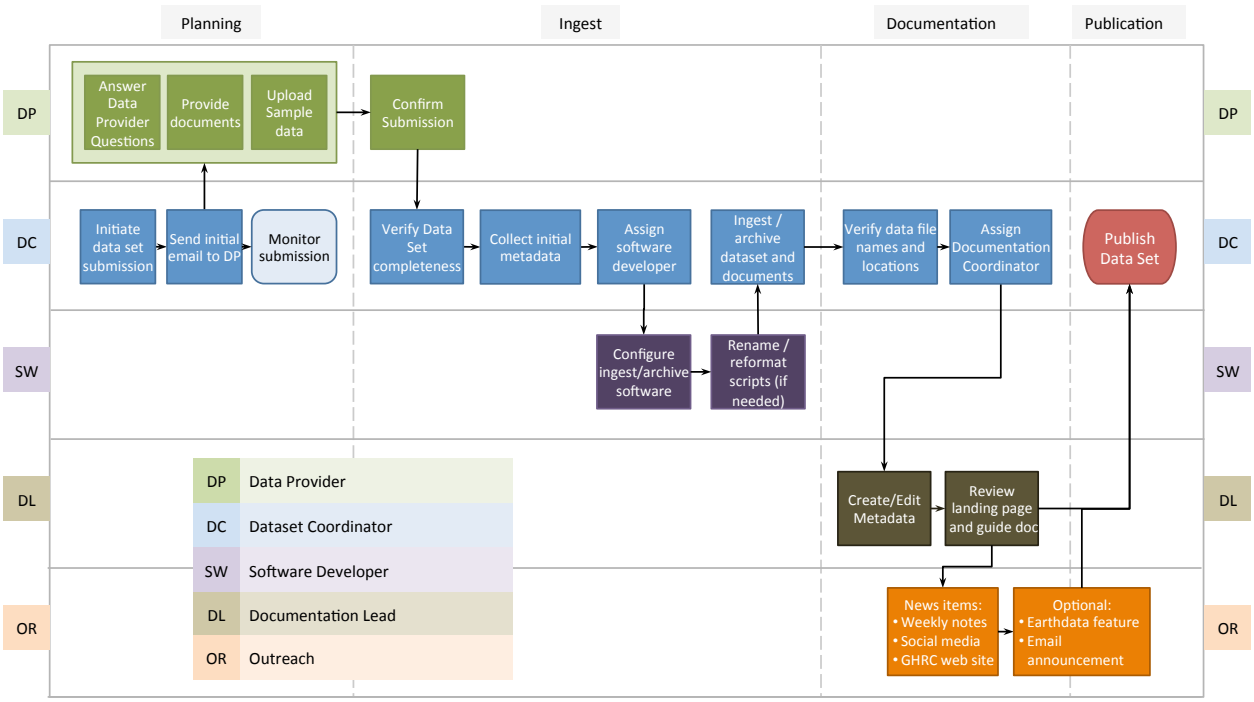


Figure 2. Process for acquiring and publishing a new dataset

Planning Phase

First, GHRC staff will engage with the science team in order to better understand the data, and to review data formatting and metadata requirements based on the agreed-upon level of service for the dataset. This collaboration may include:

- Attending science team conferences/telecoms
- Email correspondence, virtual or face-to-face meetings with science data providers in order to
  - Review detailed metadata needs listed in Internal Data Information Worksheet
  - If the data collection includes multiple datasets, prioritize datasets for publication and distribution
  - For each dataset, define data citation including dataset title and authors
  - Provide information community standards for data format and metadata

