Preserving and Sustaining NASA Data: Quirks and Trials

Ed Grayzeck

NASA Goddard Space Flight Center

Head, National Space Science Data Center

Program Manager, Planetary Data System

Sept 2008

NSSDC Evolution

- Archive for 40+ years of NASA space and earth science data with the latter split off in the early 1990s
- Established as permanent archive (2002) for SS missions, Active Archives, Scientific Archive Research Centers, Virtaul Observatories, Resident Archives
- DATA SOURCES
  - 545 spacecraft
  - 1546experiments
  - 4410 distinct data collections
- ARCHIVE HOLDINGS
  - 55 TB digital data; earliest 1958, Explorer 4
  - 50,000 media – DLTs, tapes, CDs, DVDs [includes Earth Science tapes]
  - 30,000 tape media – 7&9 trk [working with GSFC/Earth Science to restore]
  - analog holdings (e.g. 700,000 photos)
- ARCHIVE FORMAT
  - Archive Information Packages (7.4 TB so far) with extracted attributes to make them media & platform independent following the OAIS functional model

NSSDC Today

NSSDC provides

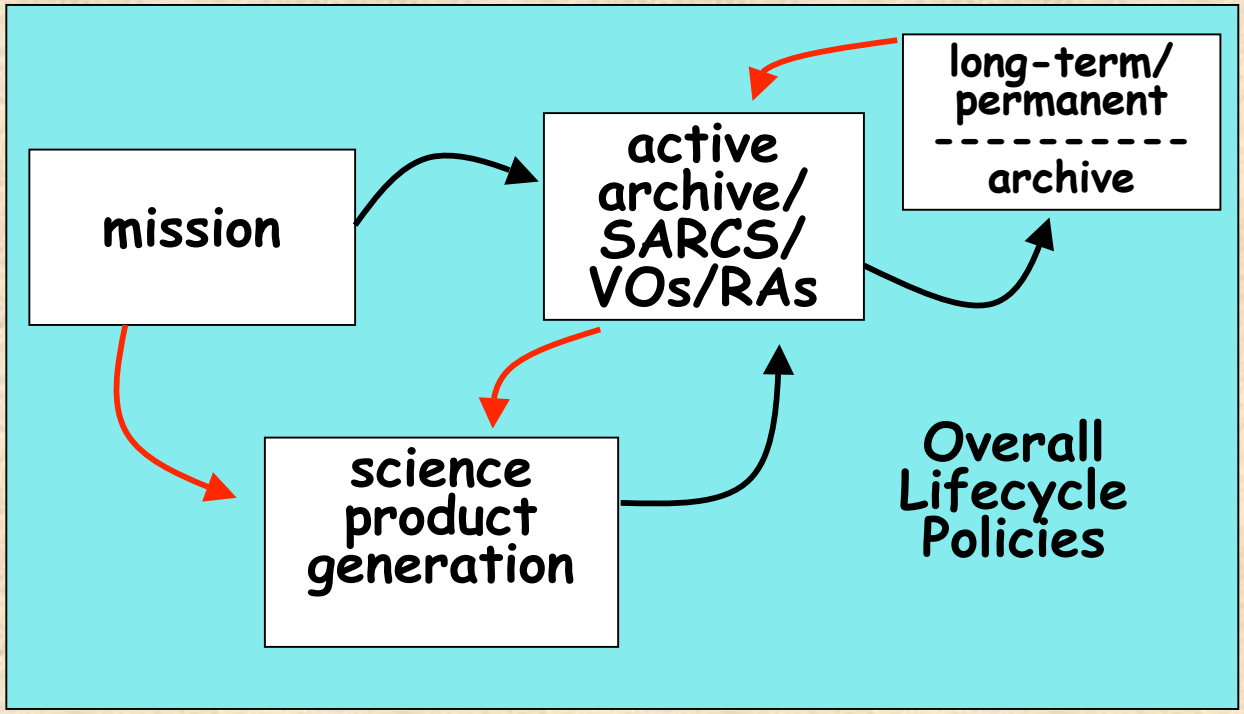
- **A permanent archive writing Archival Information Packets (AIP)**
- **An "active archive" of science data from past missions not in AAs, SARCs**
- **A distributed information framework (for active missions that do not fall under AAs or SARCs)**

