The Lifecycle of NASA's Earth Science Enterprise Data Resources

Kenneth R. McDonald[1], Richard A. McKinney[2], Timothy B. Smith[2], Robert Rank[3]

[1]NASA/Goddard Space Flight Center
Mail Code 423, Greenbelt MD 20771 USA
(ken.mcdonald@nasa.gov)

[2]USGS/EROS Data Center/SAIC
Sioux Falls, SD 57198 USA
(mckinney@usgs.gov), (tbsmith@usgs.gov)

[3]NOAA/NESDIS
5627 Allentown Road #100, Suitland MD 20746 USA
(Robert.Rank@noaa.gov)

Introduction

A major endeavor of NASA's Earth Science Enterprise (ESE) is to acquire, process, archive and distribute data from Earth observing satellites in support of a broad set of science research and applications in the U. S. and abroad. NASA policy directives specifically call for the agency to collect, announce, disseminate and archive all scientific and technical data resulting from NASA and NASA-funded research. During the active life of the satellite missions, while the data products are being created, validated and refined, a number of NASA organizations have the responsibility for data and information system functions. Following the completion of the missions, the responsibility for the long-term stewardship of the ocean and atmospheric, and land process data products transitions to the National Oceanic and Atmospheric Administration (NOAA) and the U.S. Geological Survey (USGS), respectively.

Ensuring that long-term satellite data be preserved to support global climate change studies and other research topics and applications presents some major challenges to NASA and its partners. Over the last several years, with the launch and operation of the EOS satellites and the acquisition and production of an unprecedented volume of Earth science data, the importance of addressing these challenges has been elevated. The lifecycle of NASA's Earth science data has been the subject of several agency and interagency studies and reports and has implications and effects on agency charters, policies and budgets and on their data system's requirements, implementation plans and schedules. While much remains to be done, considerable progress has been made in understanding and addressing the data lifecycle issues.

Lifecycle of NASA's Earth Science Data

One way to view the life cycle of remotely sensed Earth science data is to identify the major functional areas or facilities where the data resides and is transformed during its lifetime. The breakdown that we will use includes mission planning, development and operations, science data production, active archive and long-term archive. While the physical configuration and organizational responsibilities of these facilities can vary from mission to mission, all of the functions are required. Following is an overview of the responsibilities of each area.

Mission Planning, Development and Operations

The mission planning and development functions have traditionally focused on the lifetime of the mission, from conception, to launch, through data reception and production, to and including insertion into the active archive. It includes the planning, design, development, funding, and any other aspects of preparation. After the pre-launch, launch and a commissioning period, the mission will transition into its operational phase. Typically, a mission's life span is only a few years after which the mission ends to allow for a final reprocessing of the data. The mission operations functions include both the command and control of the spacecraft and the instruments and the acquisition and low level processing of all of the data and its delivery to the science processing and active archive facilities. In addition to the science data streams, all of the orbit, attitude and engineering data are also delivered. Mission operations also include processing the raw data into a form usable by the science data users (Level 0).

Science Product Generation Responsibilities

The science product generation receives the Level 0 data and applies various algorithms to produce high-level products in support of land, oceans, atmosphere, hydrology, and other scientific studies as well as a wide range of applications (e.g. agriculture, disaster management, environmental health) that can use the science products. The generated products are distributed primarily to the active archive centers but also can be distributed directly to users.

Active Archive Responsibilities

The active archive is the system for processing support, archiving, documenting, and distributing the data and information for the life of the mission. The active archive's major role is to serve the day-to-day needs of the mission. At this stage in its life cycle, the data are being regularly processed and reprocessed using on-going data validation and quality assessment results while also being provided to the broader community of science and applications users. The role of the Active Archive typically involves three components: scientific stewardship, customer service, and IT infrastructure requirements. The active archive supports the routine operations of data acquisition, data processing, data re-processing, and data staging for archive products requested for real-time and historical data from the mission. The active archive keeps track of the various processing algorithms that may be involved in the project and the appropriate metadata and browse links within the system. It is the system hub for the mission's product generation system and it is the fundamental source of information and data for Long Term Archive (LTA) transfer activities.

LTA Responsibilities

The LTA is the archive in which stewardship of data, products, information, and documentation is held on a permanent basis or until the data is considered to be of no value. The stewardship entails preservation, maintenance, and access of the data to ensure integrity and quality of the data as the documentation indicates. The LTA is generally populated from the active archive.

