Implementation of the Open Source Document Preservation System at the NASA GES DISC

Mo G Khayat[2, 1], Steven J Kempler[1], Barbara DeShong[2, 1], Ed Esfandiari[2, 1], James Edward Johnson[2, 1], Irina V Gerasimov[2, 1], Michael R Berganski[2, 1], James G. Acker[2,1]

1. Goddard Earth Sciences- Data & Information Services Center (GES DISC), NASA Goddard Space Flight Center, Greenbelt, MD, United States.
2. ADNET Systems Inc., Rockville, MD, United States.

NASA's earth observation missions commenced with the Television Infrared Observation Satellite (TIROS) series in the 1960s and continued with the Nimbus and Landsat satellite missions beginning in the 1960s. The Nimbus satellites inaugurated multi-sensor missions for environmental remote sensing. In the ensuing four decades, NASA's Earth science activities have led to increasingly sophisticated satellite instruments, much larger data volumes, more complex data analyses, and a diverse suite of data products generated with sophisticated data algorithms. NASA now has at its disposal a huge amount of information about the state of our planet obtained from the vantage point of polar and low earth orbit satellites. For scientists seeking to study Earth's changing climate, having long-term time series of data on key climate variables is crucial. The data from these missions constitute a vital archive for Earth science research.

However, making full use of this information is far from trivial. The past four decades have seen a veritable revolution in information technology as digital storage media, for example, have evolved from magnetic tapes to optical compact discs and solid state technology. There has been a similar rapid evolution of both the amount of data collected and the means used to collect and preserve data from satellites. Over time, data collected by an instrument and stored in one media type (film, magnetic tapes, solid state hard drives, etc.) has to undergo a refresh of media type in order to be useable. This constant refresh is one important aspect of preservation; however, the complexity of the usability of this preserved data is compounded when users encounter a scarcity of documentation to describe the data or the processing steps employed in the algorithms or the generation of higher level products. This issue becomes more acute as the original principal investigators or scientists most familiar with these details move on to other project or leave behind inadequate documentation for later generations. It has become imperative therefore to not only preserve the original data, but also any relevant documentation that is produced in the lifecycle of a mission.

Many Earth Observing System (EOS) missions have either reached the end of their active life or are nearing it. Preservation of data and artifacts from these missions is critical to long-term studies of our planet's climate and to aid future generation's ability to understand climatic changes. Data from these legacy missions provide valuable time comparisons of environmental conditions without which our climatology analysis would be incomplete. NASA has recognized the importance of preserving Earth Science data and documentation by maximizing the capture and safekeeping of mission related artifacts (actual data proper, as well as metadata, documentation and intermediate processing software related input/output). "NASA Earth Science Data Preservation Content Specification" (423-SPEC-001) serves as the leading guidelines for preserving the Earth Observing System Data and Information System (EOSDIS) data (Reference 1). We will present an architectural overview of the Goddard Earth Sciences – Data and Information Service Center (GES DISC) preservation system, which is capable of long-term archive of documentation artifacts and other associated digital content. The GES DISC is one of the EOSDIS data centers that have been actively pursuing this preservation specification over the last few years. We designed this system based on Fedora Commons, an open-source physical objects preservation and management software. We will present our concept of operations for user access and details of how our implementation takes into account document access, based on considerations of proprietary or sensitive information. The first mission to make use of the GES DISC preservation system is the High Resolution Dynamics Limb Sounder (HIRDLS) on the Aura spacecraft.

