Data and Documentation Preservation Systems at the NASA GES-DISC

ESIP Summer 2014 Meeting
Frisco, Co August 9, 2014

mo.khayat@nasa.gov
http://disc.gsfc.nasa.gov

NASA/Goddard Earth Sciences Data and Information Services Center (GES DISC)

An open-source system for preserving data and documentation from HIRDLS, TOMS, UARS and other NASA earth science missions

Mo Khayat[1,2], James Johnson[1,2], Irina Gerasimov[1,2] Jennifer Wei[1,2], Barbara Deshong[1,2], Steve Kempler[1], Ed Esfandiari[1,2], Michael Berganski[1,2]

[1]NASA Goddard Space Flight Center, [2]ADNET Systems Inc.

Introduction

Many NASA Earth Observing System (EOS) missions have either already reached the end of their active life or are nearing it. Preservation of data products is a fairly well defined task for the NASA EOS Data Centers or DAACs. However, supporting documentation and other artifacts from these missions are also critical to the long-term studies of our planet's climate, and to aid future generation's ability to understand climatic changes. The challenge is how to preserve these items along with the traditional data products.

The Goddard Earth Sciences Data and Information Services Center (GES-DISC) has implemented a Repository System to facilitate the long-term archive of documentation artifacts and other associated digital content. The GES-DISC designed this system based on Fedora Commons, an open-source repository management software, for cost savings and flexibility.

The first mission to utilize the GES-DISC Repository System was the recently completed High Resolution Dynamics Limb Sounder (HIRDLS) on the Aura spacecraft. Data and documentation from the Upper Atmosphere Research Satellite (UARS) and the Total Ozone Mapping Spectrometer (TOMS) have recently been completed as well. The GES-DISC is negotiating the transfer of data preservation items from the current Microwave Limb Sounder (MLS) on Aura, and the Atmospheric Infrared Sounder (AIRS) missions before they end.

NASA Earth Science Data Preservation Content Specification (423-SPEC-001)

Being able to understand and interpret data from these older missions after the experts familiar with them have moved on or are no longer available is a concern to NASA. The importance of preserving the Earth Science data and documentation resulted in the issuance of the "NASA Earth Science Data Preservation Content Specification" (423-SPEC-001) for the Earth Science Data and Information System (ESDIS) supported data centers. Documents and data in the GES-DISC repository are archived and classified according to the 423-SPEC-001 into 9 categories listed in the table below for each mission.