A Call for Action

NASA Policy Directives specifically call for the agency to "collect, announce, disseminate, and archive" all scientific and technical data resulting from NASA and NASA-funded research. Various scientific and policy-making groups have reviewed and defined the requirements for essential data systems and services needed to ensure an LTA for satellite data record in support of climate research. Most significantly, the U.S. Global Change Research Program (USGCRP) evaluated the purpose of an LTA program for Earth observation data and derived products, lessons learned from current and past experience, and the vision, guiding principles, and essential functions necessary for the success of such a program. The USGCRP concluded [1] that a national LTA program must:

- Enable and facilitate the best possible science and highest-quality assessment for making policy and business decisions
- Document the Earth system variability and change on global, regional and local scales, building and maintaining a high quality base of data and information and establishing the best possible historical perspective critical to effective analysis and prediction
- Ensure archive holdings, facilities, and services are actively promoted and made readily available to the maximum number of users
- Enable and facilitate future research.

A report from the National Research Council (NRC) Committee on Earth Studies [2] has also contributed to the definition of requirements for essential data system services need ensure a long-term satellite data record in support of climate research. The CES report defined following principles to help ensure the preservation of the climate record:

- Accessible and policy-relevant environmental information must be a well-maintained part of our national scientific infrastructure.
- The federal government should 1) provide long-term data stewardship, 2) certify open, flexible standards, and 3) ensure open access to data.
- Because the analysis of long-term data sets must be supported in an environment of changing technical capability and user requirements, any data system should focus on simplicity and endurance.

- Adaptability and flexibility are essential for any information system if it is to survive in a world of rapidly changing technical capabilities and science requirements.
- Experience with actual data and actual users can be acquired by starting to build small end-to-end systems early in the process.
- Multiple sources of data and services are needed to support development of climate data records.
- Science involvement is essential at all stages of development and implementation.

More recently, the Earth Observing System Science Working Group on Data [3] offered the following recommendations relevant to the Earth Science Data Lifecycle:

- Science stewardship: NASA, in conjunction with NOAA and USGS, should determine what it is and how it works, at the various stages in the life of the LTA, who is responsible for it, and who funds it.
- The transfer of data sets to the LTA should begin as soon as feasible.
- Requirements of the individual instruments for the LTA should be communicated knowledgably to NOAA and USGS.
- Coordinated schedules and goals for working with NOAA and USGS to effect the initial LTA agreements, planning, and transfer.
- Instrument teams and DAACs should participate in advisory panels and committees within NOAA/USGS to specify and administer the LTA.
- EOS teams should identify and recommend LTA levels of service that should be provided by NOAA and the USGS.
- Provision is required for NASA investigators to have ongoing access to the data sets under similar conditions. to the present.

Data Lifecycle Studies and Analyses

In early 2002, NASA Earth Science Enterprise (ESE) initiated a series of studies under its Strategic Evolution of ESE Data Systems (SEEDS) activity. The purpose of the studies was to address particular topics that would help the enterprise evolve its data systems capabilities in the future. One of the topics identified was the long-term preservation of the NASA ESE data collections. While the motivation for this study arose in response to concerns with the long-term preservation of the science data and it is true that many of the associated challenges are related to the transfer of data from the active archives to the LTAs, it became clear that effectively satisfying the LTA requirements levies new or raises the importance of existing requirements on the other functional areas. Therefore, the study was broadened to be the SEEDS Data Lifecycle (DLC) Study.

The approach taken by the SEEDS DLC study team was first to thoroughly review all of the documentation relevant to data preservation that had been produced by previous study teams and review committees. This also included the review of mission and project requirements documents and relevant interagency Memoranda of Understanding. Then the study team generated a preliminary set of responsibilities for each of the data lifecycle functional areas associated recommendations. This draft study report formed the basis of the team's interactions with the participants of the several SEEDS Community Workshops in 2002-2003. The participants in these workshops represented data producers and users from Earth science research and applications communities, data systems and especially data management specialists and members of NASA's partner agencies and organizations. The input from the community interactions was incorporated into the DLC studies findings and recommendations (http://eos.nasa.gov/seeds/).