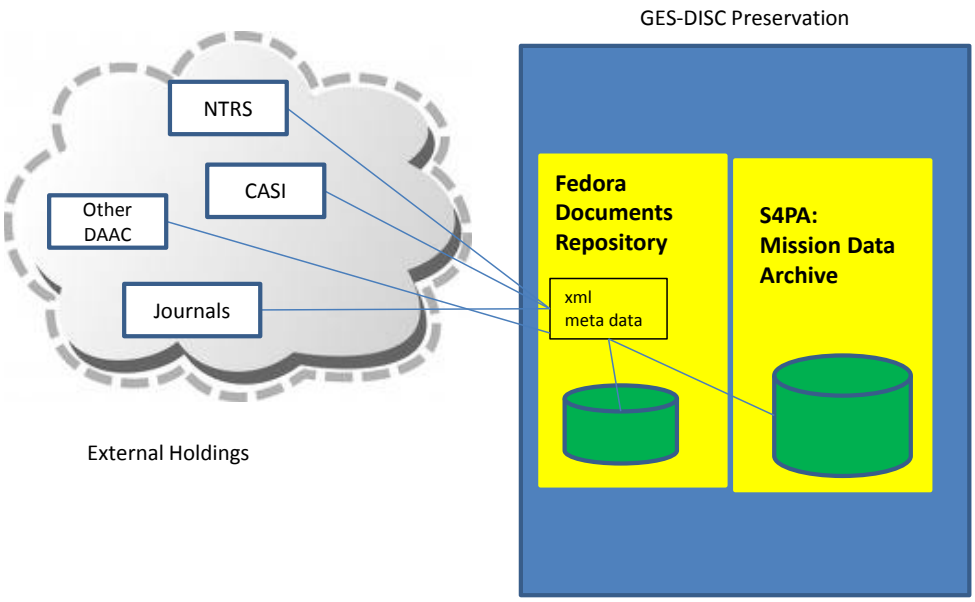

Figure 1 – A pictorial overview of the GES DISC mission data and documentation preservation systems. The physical artifacts could either reside at the GES DISC or be external, for example at the NASA Technical Reports Server (NTRS), the NASA Center for AeroSpace Information (CASI) or any of the Distributed Active Archive Centers (DAAC) or any number of scientific and technical journals.

Data Preservation at GES DISC

The GES DISC began archiving data in the early 1990s, starting with data from the Upper Atmosphere Research Satellite (UARS) and the Total Ozone Mapping Spectrometer (TOMS). With that experience, GES DISC soon established a niche in archiving atmospheric chemistry datasets. It now archives data for three of the instruments on Aura: HIRDLS; Microwave Limb Sounder (MLS); and Ozone Monitoring Instrument (OMI). In addition to these, GES DISC archives precipitation data sets, typified by the Tropical Rainfall Measuring Mission (TRMM) and the forthcoming Global Precipitation Measurement (GPM) mission among others. GES DISC also archives data from the Modern Era Retrospective-analysis for Research and Applications (MERRA). The GES DISC also develops tools and services that enable users to search for, order, download, and visualize NASA Earth science datasets.

The NASA GES DISC recently completed a data preservation campaign for HIRDLS—a 21-channel (6.12 - 17.76 μm) limb scanning infrared radiometer measuring emission from Earth's limb. HIRDLS made measurements of temperature, cloud top pressure, geopotential height, trace constituents (e.g., ozone, nitric acid, chlorofluorocarbons), aerosols, and cirrus clouds from the middle troposphere to the mesosphere. HIRDLS' high-resolution measurements provided new insights into troposphere–stratosphere interactions, chemical reactions in the atmosphere, cyclic weather events, and air pollution impacts.

The GES DISC preservation system consists of both data products as well as associated documentation that pertains to these products. The mission data archive system consists of an in-house developed system called the Simple, Scalable, Script-Based, Science Product Archive (S4PA). The S4PA archives all mission data processing levels from L0 to L3+. The S4PA is further described in Reference 2. To augment the mission data archive system, the GES DISC also implemented a documentation and digital objects preservation system based on the Fedora Commons open source system. We chose Fedora Commons for its flexible and extensible architecture. The Fedora Commons may include actual digital copies of items stored locally at the GES DISC, or point to external objects stored at other NASA centers (DAACs), the NASA Center for AeroSpace Information (CASI), the NASA Technical Reports Server (NTRS), or any of the commercial journal and scientific publications. Metadata associated with the data products link the associated documentation or other reference items to the mission data products. Figure 1 shows a pictorial relationship between the S4PA mission data archive and the Fedora document preservation system.

