Performing a Migration in the Framework of the OAIS Reference Model: NSSDC Case Study

Donald Sawyer [(1)], H. Kent Hills [(2)], Pat McCaslin [(3)] John Garrett [(4)]

[(1)] National Space Science Data Center (NSSDC)
Code 690.1 NASA/Goddard Space Flight Center
Greenbelt, MD 20771 USA
Donald.M.Sawyer@nasa.gov

[(2)] QSS Group, Inc./NSSDC
Code 690.1 NASA/Goddard Space Flight Center
Greenbelt, MD 20771 USA
Hills@mail630.gsfc.nasa.gov

[(3)] QSS Group, Inc./NSSDC
Code 690.1 NASA/Goddard Space Flight Center
Greenbelt, MD 20771 USA
McCaslin@mail630.gsfc.nasa.gov

[(4)] Raytheon, Inc./NSSDC
Code 690.1 NASA/Goddard Space Flight Center
Greenbelt, MD 20771/USA
John.Garrett@gsfc.nasa.gov

Abstract. The migration of scientific and technical data for its long-term preservation is a common practice in digital archives that have been in existence for as short as ten years, due to the rapid pace of technology change. Until the development of the "Reference Model for an Open Archival Information System (OAIS)", ISO 14721 (2003), there has not been a widely accepted framework in which to discuss the various practices of existing archives. This paper expands on the discussion regarding the various types of migration described in the OAIS reference model in terms of the OAIS functional modeling and information modeling as applied to scientific and technical data, and to their related documents. Using this framework, the experience at NASA's forty-year old National Space Science Data Center, of using the OAIS reference model concepts to describe and evolve its systems and to help inform its on-going migration of data from legacy 9-track and 3480 cartridges to Digital Linear Tapes, is described.

1 Introduction

The National Space Science Data Center (NSSDC), located at NASA's Goddard Space Flight Center, has been a digital archive for science data for some 40+ years. It has accomplished several data migrations during this period as described in our paper [1] at the PV-2004 conference. The focus of the present paper is to first (section 2) consider typical science and technical data migrations from the perspectives of the Reference Model for an Open Archival Information System (OAIS) [2], hereafter just 'OAIS Reference Model'. The second focus (section 3) is to then look at actual experience of a new migration of science data at NSSDC in this same framework, identifying strengths and weaknesses, and lessons learned.

Science and Technical data can take many forms and have a wide variety of content types. However for our purposes, this data is characterized as mostly numeric, often binary, with some embedded text, and composed of a large number of data objects (typically files) with similar data structures. It must be accompanied with adequate documentation so that the structure and meaning of the digital objects can be understood by those communities that the archive is serving.

2 OAIS Migration Context

The OAIS Reference Model has been widely adopted as a conceptual framework, providing common terms and concepts, by all types of digital archives and repositories regardless of discipline. It defines information modeling concepts to facilitate addressing the meaning of information and the types of information needed to help preserve other information. It defines a functional model of an archive to facilitate discussions on archival activities. It goes into some detail in discussing migration as a key preservation activity of many digital archives. These views are elucidated in the following sub-sections and are related to our characterization of scientific and technical data.

2.1 Information View

The OAIS Reference Model defines an Information Object as a fundamental concept. As shown in Fig. 1, a Data Object together with its Representation Information yields an Information Object. The Data Object may be a physical object or a digital object. The Representation Information describes how to interpret the structure of the digital object and the meaning associated with this structure. Since the Representation Information is also an Information Object, it can be viewed as composed of its own Data Object and associated Representation Information. This is recursive until the data object no longer needs Representation Information in practice, such as when it is a physical document written in a sufficiently understandable manner or it is a digital document whose format, such as text, is renderable by widely available tools and whose content is sufficiently understandable. Since Representation Information may be composed of multiple Information Objects, the recursion leads to a Representation Net. Its termination, in practice, is an archival decision that should be made while considering the need to preserve information into the indefinite future. It is likely to need periodic review for adequacy.