The set of recommended responsibilities that were identified for each functional area to ensure the long-term preservation of NASA Earth science data is too lengthy to repeat in this paper. In general, they provided detailed guidelines with respect to activities of planning, documentation and communication and the interactions between them. With the diverse set of organizations that are responsible for data over its lifetime and lengthy timeframes involved, it is essential that a comprehensive plan be developed and documented so that all parties are aware of their respective responsibilities. As data and its associated software move through its lifecycle, it is essential that all associated documentation be produced and conveyed with it to ensure that the data is not only preserved but also usable and well understood. Finally, all of the organizations responsible for the data must communicate among themselves and their user communities to ensure a smooth transition of the data and a set of data services responsive to users needs. These interactions must also be documented in the form of interface control documents among functional elements and user requirements and system operational concepts.

In addition to the specific responsibilities delineated for each functional area, the following general policy recommendations were also outlined:

- There should be an archive defined for each data product. Identify the active and LTA at the beginning of the mission, noting that for some products the active and LTA may be the same.
- "Data-buys" or any proprietary mission needs a time period to be defined after which the data becomes the property of NASA. Intellectual property rights issues need to be addressed.
- During the proposal process, the new project needs to commit/agree to a plan for a data lifecycle strategy for the project data.
- A process should be established for the science assessment of the long-term potential of data and data products to assist in making decisions on data stewardship in a resource-constrained environment.
- All archive data collections should be complete, including the archiving of the required ancillary data, project and data set documentation, and the science production software.
- Once a physical transfer of any data has occurred and been formally accepted by the archiving site, these data become the responsibility of the accepting party.
- Data may be transferred to other archives such as the National Archives and Record Administration (NARA) when deemed appropriate.
- Data should be available throughout its life cycle without loss or degradation in quality.
- Throughout a product's life cycle, a point-of-contact should be provided that could answer questions about the data or use of the data.
- Determine the relevance of Data Quality Act (67 FR 8452) and resolve issues related to watermark, provenance, reproducibility of data, peer review, integrity of data, and supporting information.

A final recommendation that was principally developed as a result of the community interactions was that NASA needs to form and maintain a Data Lifecycle Working Group with broad representation from data producers and users and data management specialists from across government agencies, academia and industry. It is clear that with the timescales that are being considered, even when the aforementioned data lifecycle plans are developed and base-lined, they will likely need revision as agencies' priorities and budgets as well as users' needs and expectations change. Some sort of working group to assist in the studies and analyses to support NASA and its partner agencies in addressing those changes will be essential.

Current Data Lifecycle Activities

NASA

Over the past fifteen years, NASA has been designing, developing, operating and evolving the Earth Observing System Data and Information System (EOSDIS) to acquire, process and archive its Earth science data resources. Since 1994, an operational prototype of EOSDIS (Version 0) largely based on existing components has been handling the heritage data collections of the agency. The data are archived at nine Distributed Active Archive Centers (DAACs) across the USA that were selected on the basis of their existing expertise in producing and managing Earth science data for particular science disciplines. The EOSDIS Core System (ECS) was installed at four of the DAACs in 1999 to support the Landsat 7 and Terra missions, the latter being the first of the EOS platforms. Since then, EOSDIS has been performing the mission operations, science processing and active archive functions for those and most of the subsequent missions of the EOS era (a select few missions have performed some or all of those functions outside of EOSDIS).

Today, EOSDIS is fully operational and handling an unprecedented amount of data from numerous instruments on multiple missions. The majority of that data is archived in the ECS DAACs, which with new data and reprocessing campaigns, are ingesting approximately 4.5 TB per day and have a collective near-line and on-line archive that is greater than 3 PB. The system also has a large and active user base. Last year, all nine of the DAACs satisfied over 220,000 users orders for data and distributed in excess of 440 TB. In addition to distributing the data, the DAACs also have a full contingent of trained user support personnel to assist in the acquisition and use of the data.

Over the course of its development EOSDIS architecture has evolved, but the major functions and the associated requirements have remained intact. As an example, in its original design, the EOS science data production was to be done by the ECS at the DAACs using science-processing software produced by the instrument teams. Today, most of this processing is done by Science Investigator-led Processing Systems (SIPS), usually co-located with the instrument teams, and the products are delivered to the DAACs for archive and distribution. However, in addition to delivering the

science data, the teams are also responsible for delivering software and documentation that are essential in assuring the long-term preservation of the data.