    - Agree on file naming convention
    - Develop a data and information product release plan (data publication)
    - Agree on archiving requirements including criteria for versioning and retiring datasets in the collection
    - Identify unique tools and services appropriate to the data collection
    - Identify applicable downloadable tools, clients and online services
- For data collections with a higher level of service, GHRC and science team will work together to develop a dataset or collection specific Data Management Plan following the "Guidelines for Development of a Data Management Plan (DMP)" document from the Earth Science Division of NASA's Science Mission Directorate and GHRC's Internal Data Information Worksheet. The DMP should identify all science artifacts needed using the NASA Earth Science Data Preservation Content Specification (PCS).
- GHRC may also develop an Interface Control Document or Inter-Project Agreement with data provider, if applicable.
- Develop preservation policies following the PCS, considering these factors:
    - Experiment repeatability – software packages needed for production processing
    - Items archived by other agencies
    - Version obsolescence rules
    - Golden copy of each version (for processing)
    - Algorithm History
    - Retention of data used in publication
    - Strategies for retiring datasets

At the end of this phase, the data provider will have supplied initial metadata, documentation and sample data files.

Ingest and Documentation Phase

This phase entails planning and implementation of GHRC procedures to transfer data files to the DAAC, make any adjustments needed to file format or names, stage files to data server and archive, create metadata and generate metrics.

- Plan for data ingest
    - Acquire sample products to drive migration testing
    - Request Digital Object Identifier (DOI)
    - Create an Internal Data Information Worksheet to identify
        - data source and ingest protocols
        - basic collection level metadata
        - processing steps if applicable
        - local data server and archive locations
    - Implement and test data ingest software and procedures
        - data acquisition and staging
        - renaming or reformatting data files if needed

        - granule level metadata
        - ingest/archive metrics
- Ingest new products
    - Create collection level metadata record in GHRC catalog
    - Execute and monitor ingest software defined above

Publication Phase

Data publication includes making data and documentation available online, configuring any applicable data services, registering new datasets in local and NASA ESDIS search systems, and notifying the user community.

- Enable Data Services
    - Accessibility
        - Publish metadata to the ESDIS Common Metadata Repository (CMR – ECHO and GCMD)
        - Publish metadata in HyDRO
        - Publish DOI
        - Provide direct file downloads
        - Configure OPeNDAP access
    - Usability
        - Publish PI-provided documentation and online Data User Guide document.
        - Provide links to any visualization tools
- Advertise new and updated products
    - Weekly notes
    - Feature on GHRC web site
    - Email to community and ESDIS outreach coordinator (Optional)
- Issue notification to stakeholders to release producer from primary access and discovery responsibilities

Data Versioning

GHRC follows NASA ESDIS's guidelines on data versioning, as defined in the NASA Earth Science Data Preservation Content Specification.

> For all products held in the archive, documentation of processing history and production version history, indicating which versions were used when, why different versions came about, and what the improvements were from version to version. For all products held in the archive, the versions of source code used to produce the products should be available at the archive. Granule level metadata should indicate which version of software was used for producing a given granule. In the case of some datasets all versions of products may be maintained. In other cases, only the latest and penultimate versions may be maintained, with some samples of product granules of each of the historical versions. In the case where

different versions of ancillary, input data, or calibration were used, the history of those changes should be available as part of the processing history.

A new dataset or dataset version is determined by an update to the science algorithm, not subsetting or format translation that preserve science data values. A separate DOI is obtained for each version of a dataset. GHRC will maintain at least the current and most recently superseded published versions of any given dataset in the archive. Decisions about whether to maintain earlier versions of the data will be made on a case by case basis, depending on data volume, science value, and usage metrics. Only the latest version of a dataset will be available on line and listed in data search catalogs. However, the DOI for each version will be maintained indefinitely, with a landing page providing information about that particular version of the data and referencing the current version. When a new version of a dataset is received, GHRC will confirm with the PI whether an earlier version can be distributed to users on request, or if access should be restricted. The process for acquiring, documenting and publishing a new version of a dataset is similar to that for acquiring a new dataset. The red circles in Figure 3 indicate the major differences. Typically the level of effort is much lower, because metadata, documentation and data handling software should be very similar to what was used for previous versions.