1. Category	2. Content Item	3. Definition/Description
Preflight/Pre-Operations Calibration	Instrument Description	Documentation of Instrument/sensor characteristics including pre-flight or pre-operational performance measurements (e.g., spectral response, instrument geometric calibration (geo-location offsets), noise characteristics, etc.).
	Preflight/Pre-operational Calibration Data	Numeric (digital data) files of Instrument/sensor characteristics including pre-flight or pre-operational performance measurements (e.g., spectral response, instrument geometric calibration (geo-location offsets), noise characteristics, etc.).
Science Data Products	Raw Data and Derived Products	Raw data are data values at full resolution as directly measured by a spaceborne, airborne or *in situ* instrument. Derived products are higher level products (level 1b through 4) where calibration and geo-location transformations have been applied to generate sensor units, and/or algorithms have been applied to generate gridded geophysical parameters.
	Metadata	Information about data to facilitate discovery, search, access, understanding and usage associated with each of the data products.
Science Data Product Documentation	Product Team	Names of key science team leads and product team members (development, help desk and operations), roles, performing organization, contact information, sponsoring agencies or organizations and comments about the products.
	Product Requirements	Requirements and designs for each science data product, either explicitly or by reference to the requirements/design documents. Product requirements and designs should include content, format, latency, accuracy and quality.
	Processing and Algorithm Version History	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.
	Product Generation Algorithm	Detailed discussion of processing algorithms, outputs, error budgets and limitations. Processing algorithms and their theoretical (scientific and mathematical) basis, including complete description of any sampling or mapping algorithm used in creation of the product , geo-location, radiometric calibration, geophysical parameters, sampling or mapping algorithms used in creation of the product, algorithm software documentation, & high-level data flow diagrams. Description of how the algorithm is numerically implemented.
	Product Quality	Description of the impact to product quality due to issues with computationally intensive operations (e.g., large matrix inversions, truncation and rounding). Documentation of product quality assessment (methods used, assessment summaries for each version of the datasets). Description of embedded data at the granule level including quality flags, product data uncertainty fields, data issues logs, etc. Relevant test reports, reviews, and appraisals.
	Product Application	Useful references to published articles about the use of the data and user feedback received by the science and instrument teams about the products. Includes reports of any peculiarities or notable features observed in the products.
Mission Data Calibration	Calibration Method	The methods used for instrument/sensor radiometric and geometric calibration while in operation (e.g., in orbit). The source code used in applying the calibration algorithms. Documentation of in-line changes to calibration or to instrument or platform operations or conditions that occur throughout the mission.
	Calibration Data	Instrument and platform engineering data collected during operations (e.g., on orbit), including platform and instrument environment, events and maneuvers; attitude and ephemeris; aircraft position; acquisition logs that record data gaps; calibration look-up tables; and any significant external event data that may have impacted the observations.
Science Data Product Software	Science data product generation software and software documentation	Source code used to generate products at all levels in the science data processing system. Software release notes, including references to versions of operating systems, compilers, commercial software libraries used in the code. Versions of science data product software should be archived for each major product release. A major product release is characterized by the appearance of peer reviewed publications where reported results are based on the product version.
Science Data Product Algorithm Inputs	Ancillary data and documentation	Complete information on any ancillary data or other data sets used in generation or calibration of the data set or derived product, either explicitly in data descriptions or by reference to appropriate publications. Ancillary data should be stored with the products unless it is available from another permanent archive facility.
Science Data Product Validation	Datasets and documentation	Accuracy of products, as measured by validation testing, and compared to accuracy requirements. Description of validation process, including identification of validation data sets, measurement protocols, data collection, analysis and accuracy reporting.
Science Data Software Tools	Software and documentation	Product access (reader) tools. Software source code that would facilitate use of the calibration data, ancillary data and the data products at all levels. Includes software source code useful for creating programs that will read and display the calibration data, ancillary data and product data and metadata values. Commercial tools should be identified with appropriate references.

Fedora Commons Interface

The GES-DISC used Fedora Commons, an open-source repository management system that is used in many universities, research centers, and libraries. It comes with a simple web-based GUI interface which provides for easy administration of the system. The GUI also allows one to enter objects or datastreams (these can be of any type document, image, source code, binary data, etc.) into the system. The system uses XML to manage the objects. The GES-DISC has also developed a command line script to allow batch ingest of objects into the Fedora Repository.

Figure 1. Internal GUI used to ingest and archive records into the repository (left), and internal GUI used to search and retrieve records from the repository (right).

Public Access of Preservation Documents

External users access the publicly available documents by visiting the mission specific documentation page for that instrument. The Fedora repository system is at the backend and makes access to the linked documents possible. Note that restricted objects (ITAR, proprietary, or software) are not accessible through the public interface. Three missions are now public:

HIRDLS http://disc.sci.gsfc.nasa.gov/Aura/additional/documentation/hirdls-preservation-documents
TOMS http://disc.sci.gsfc.nasa.gov/acdisc/documentation/toms-mission-preservation-documents
UARS http://disc.sci.gsfc.nasa.gov/acdisc/documentation/uars-mission-preservation-documents