- **The ADEC has 4 "SARCs", organized by wavelength with other entities such as NSSDC, CDS, and missions (SWIFT); MSC is new.**

  Data held at SARCs considered long term; NSSDC as backup
- **Continuation work of PDS with international partners, e.g. ESA PSA**
  - **Has 5 science and three support nodes as long term**
  - **Creation of Data Nodes (currently 4) to streamline interaction with missions**
  - **MOU PDS-NSSDC partnership to provide second site archive**
- **Heliophysics has SDAC and SECAA but missions interact directly**
  - ***NSSDC is the primary archive with distribution of legacy data***
  - ***Heliophysics Data and Model Consortium (HDMC) formed June 2008 to coordinate Virtual Observatories and Resident Archives***

Data Life Cycle

A simple model showing the four major lifecycle entities within a context of an overall set of guiding policies



Some of these functions may be grouped together in any given mission or project

Life Cycle Background

- Astrophysics missions often work through existing data centers to host the data archiving and distribution, and the data quality relies on the process specified in the Project Data Management Plan (PDMP), which may be an extension of the data center review. The Astrophysics Data centers Executive Council (ADEC) coordinates
- Planetary missions are usually led by a team with a Principal Investigator (PI) and a large number of Co-Investigators and the data may be hosted by a discipline node, subnode, or data node of the Planetary Data System (PDS). An NRA requires the data be put into the PDS following the outline in the Proposers Archive Guide which requires both a Data Management Plan and Archive Plan.
- Heliophysics has PI led missions with many Co-Is widely distributed and data flow is specified in a PDMP
  - The data may be hosted at a central location or be distributed amongst a group of institutions
- Space Science data is diverse from many sources and may be duplicative

NSSDC Services

NSSDC Archival Storage Service Levels	
Permanent Archive: AIPs	Preservation of digital data in Archival Information Packages delivered by a data producer or created at NSSDC. AIPs are re -written to new media within six years. Data is disseminated by NSSDC if not available thro ugh an active archive or per Memorandum of Understanding (MOU) .
Permanent Archive: non-AIP digital data	Preservation of non -packaged data on various media types. Data will eventually be migrated from legacy media to AIPs. Data is disseminated by NSSDC if not a vailable through an active archive or per MOU.
Second Archive	Storage of digital data on distributable media that is also held by another archive. No media refreshment is performed. NSSDC may disseminate the data if authorized to do so by the primary archive as per MOU.
Backup	Storage of digital datas to support another archive's contingency plan per MOU. Data will not be disseminated by NSSDC.
Analog Archive:	Preservation of analog data on a variety of media with selected refreshment and selected digitization. Selected retention of original analog data after digitization. Data are copied and disseminated by NSSDC..

Data Stewardship

- Data stewardship at NSSDC is taken to mean that NSSDC is active in acquiring and preserving the bits, and usually the information content, of data derived from, or supporting, NASA's space science missions and related research. It attempts to ensure that the information content remains effectively useful to the evolving needs of space science data users. To this end, it has defined three categories of digital service:
  - Permanent Archive: Long-term curation of digital data
    ****
  - Content information preserved, not necessarily in the original format
  - Data may be repackaged and/or transformed to maintain accessibility and usability
  - Data access may be restricted per MOU
  - Data submitted in AIP (preferred) or non-AIP form
    ****
  - Second Archive: Preservation of data also held by another archive
  - Backup Archive: Storage of digital data in support of another archive's contingency plans