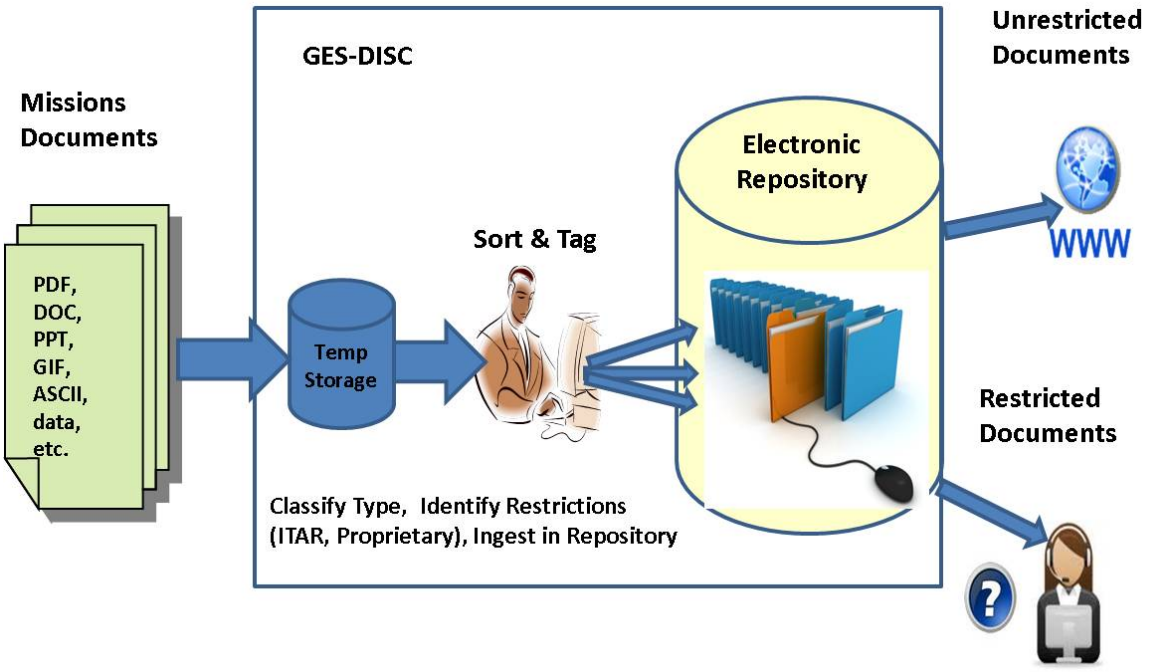

Figure 2- An overview of the physical objects sorting, tagging, storage in archive and distribution system. Restricted documents are currently only available by contacting the GES DISC user services.

The preservation items and documentation slated for preservation for EOSDIS encompass eight different content elements as identified by 423-SPEC-001 and as summarized in Table-1. One task of the long term preservation involves the sorting of the large number of preservation artifacts and determining what is relevant for preservation. After this determination has been made, the next activity involves classifying these items according to one of the eight preservation categories listed in Table 1. Figure 2 provides a pictorial presentation of the process that is typically used to identify, sort and tag the many items slated for long term preservation.

The physical objects are typically unrestricted documents intended for public distribution as they provide detailed information for the data products accessible to the public. These objects cover all aspects of a satellite remote-sensing mission. The preserved data types can include a wide range of content: instrument specifications, prelaunch calibration reports, algorithm input data, algorithm software, data product documentation (i.e., what the data product represents and what it will be used for), data acquired on data validation campaigns, calibration data collected during the mission, and several other types. One of the complexities of this and similar data preservation efforts is that the data are definitely not all numbers and data plots; they could also be text, software code, diagrams, or images, or something else. Occasionally however, some of the preservation artifacts may contain specific proprietary information (such as manufacturer specific information used in fabrication or design of instruments) or information that is restricted for distribution or subject to the US government import and export restriction (the International Traffic in Arms Regulations -ITAR). Those documents are tagged internally in the preservation system as restricted and are only available by directly contacting the GES DISC User Services. The distribution of these documents is limited and subject to verification of compliance to the applicable restrictions.

Preservation Category	Description
Preflight/Pre-Operations Calibration	This element may include instrument specifications, calibration reports, and prelaunch performance measurements.
Science Data Products	This element can include data from the instrument at all processing levels from the Level 0 raw data to Level 3 global and Level 4 model data, as well as metadata required to allow both search and access *for* the data and understanding *of* the data.
Science Data Product	Many different types of information are included under this data preservation

Documentation	element, including the names of science team members, product requirements, data processing history, algorithm history, detailed algorithm descriptions, and data quality assessment.
Mission Data Calibration	There are two main categories intended for preservation here. One category is descriptions of the calibration methods used for the mission, and the second category is the actual calibration data.
Science Data Product Software	Data collected for this element consists of the software (both description of and the actual code) for the generation of the data product. It is desirable to capture as many different software versions corresponding to the corresponding data product releases as possible
Science Data Product Algorithm Input:	Many remote sensing algorithms require other data (ancillary data) as input to calculate a particular data product. This information includes full descriptions of the input data and attributes covering all input data used by the algorithm, including primary sensor data, ancillary data, forward models (e.g. radiative transfer models, spectral line-lists, optical models, or other model that relates sensor observables to geophysical phenomena) and look-up tables.
Science Data Product Validation	Data types that are classified under this element include the data collected on validation campaigns, accuracy reports, characterization and description of the validation process, ongoing calibration and validation results, and methods used to maintain accurate calibration of the instruments collecting the validation data.
Science Data Software Tools	This often-overlooked (or undervalued) element refers to the tools (mostly software but possibly including hardware) required to read and/or display data collected under the other elements. Data can be in many different formats, requiring specific tools to read and use the data. If these tools are not preserved along with the data, having just the data becomes useless.

Table 1- The classification of objects for preservation according to the EOSDIS specification.