Understandably, the focus of NASA to data has been on getting a stable, operational system in place to support the mission operations, science data production and active archive functions for the EOS missions and that has been accomplished. However, the requirement for transitioning data to its long-term archive is the life of the mission plus four years, which in the case of Terra would be about 2010. To meet that schedule, a general guideline has been to begin to transfer data three years after its acquisition. For a number of reasons this guideline has not been met, but the issue has risen in priority and NASA is actively working with its partners to accelerate the process.

An overview of the process that is being defined to govern the transfer of data from the active archive to the LTA and which also identifies the issues that need to be addressed is illustrated in Fig. 1. The key to this transition is to develop a detailed plan for the transition. To do this, existing plans and agreements must be reviewed and updated according to current agency requirements, priorities and budgets. Part of the trade space in making these decisions is determining the appropriate level of service for each data product with the understanding that this will likely vary from product to product. It will also be necessary for a process to be defined to make the decisions on the products that is transparent and documented and gets appropriate input from each of the involved agencies and the communities that they serve. This process must also work within the existing management processes that have been established by the agencies. NASA and its operational partners, NOAA and the USGS are working together to address these issues and some initial plans and transitions are underway. However, work on the transition of EOS data is still in the earliest stages.

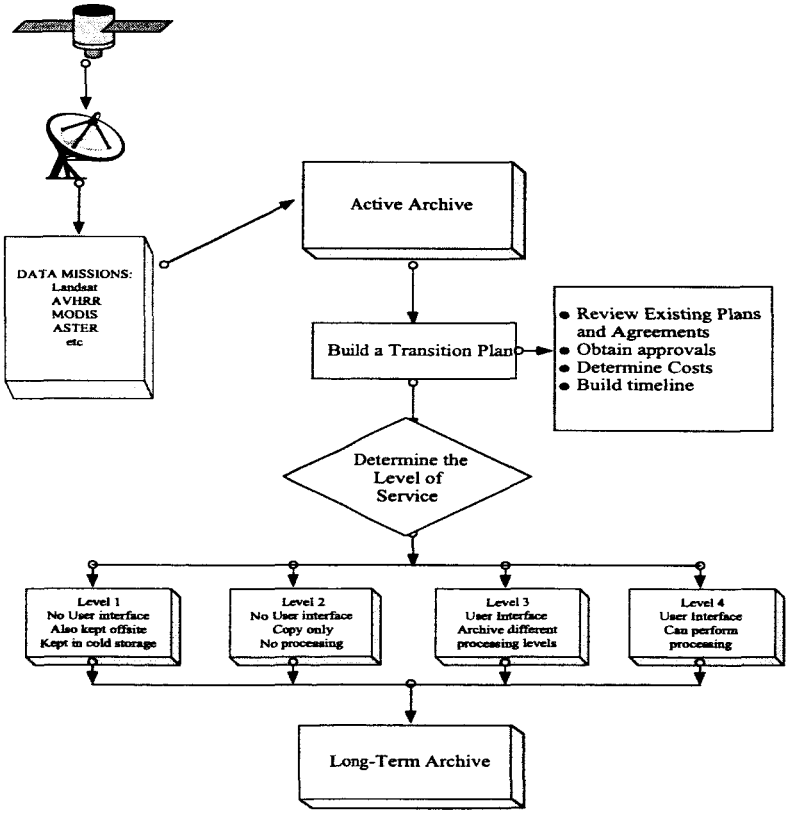

Figure 1. Planning process for transferring data to the LTA

USGS

The USGS has a mandate from the Department of Interior (Land Remote Sensing Policy Act of 1992; Public Law 102-555, 15 U.S. Congress 5601) to establish a public-domain archive of satellite data of the Earth's land surface. The resulting National Satellite Land Remote Sensing Data Archive (NSLRSDA), managed by the USGS EROS Data

Center, includes land and coastal observation data captured by satellites and augmented with data from other government, commercial, and international sources. The major focus of NSLRSDA is to manage and preserve those data so that they remain accessible to a broad range of users over time. This entails employing state-of-the-art transcription and archiving technologies, transferring the data to new media as needed, investigating and implementing advanced media and data storage systems for long-term data preservation, and exploring new ways to make data accessible. A fifteen member Advisory Committee, in accordance with provisions established by the Federal Advisory Committee Act, has also been established to provide advice and council on guidelines and rules relating to NSLRSDA data acceptance, maintenance, preservation, and access management policies The Advisory Committee meets twice a year (http://edc.usgs.gov/archive/nslrsda/advisory/charter7.html).