Streamlining Ingest:Approach

- Following NSSDC standard, 2 phases
  - Preliminary Phase: Initial interaction with data provider to reach preliminary agreement to proceed
  - Formal Phase: Establish sufficient agreement with data provider, including supporting data documentation, to set up data 'pipeline'
    - Complexity depends on archival service and extent of data deliveries over time
- Work with NSSDC curation scientists, operations, and technologists to develop agreed set of attributes need to support the two phases
- Work with sister organization, Space Physics Data Facility, as friendly data providers to validate approaches
- Work with Planetary Data System in developing and using remote, configurable, software installation for actual data transfers

LESSONS LEARNED: Common Methods

- Help scientists locate data required for a given study.
- Provide scientists with access to those data.
- Assure that those data are useable.****
- Preserve the data forever.*
- Aid scientists in using the data

Retention of datasets

- Definitive datasets (and derived datasets with most science potential for missions where a definitive dataset does not exist) are archived with current standards and retained indefinitely
  - Ensuring continuity requires periodic renewal cycles
  - NASA HQ can approve to release a dataset rather than renew if there is a financial justification and upon soliciting the consultation of the representative community of scientists
- Other datasets
  - Data from NASA's in-flight laboratory-like experiments will be retained according to the same principle as flight data
  - Pre-definitive data from a given mission will be released six months after project personnel certify that the definitive datasets created there from faithfully replicate their content
  - Derived datasets will be retained indefinitely as long as they are scientifically useful
- How to determine if data are scientifically useable ****
  - NSSDC formed Assessment Teams from the community in each scientific area to set priority on retention; concluded Aug 2008

NASA Unique Lunar Data

- Goal is to have digital data in older formats restored to be readily accessible for use by scientists and engineers.
  - Data will be available online and archived at NSSDC and eventually with PDS.
- Only direct data on the lunar surface environment, will allow improvements in design and scope of future lunar instruments. Data sets recommended for restoration by external reviewers
- Special request - Apollo (ALSEP) microfilm scanned (LRO)

LUNAR DATA PROJECT / DATA NODE PROGRESS

- Cold Cathode Ion Gage - Completed, submitted to PDS (2 data collections)
- Solar Wind Spectrometer - Review completed, working off liens for PDS
- X-Ray Spectrometer - Review completed, working off liens for PDS
- Dust Detector - 26 microfilm reels scanned, OCR work in progress
- CPLEE - Magnetic tapes read, CDF's being created for PDS conversion
- Soil Mechanics - Microfilm scanned, detailed metadata being created
- Suprathermal Ion Detector - LASER proposal funded
- Infrared Radiometer - Magnetic tapes read for W. Mendell (Johnson SC)
- Lunar Atmosphere Composition Experiment - LASER proposal funded
- Heat Flow - Magnetic tapes read and converted, other data pending
- Magnetic Fields - LASER proposal funded (UCLA group)