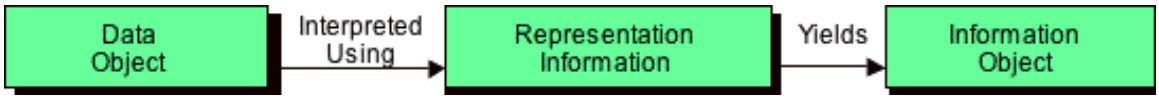

Fig. 1. An Information Object is composed of a Data Object and the Representation Information that allows the Data Object to be sufficiently understandable. The Data Object may be a physical object or a digital object. Representation Information is itself an Information Object, leading to recursion. (Fig. extracted from reference [2])

For example, a science data file containing observations of magnetic field values taken in space is written as a sequence of fields repeating throughout the file. Its Representation Information may be in the form of a paper document describing the format of the file, the meaning of the fields, how they relate to one another, and how the observing instrument made its observations. This may also be in the form of a digital file in the PDF format, in which case the Representation information also needs to include a description of the PDF format, or a well understood pointer to such a description. An identification of a proprietary program that reads a PDF is not suitable, but a description of the algorithm actually used for reading the PDF would also be suitable.

The OAIS Reference Model also identifies several categories of information objects. Information that is the primary target of preservation is called Content information. Information associated with the process of preserving the Content Information is called Preservation Description Information (PDI), and it is broken down into Provenance Information, Context Information, Reference Information, and Fixity Information as follows [2]:

- **"***Reference Information:* This information identifies, and if necessary describes, one or more mechanisms used to provide assigned identifiers for the Content Information. It also provides those identifiers that allow outside systems to refer, unambiguously, to this particular Content Information. Examples of these systems include taxonomic systems, reference systems and registration systems. In the OAIS Reference Model most if not all of this information is replicated in Package Descriptions, which enable Consumers to access Content Information of interest.
- *Context Information:* This information documents the relationships of the Content Information to its environment. This includes why the Content Information was created and how it relates to other Content Information objects existing elsewhere.
- *Provenance Information:* This information documents the history of the Content Information. This tells the origin or source of the Content Information, any changes that may have taken place since it was originated, and who has had custody of it since it was originated. This gives future users some assurance as to the likely reliability of the Content Information. Provenance can be viewed as a special type of context information.
- *Fixity Information:* This information provides the Data Integrity checks or Validation/Verification keys used to ensure that the particular Content Information object has not been altered in an undocumented manner. Fixity Information includes special encoding and error detection schemes that are specific to instances of Content Objects. Fixity Information does not include the integrity preserving mechanisms provided by the OAIS underlying services, error protection supplied by the media and device drivers used by Archival Storage. The Fixity Information may specify minimum quality of service requirements for these mechanisms."

Examples of PDI for space science data are given in Table 1 as extracted from OAIS Reference Model [2].

Table 1. Examples of Preservation Description Information for space science data, extracted from reference [2]

Provenance	Context	Reference	Fixity
• Instrument description • Processing history • Sensor description • Instrument mode • Decomutation map • Software interface specification	• Calibration history • Related data sets • Mission description • Funding history	• Object identifier • Journal reference • Mission, instrument, title, attribute set	• CRC • MD5 • Reed-Solomon coding

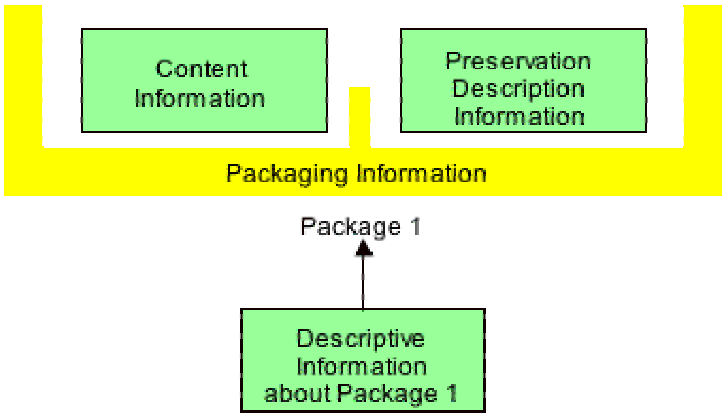

Fig. 2. An Archival Information Package consists of the Content Information and its Preservation Description Information. These Information Objects are bound by Packaging Information and are associated with Descriptive Information used to support searching for the Content Information of interest within an OAIS archive. (Fig. extracted from reference [2])