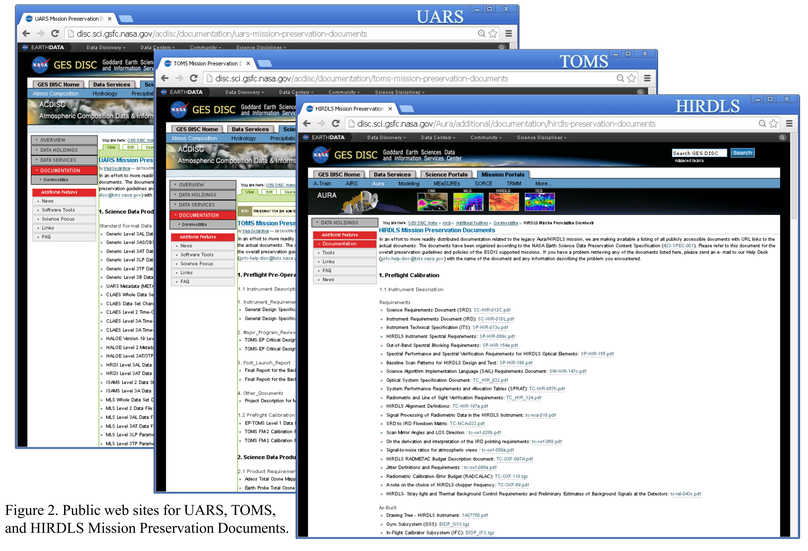


Figure 2. Public web sites for UARS, TOMS, and HIRDLS Mission Preservation Documents.

GES DISC Preservation Implementation

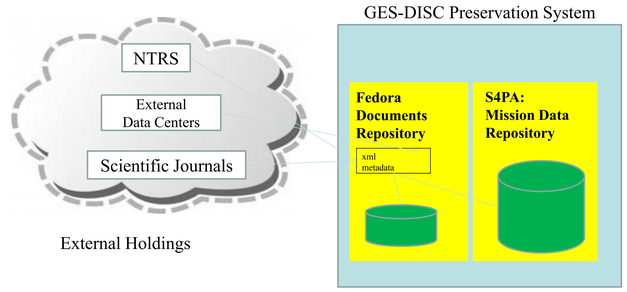
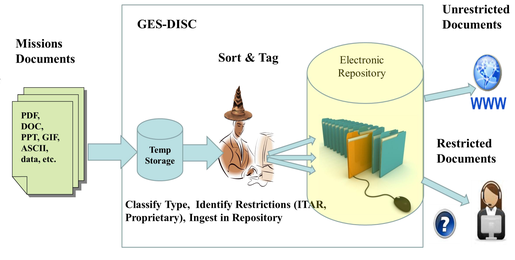


Figure 3 - Overview of the GES DISC data and documentation preservation systems. Artifacts may reside at the GES DISC or be external, (e.g. NASA Technical Reports Server (NTRS), another Data Center, or in a scientific or technical journal

1. Identify documentation
   GES DISC Science Support staff identified specific information needed per mission in Data Preservation Mission List by working closely with original mission teams to sort out documents for preservation.
2. Specify and implement preservation environment
   Local archive based on open-source system Fedora Commons.
   Implementation is complete for HIRDLS, TOMS, and UARS datasets. Other missions AIRS, MLS, TRMM, etc. are in progress.
   Exploring NASA Technical Reports Server (NTRS) and NASA Aeronautics and Space Database (NA&SD) as repository for restricted documents
3. Retrieve documentation
   Public documents accessed by users on mission portal pages.
4. Implement retrieval and distribution services
   1. *Access for internal GES-DISC users*
   2. *External Access via WWW for unrestricted documents*
   3. *External Access for restricted documents (ITAR) via User Services contact*
   4. *External Access for restricted documents via authentication (TBD)*
   5. *Iterate with other DAACs/community*

Figure 4 - Overview of the physical objects sorting, tagging, storage in archive and distribution system.

Restricted documents are currently only available by contacting the GES DISC user services.



Lessons Learned, Challenges, and Future Plans

- Heritage missions require extensive work to identify and classify documents
- Restricted (ITAR or Proprietary) vs. Unrestricted requires special handling
- Limited utility of NASA infrastructure like NTRS (not capable of accepting all STI)
- Incorporate DOI metadata into repository (if available)
- Level of service to provide external users if NTRS not a viable option
  No distribution vs. case-by-case (subject to export control rules, authentication)?

References

NASA Earth Science Data Preservation Content Specification (423-SPEC-001) H. K. Ramapriyan, EOSDIS Project Office, NASA GSFC
https://earthdata.nasa.gov/sites/default/files/field/document/423-SPEC-001_NASA%20ESD_Preservation_Spec_OriginalCh01_0.pdf

Evolution of Information Management at the GSFC Earth Sciences (GES) Data and Information Services Center (DISC), IEEE Transactions on Geoscience and Remote Sensing, Volume 47, Issue: 1, 2009