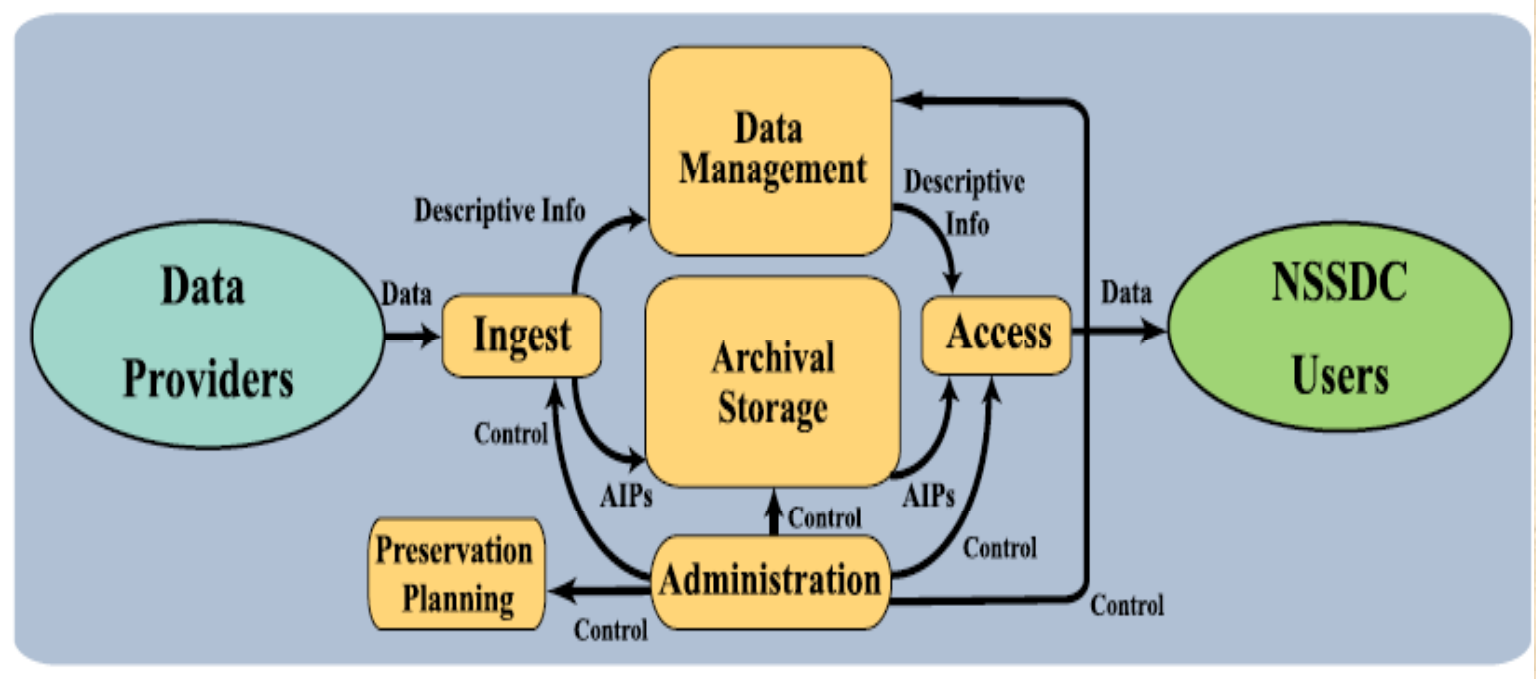
Resident Archive Initiative

Current holdings in solar/space phsyics data including 684 missions. There are new missions for which NSSDC is named as the permanent archive. To streamline the archiving process from mission to NSSDC, there is an initiative to form RA(s) which have the following functions consistent with the Heliophysics Data Environment Management Policy:

1.Ensure that the mission data are served to the general space and solar physics community in an efficient and scientifically useful interoperable manner consistent with the community data environment standards
2.Maintain the integrity of the data by safeguarding against data loss which could be effected by the use of mirror sites
3. Provide expert assistance with data issues
4.Document the data (including mission and PI information) as required to maintain independent usability
5.Obtain community feedback to ensure success
6.Make sure the data will be archived after the RA is no longer needed (e.g., transferred to another RA,or NSSDC