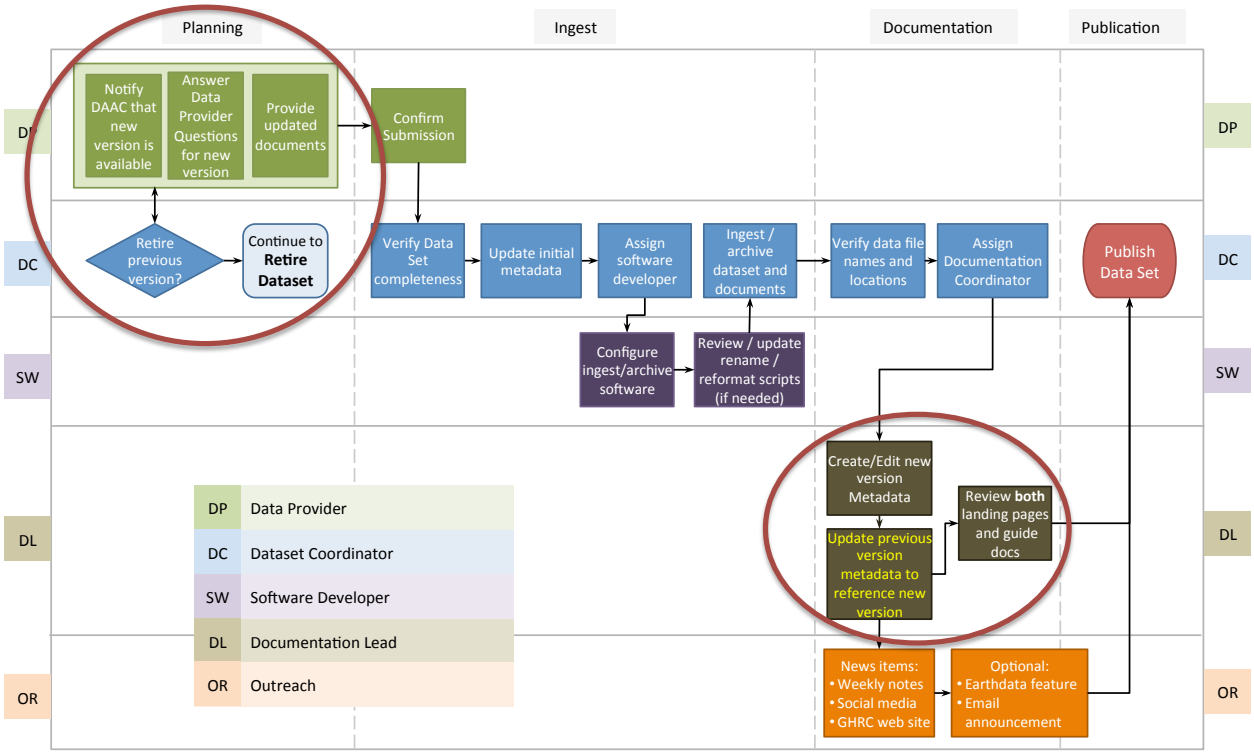


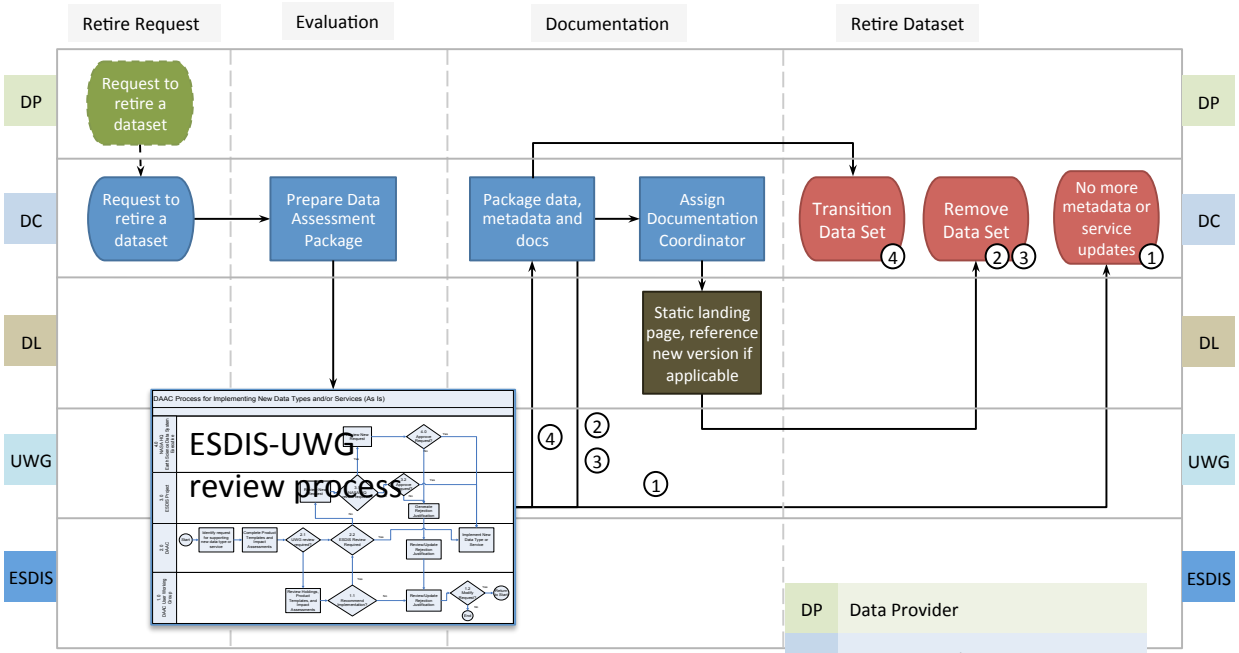
Figure 3. Process for dataset version updates

In some cases, the GHRC will publish and distribute data in a different format than that provided by the PI. For example, we may translate a dataset to a standard format like

netCDF for long-term archive and distribution. The original data as provided by the PI will be archived for provenance and preservation purposes. Both the original and standard format of the dataset may be published and distributed, particularly if the original data format is widely used in the community. However, if the original data formatting is preserved, but augmented with additional in-file metadata (e.g., CF-compliant metadata added to a netCDF file), the original data may be discarded after automated checks to make sure all data values are unchanged.

Strategies for Retiring Datasets

Occasionally a GHRC dataset may be deprecated to a lower level of service upon request of the data provider, due to a decline in usage metrics or science relevance, or because it has been superseded by another dataset. Such deprecation may involve deactivation of data services, a notation in the data catalog recommending use of another dataset instead of this one, removal of a dataset from the catalog, or even removal from the archive altogether (Figure 4). As online storage costs continue to decrease, it is unlikely that a deprecated, but still publicly available, dataset would be removed from the GHRC's public data server.



Options to retire a dataset:

① Leave data available online with low level of service
② Remove data from online server and public catalog, keep on archive
③ Remove from online server, catalog and archive
④ Transition to long term archive

Figure 4. Process for retiring a GHRC dataset

Any datasets slated for deprecation will be reviewed by the GHRC User Working Group. The review package will include a brief description of the dataset, proposed rationale for deprecation, definition of the proposed reduced level of service, and an estimate of cost savings based on service reduction. If the UWG recommends removal of a dataset or transition to a long term archive (both higher cost options than leaving the dataset online with a lower level of service), ESDIS will be asked to review this recommendation.

References – Applicable NASA Requirements and Guidelines

Guidance for Development of a Data Management Plan: http://science.nasa.gov/earth-science/earth-science-data/data-management-plan-guidance/

NASA Earth Science Data Preservation Content Specification: https://earthdata.nasa.gov/standards/preservation-content-spec

Digital Object Identifiers for EOSDIS: https://wiki.earthdata.nasa.gov/display/DOIsforEOSDIS/Digital+Object+Identifiers+%28DOIs%29+for+EOSDIS

Requirements for Archiving, Distribution and User Services Document (423-10-69) for NASA DAACs

Peng, G., Privette, J. L., Kearns, E. J., Ritchey, N. A., & Ansari, S.. (2015). A Unified Framework for Measuring Stewardship Practices Applied to Digital Environmental Datasets. Data Science Journal, *13*(0), 231–253. DOI: http://doi.org/10.2481/dsj.14-049