GES DISC Document Preservation

The GES DISC Preservation system is built using the Flexible Extensible Digital Object Repository Architecture (Fedora) which is an open source community built software product. Fedora provides for an expandable system built on a powerful digital object model and contains an extensible metadata management capability with easy integration of Web services (SOAP and REST). It also provides for version management of the digital objects and has a relatively simple user interface (Figure 3) and configurable security architecture. Each object slated for preservation is assigned a Persistent ID (PID) as the unique digital object identifier within the system. Fedora Commons also contains an internal database with a XML like schema (FOXML) and scripting capability which enables ingest of large quantities of objects more efficiently.

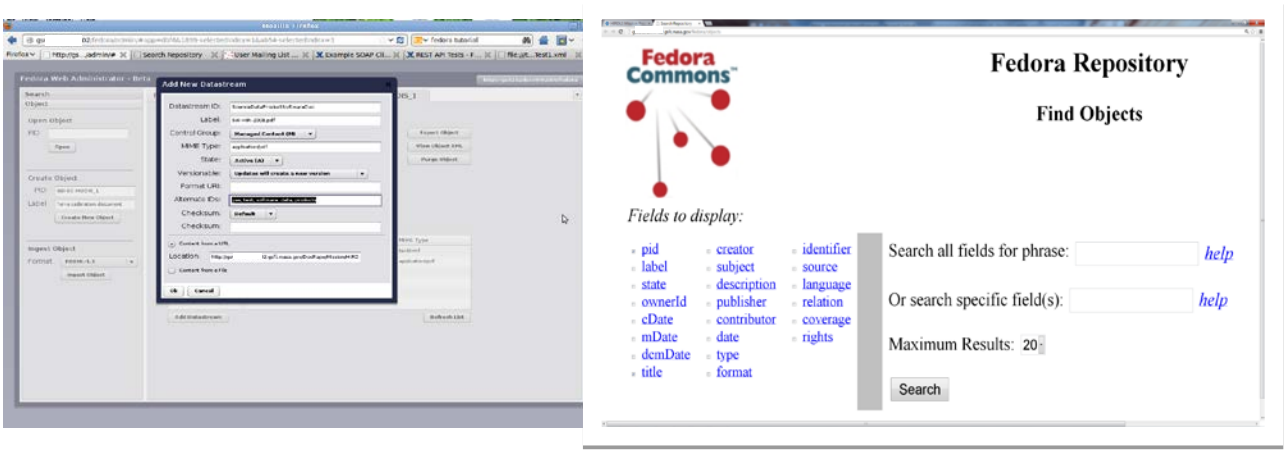

Figure 3- Fedora Commons provides a simple GUI interface for use which provide for an easy administration of the system. In addition the interface supports command line and scripted entry which is more convenient for larger ingest jobs.

The HIRDLS data preservation effort provides access to applicable mission documentation on a simple portal page: ***http://disc.sci.gsfc.nasa.gov/Aura/additional/documentation/hirdls-preservation-documents***. Users can easily access documents or other publications that provide detailed information on the mission, its products, and the algorithms used to process and produce the higher level products. Figure 4 shows the interface that is used by the users to browse and download the publicly available documentation for the HIRDLS mission.

Conclusions and future plans

Because data from NASA's missions are valuable scientific resources, data preservation efforts such as that for HIRDLS are intended to allow scientists to use the data in the future for comparisons with newer instrument datasets, as well as with evolving and new data analysis methods. In turn, this will increase the usefulness of Earth observations from upcoming missions by creating an improved historical comparison capability and a much better characterization of trends in Earth system data records. The results of such research allow insight into the changes affecting Earth's vital ecosystems and the natural support systems on which humanity relies. In the process of setting up the physical objects preservation system for documentation, using an open source system, the GES DISC staff had to self-train on the Fedora Commons system using the publicly available documentation and overcome a large training curve largely by trial and error. Although there is now an expanding community of users, overcoming local configuration and integration issues requires some initial investment of time. Once we had setup a prototype that could be tested and used in a sandbox setting, setting up and ingesting documents into the preservation system became more manageable.

The HIRDLS prototype provided us with the initial experience to setup a local archive for documents which will be the basis for completing the preservation activity for the other missions at the GES DISC. Our plans include extending these services for mission documentations for the Upper Atmosphere Research Satellite (UARS), Total Ozone Mapping Spectrometer (TOMS), Microwave Limb Sounder (MLS), Ozone Monitoring Instrument (OMI), Atmospheric Limb Sounder (AIRS), among others in the near future. One of the challenges still remaining is to setup a registration and a user verification mechanism that could be used to take proper steps and precautions for the distributing ITAR and proprietary documents. This is an ongoing activity as attempt to navigate the requirements and regulations surrounding both the ITAR and proprietary documents as well as the privacy issues with implementing a user registration system.

Figure 4- The portal page that is used by users to access and download the publicly available documentation for HIRDLS.

References:

1. 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***

2. 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