First RA was the Sampex Data Center at ACE Science Center

NSSDC: Functions



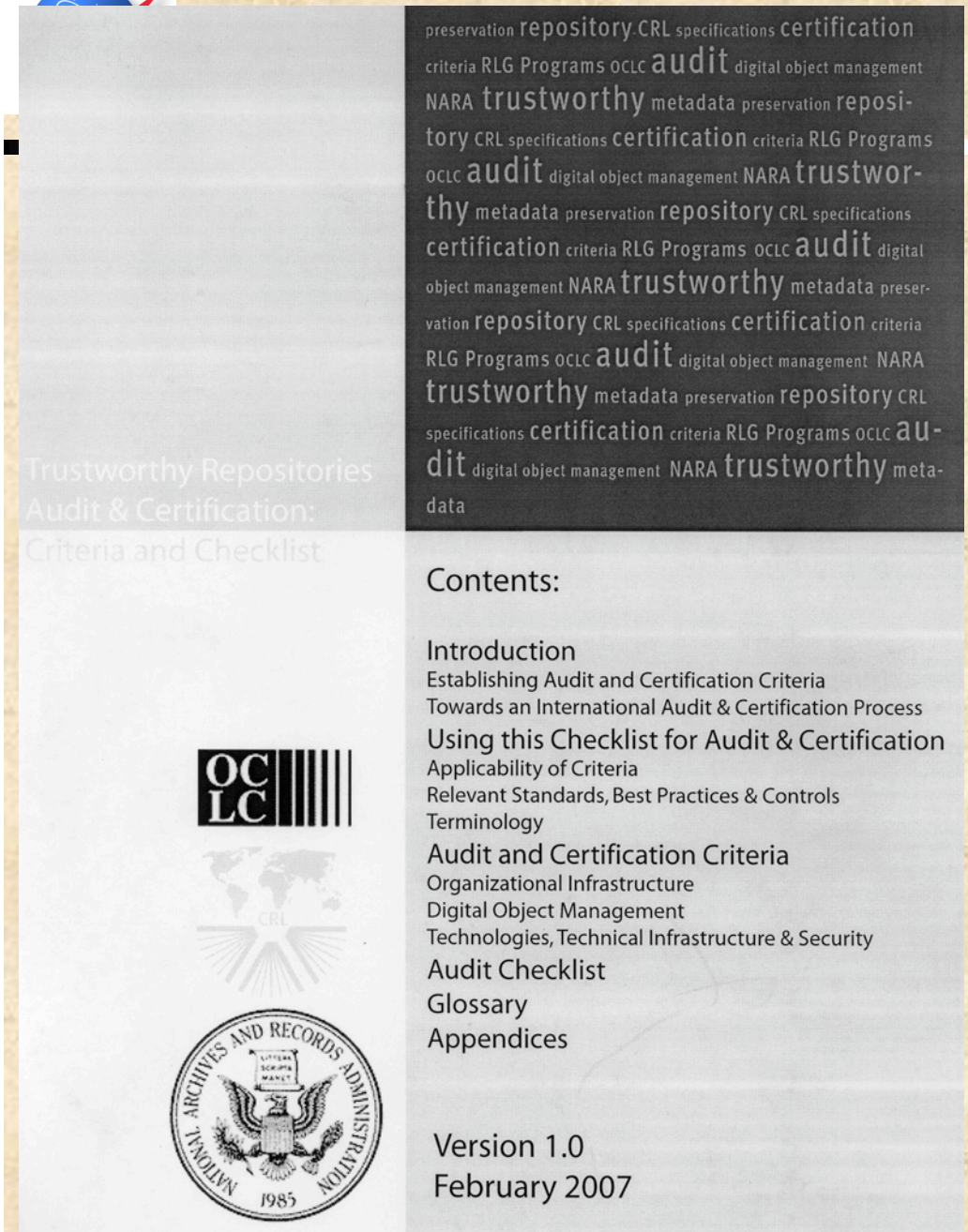


Table of Contents

TRAC and Resident Archives

- An Initial NSSDC Response to the Trustworthy Repositories Audit & Certification: Criteria and Checklist Document

  Don Sawyer and Ed Grayzeck, December 16, 2007

- The NSSDC has been active for a number of years in supporting various standards intended to benefit the NASA space science community. One related activity has been participation in the development of the "Trustworthy Repositories Audit & Certification: Criteria and Checklist" document, published February 2007, and commonly referred to as the TRAC document (http://www.crl.edu/PDF/trac.pdf). It was developed by invited individuals from a broad range of repositories, including science data centers, national libraries, and national archives. It was always intended that this work should be an important input to an ISO standard, and that has happened under the Consultative Committee for Space Data Systems (CCSDS). A Repository, Audit, and Certification (RAC) group has been formed that is fully open to all participants, and the major thrust has been to use the TRAC document as its primary working document. (see http://wiki.digitalrepositoryauditandcertification.org/bin/view/Main/WebHome).