The OAIS Reference Model also defines the concept of an Information Package and specializes this for archival storage as an Archival Information Package (AIP). The AIP is further specialized as either an Archival Information Unit (AIU) that is the smallest collection of information with full PDI, or an Archival Information Collection that consists of multiple AIUs with supporting PDI. Fig. 2, extracted from OAIS Reference Model [2], shows important AIP Information Objects and their relationships.

The Content Information and PDI of an Archival Information Package are kept together by virtue of the Packaging Information. Associated with the AIP is Descriptive Information that is used by the archive to support searching for the Content Information of interest. For example, Descriptive Information for science data files containing magnetometer data would typically include time range covered, the locations of the observations, and the types of magnetic field components available.

2.2 Functional View

The OAIS Reference Model provides a functional view, at three increasing levels of detail, to support an understanding and comparison of archival activities. Fig. 3 shows an overall view at the middle level of detail and it identifies the six functional areas within the OAIS archive.

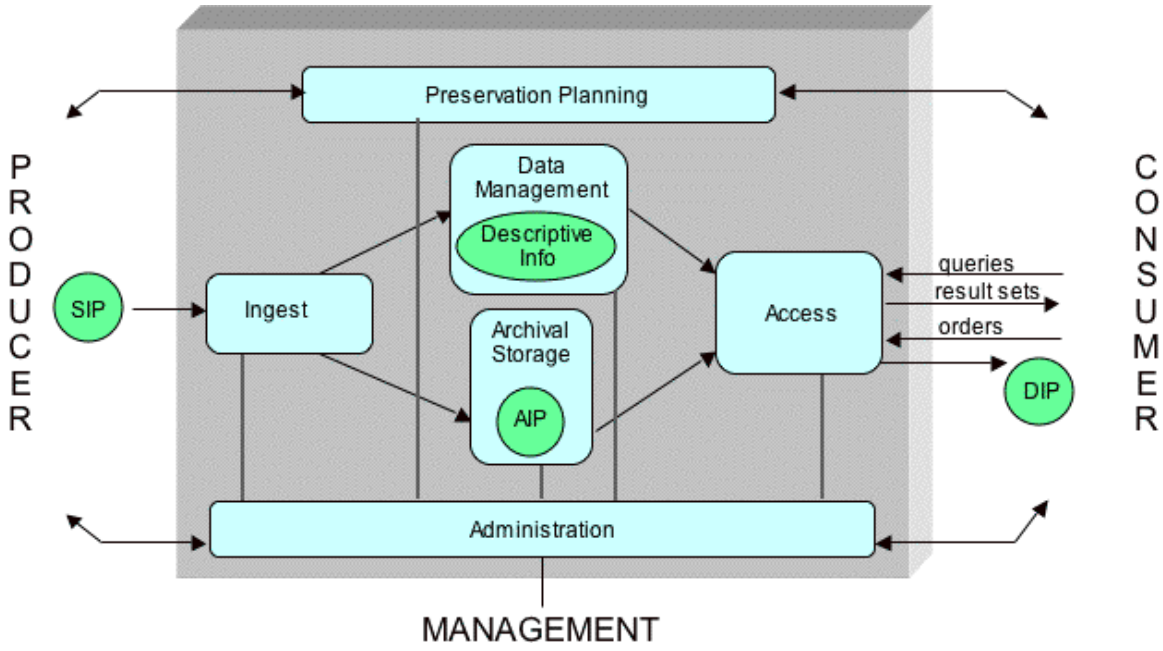

Fig. 3. Functional view of OAIS Reference Model at intermediate level of detail showing the six major functions. The major information flow, from the Producer, through the archive, to the Consumer is also shown. (Fig. extracted from reference [2])