The USGS EROS Data Center utilizes National Archives and Records Administration (NARA) endorsed processes for NSLRSDA and other USGS LTA collections to recommend which data sets are kept or what data may be delivered to NARA in accordance with records management practices and schedules.

The Land Processes (LP) Distributed Active Archive Center (DAAC) was established as an active archive at the USGS's EDC as part of NASA's Earth Observing System (EOS) Data and Information System (EOSDIS) initiative to process, archive, and distribute land-related data collected by EOS sensors, thereby promoting the interdisciplinary study and understanding of the integrated Earth system. The LP DAAC, by design, is a temporary archive until the end of missions, at which time the data is migrated to a long-term archive.

Prior to the launch of Landsat 7, Terra and Aqua, the LP DAAC distributed a variety of prototype and precursor data and products commonly known as Version 0 Data and Products. These data and products have transitioned from the LP DAAC to the USGS long-term archive. The Version 0 products include:

- Advanced Solid-state Array Spectrometer (ASAS)
- Advanced Very High-Resolution Radiometer (AVHRR)
- Global Composites
- Global Normalized Difference Vegetation Index (NDVI) CD
- Orbital Segments
- Aircraft Scanners
- Global Land Cover Characterization (GLCC)
- Global Land Cover Test Sites (GLCTS)
- GTOPO30 and Hydro 1k
- Landsat 7
- NASA Landsat Data Collection (NLDC) MSS/TM
- North American Landscape Characterization (NALC)
- Spaceborne Imaging Radar-C (SIR-C)
- SIR-C Educational CD

In order to maintain consistency and cohesiveness for the science users, the Version 0 data is still searchable and orderable through the EOS Data Gateway (EDG), formerly known as the Version 0 Information Management System (V0 IMS). As the priority dictates, the Version 0 data will be moved from EDG access to an appropriate USGS client for user access.

NOAA

The National Environmental Satellite, Data, and Information Service (NESDIS) is responsible for the collection, archiving, and dissemination of environmental data collected by a variety of *in situ* and remote sensing observing systems operated by the National Oceanic and Atmosphere Administration (NOAA) and by a number of its partners [e.g., National Aeronautics and Space Administration (NASA)]. To prepare for large increases in its data holdings, NESDIS initiated the planning and development for a Comprehensive Large Array-data Stewardship System (CLASS) that provides archive and access services for these data. CLASS must be able to handle the data flow from current satellite-based systems [e.g., Polar-orbiting Operational Environmental Satellite (POES), Geo-stationary Operational Environmental Satellite (GOES), and Defense Meteorological Satellite Program (DMSP)] ground-based systems [e.g., Next Generation Weather Radar (NEXRAD)] and *in situ* systems [e.g. Automated Surface Observing System (ASOS)]. It must also be structured to handle the large increases in data that will come from planned satellite launches [e.g., Meteorological Operational satellites (Metop), National Polar-orbiting Operational Environmental Satellite System (NPOESS), NPOESS Preparatory Project (NPP), and Earth Observing System (EOS) satellites].

CLASS will be operational at two locations, the Office of Satellite Data Processing and Distribution (OSDPD) facility at Suitland, MD and the National Climatic Data Center (NCDC) facility at Asheville, NC. Each facility has similar hardware and identical software, is capable of assuming the overall CLASS load at any given time, and is operational at all times. During normal operations, both facilities share the processing load. The most important difference between the two CLASS facilities is the tape robotic systems and associated Hierarchical Storage Manager (HSM). The interface between CLASS and NCDC's tape robotic systems is defined and described in the CLASS-NCDC Archive and Recall System (NARS) Interface Control Document (ICD).

CLASS provides life cycle capabilities for archiving, distribution, preservation, and operation, such that all approved campaign array-data may be preserved as defined by existing National Archives & Records Administration (NARA) and NESDIS archive policies, distributed as requested to customers, and available for disaster recovery. The scope of these capabilities includes the ability to scale system functionality to continuous growth in campaigns and the preservation needs of the data.

Multiple departments of NOAA contribute to the CLASS design process. The technical management team includes members from the Office of System Development (OSD), National Climatic Data Center (NCDC) and the National Geophysical Data Center (NGDC).