- NSSDC has at least two motives for beginning to evaluate itself against these requirements at this time:

  - As a participant in TRAC and RAC, such responses should provide a practical view that will inform the evolution of the standard (John Garrett continues with RAC WG)

  - NSSDC is being viewed as a party who should provide guidance and some technical oversight to Heliophysics 'Resident Archives'. There has been some thought that a document, possibly modeled after some form of TRAC, may serve as a good guideline basis.

TRAC document mapped to RAs

- B. Digital Object Management
- B1. Ingest: acquisition of content
- B1.1 Repository identifies properties it will preserve for digital objects.
- B1.2 Repository clearly specifies the information that needs to be associated with digital material at the time of its deposit .
- B1.4 Repository's ingest process verifies each submitted object for completeness and correctness as specified in B1.2.
- B1.5 Repository obtains sufficient control over the digital objects to preserve them.
- B1.6 Repository provides producer/depositor with appropriate responses at predefined points during the ingest process
- B1.7 Repository can demonstrate when preservation responsibility is formally accepted for the contents of the submitted data objects.
- B1.8 Repository has contemporaneous records of actions and administration processes that are relevant to preservation.

  …….
- B6. Access
- B6.1 The repository documents and notifies the designated community what access and delivery options exist
- B6.2 The repository has implemented a policy to record access actions
- B6.4 The repository has documented and implemented distribution policy that matches ingest policy (e.g., for proprietary cases)
- B6.10 The repository enables the dissemination of authentic copies
- C. Infrastructure
- ……
- C3 security
- C3.1 Repository maintains a systematic analysis of such factors as security needs
- C3.2 Repository has controls on security needs
- C3.3 Repository addresses security needs with appropriate staff
- C3.4 Repository has suitable written disaster recovery plan

Heliophysics RA Example: TRAC attributes *

Excerpts from the Draft Heliophysics Data Environment policy (August 2008)

An archive should include, in addition to the archival resources themselves:

- **A statement of its purpose, scope, and audience;**
- **An inventory of science data, ephemeredes, attitude, engineering, and any other (e.g., browse, higher-level, event list, or combined) products needed for scientific use of the basic products, and of the documentation associated with the production and validation of these;**
- **An inventory of the relevant documentation of the spacecraft, instruments, and instrument calibrations;**
- **SPASE descriptions of the archival resources that include Access and Information URLs that point to the resources and related documentation;**
- **Documented methods for providing archival resources to users such that they will be able to assess the utility of the scientific data and use it once accessed;**
- **Easily accessible and documented analysis tools;**
- **Documented means of serving the resources, including through VOs;**
- **A log of actions and changes that includes provenance information sufficient to know the origin and history of each resource and to assure resource and referential integrity;**
- **Documented procedures to maintain the integrity of the resources, e.g, using distant mirror sites, checksums, periodic checking, monitoring, and appropriate physical and software security measures;**
- **Written disaster recovery plans;***
- **Hardware and software upgrade plans;**
- **Procedures to obtain feedback from stakeholders (providers and end-users) and to make changes based on this;**
- **A written plan for transferring the data to another archive if necessary ***

NASA TRAC Discusisons

Initial evaluation by NSSDC

- Discussion at ADEC telecons in winter 2008

  Evaluations by ADEC members in spring 2008
- Presentation to PDS Management Council in April, 2008

  Adoption of Data Integrity & Disaster Recovery, July 2008
- Initial discussion with HDMC in June 2008

  Some TRAC attributes incorporated in draft policy Aug 2008

Future Goals for NSSDC

- Streamline permanent archive operations so all data holdings are media independent by using Archive Information Packets (AIPs)

  - Enhance software to create AIPs remotely at mission or active archives, e.g., SPDF

- Update Resident Archives, as needed, for mission data

  - INCORPORATE Space Physics Search and Exchange (SPASE)

- Enhance Active Archive interactions
  - Initiate data registry with Virtual Observatories
  - Evolve the management of RAs to be cost effective and compatible with Heliophysics Data Environment policy
  - Plan Science Archives II Workshop

Background: RA Life Cycle