In this functional view, a Producer is the role played by a person or system that submits information to the archive to be preserved. This information is submitted in the form of a Submission Information Package (SIP). A SIP is a type of information package that may already be an AIP, or it may have some of the information needed to construct one or more AIPs. As SIPs are received by the Ingest function, they are converted to AIPs and passed to Archival Storage. In synchronization, appropriate Descriptive Information is passed to Data Management to support searching for the AIPs. Archival Storage preserves the AIPs, makes appropriate backups, and refreshes the media on which they are stored as needed. It also provides the AIPs to the Access function upon request. The Access function responds to external (and internal) requests for information held by the archive. It supports a Consumer, which is the role played by persons or systems who want archived information by providing finding aids that use Descriptive Information obtained from Data Management. It may also provide processing of AIPs to satisfy the Customer's specific needs, returning one or more Dissemination Information Packages (DIPs) as appropriate.

Supporting the basic flow is the Administration function. It provides day-to-day management of the archive and coordination among the other functional areas. It also plays a key role in the migration of information within the archive, as described in the next section. Administration is supported by the Preservation Planning function that monitors the state of the community served by the archive, monitors the state of appropriate technology, and it provides recommendations on standards and techniques to be used by the archive (also see section 3.1). The Management entity is the role played by those that oversee the archive less frequently than day-to-day, and they often provide funding.

Each of the six functional areas is broken out into greater detail within the OAIS Reference Model to provide additional terms and concepts, and to help clarify the scope of these areas. It is important to note that implementations of archives are not required to group functionality as shown here. Nevertheless, having this model aids comparisons within and among archive implementations.

2.3 Migration View

Digital migration is defined to be the transfer of digital information, while intending to preserve it, within the OAIS. This transfer of information, under full control by the archive, is intended to preserve the full information content and the result is intended to be an adequate replacement of the previous form. (Note: this does not mean the previous form must be discarded, however.) Fig. 4, extracted from the OAIS Reference Model [2], provides key functional and information modeling concepts as they relate to migration perspectives. It begins by relating an externally visible identifier for Content Information to an internal 'AIP Identifier' using local 'Descriptive Information Mapping'. Then the 'AIP Identifier' is mapped to an actual AIP using local 'Archival Storage Mapping'. The AIP is expressed as Packaging Information that contains Content Information and its associated Preservation Description Information.

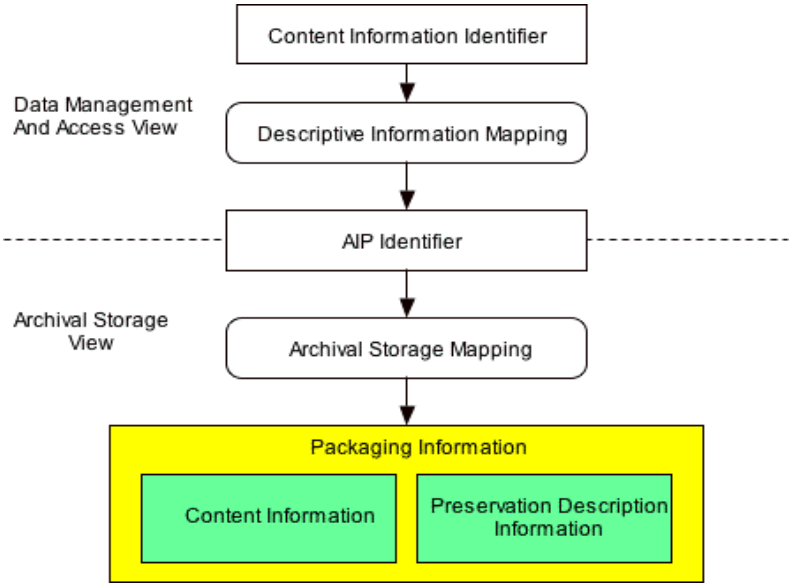

Fig. 4. Conceptual view of relationships among key OAIS information concepts, functional areas, and supporting infrastructure. The AIP has an internal unique identifier (AIP identifier) that is associated with Content Information identifiers (Reference Information) as a part of Descriptive Information to support locating the AIP of interest. Migrations may require updates to some of the entities shown. (Fig. extracted from reference [2])

The OAIS has defined several types of migration and they have different implications for the conceptual components and infrastructure identified in Fig. 4. From OAIS [2], "The primary types, ordered by increasing risk of information loss, are:

- *Refreshment:* A Digital Migration where a media instance, holding one or more AIPs or parts of AIPs, is replaced by a media instance of the same type by copying the bits on the medium used to hold AIPs and to manage and access the medium. As a result, the existing Archival Storage mapping infrastructure, without alteration, is able to continue to locate and access the AIP.