The CLASS Project Management Team (CPMT) must define the detailed policies for lifecycle management. The general organizational objectives are as follows:

- Development of software is geographically distributed, but controlled by a centralized Configuration Control Board.
- Installation of the system will be at two sites, leveraging existing facilities.
- Operations for archive and distribution of data sets is fully automatic and depends upon electronic transfer for source data.
- Operators are able to conduct supporting activities remotely, using a secure interface. Examples of supporting activities are customer help, reconfiguration of hardware for load balancing, and data migration.
- Operators coordinate with the housing facility, but system administration of hardware is the responsibility of the housing facility site management.
- Unless a disaster contingency plan is in progress, each housing facility must have one operational system.

The policies cover the following four categories of capabilities and responsibilities:

Archiving – take in data, catalog, move to archive

Primary data sources are limited to NOAA, NASA, and Department of Defense (DoD) institutions that require long-term preservation and distribution of non-classified data.

The Archive, Access and Distribution System flexibility must be sufficiently robust to cover all the needs of the seven sets of large-array data. Other sources of data (e.g., in-situ) may be added as well to the system, by application to the Data Archive Board and the Board's acceptance.

Any data set must conform to certain requirements on content, format, and handling as defined by agreement between NESDIS and the data provider to be acceptable to the system, but some changes in the system may be required to accommodate each new data set. Based upon direction from the Data Archive Board and/or CLASS Oversight Group, the CPMT initiates analyses and conducts reviews to assess the compatibility of each new data set with the Archive, Access and Distribution System, and determines the system changes required to enable the acceptance of this data set.

Preservation – backup, duplication, migration

The System provides the ability to preserve data from loss. Minimally, preservation requires that two copies of a data be kept at all times. A minimum of 50 miles must separate the storage locations of those copies.

Distribution – Internet-interface, user access, order fulfillment

Automatic distribution of electronically transferable data sets is the preferred method of delivery; the system will be able to provide offline delivery of data. Pricing is not the responsibility of the Archive, Access and Distribution System. The system supports an interface to an Order Management System (OMS) and provides necessary and sufficient information to the OMS so that the OMS may calculate pricing and carry out charging activities.

Operation – people managing the system, policies

The management team publishes and enforces policies for the operations of the resources in normal and contingency modes of operation. The team provides plans to accommodate the continuous growth of data sets and types of data sets that the system must support.

Concluding Remarks

It is clear that the members and advisory committees of the Earth science research and applications communities view NASA's Earth science data as a national resource and emphatically state that it must be preserved and made accessible to support critical science research which may have significant impacts on important policy decisions. They have defined the need for a national infrastructure for LTA of Earth observation data and have recommended that the LTA address the issues of **setting priorities** for the LTA, identifying the **data content** of the LTA (including documentation required for long term use of the data), and specifying the **stewardship and data services** to be provided by the LTA. NASA and its operational partners, NOAA and the USGS, have been working together to follow this guidance.

While concerns over the long-term preservation of data put the focus on the LTA, the studies by the agencies have recognized that this is truly a data lifecycle issue. Plans for the long-term preservation of the data should be addressed during mission formulation and reviewed throughout the data lifecycle to ensure that facilities and budgets to maintain the data are available when they are needed. Also, for the data to be continually useful, complete documentation must be generated and maintained with the data as it transitions over its lifetime. This is sometimes difficult to recognize in the earlier stages of the mission when the science teams and user support specialists of the active archives are available to share their expertise, but it is critical when those teams have disbanded and the data is in the LTA.

The three agencies have been working together to specify preliminary requirements, guidelines and plans for the transition of data from the NASA active archives to the LTA facilities. However, changes in agency directives and priorities, changes in the needs of the communities they serve, and technological advances will always be occurring and necessitate periodic reevaluations of the plans. The greatest challenges will be the definition of the processes that need to be defined to make final decisions related to these transfers, who needs to have a voice in these processes and how these processes work in conjunction with established working groups and procedures that currently govern the work of the three agencies. Each agency's actions are constrained by Administration and Congressional direction, advisory group guidance, priority, policy, and funding. It will take some considerable effort to define a new process that can be compatible with these existing mechanisms.

References

[1] "Global Change Science Requirements for Long-Term Archiving", Report of the Workshop, Oct 28-30, 1998, U.S. Global Change Research Program [USGCRP] Program Office, March 1999.
[2] "Ensuring the Climate Record from the NPP and NPOESS Meteorological Satellites", National Research Council, Committee on Earth Studies [NAS-CES] 2000).
[3] "Workshop on Long-term Archiving of EOS Data", January 29-30, 2002, EOS Science Working Group on Data, unpublished.