- *Replication:* A Digital Migration where there is no change to the Packaging Information, the Content Information and the PDI. The bits used to convey these information objects are preserved in the transfer to the same or new media-type instance. Note that Refreshment is also a Replication, but Replication may require changes to the Archival Storage mapping infrastructure.
- *Repackaging:* A Digital Migration where there is some change in the bits of the Packaging Information.
- *Transformation:* A Digital Migration where there is some change in the Content Information or PDI bits while attempting to preserve the full information content."

Unless an AIP migration involves a Transformation, it is not considered to result in a new AIP version. When a Transformation is involved, the PDI is updated to identify the previous version and to reflect the transformation that occurred. Associated Descriptive Information, used to support finding the Content Information by Consumers, should also be updated. A new (local) AIP ID is needed to reflect the new version. Although the new version is intended as a replacement for the previous version, the previous version may also be kept. However from the archives perspective, removing the previous version does not result in significant information loss.

There are two types of transforming migrations that have been defined, based on the impacts to the Representation Information. A Reversible Transformation occurs when the changes in the Representation Information have a one-to-one mapping between the original and subsequent Representation forms, which means that an algorithm can be used to convert the Content Information back to its previous form. For example, replacing ASCII characters with UNICODE characters is a Reversible Transformation. However a Non-reversible Transformation occurs whenever a Reversible Transformation cannot be guaranteed. For example, replacing an IBM 7094 floating point with an IEEE floating point value can not be reversed with an algorithm because they do not have the same range for representing values.

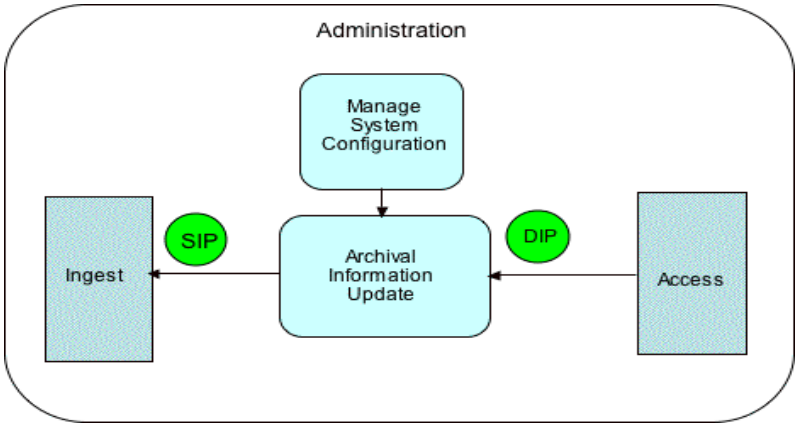

Fig. 5. The Administration function manages migrations involving repackaging or transformations of AIPs by requesting DIPs from the Access function, performing the necessary processing within the Archival Information Update sub-function, and providing the resulting SIPs to the Ingest function.

There are two other operations on AIPs that are closely related to a transforming migration, but are not classified as migrations. A new Edition of an AIP occurs when the processing of the AIP is intended to improve the information content in some way. An example, based on the science file described earlier, would be making a correction in the documentation of the format, or adding additional context information, or simply adding new processing history information. A Derived AIP occurs when the transformation involves incomplete information extraction from the source AIP, or the combining of multiple AIPs. In both these cases, new AIPs with new AIP IDs result with their PDI reflecting their processing history. A new Edition may replace a previous Edition, but a Derived AIP may not replace any of the source AIPs from which it was derived. Note that it is possible to perform a migration, followed immediately by a new Edition, all within a single processing step. The result would be a new version and a new Edition.

From an OAIS functional perspective, migrations involving repackaging or transformations are orchestrated by the Administration sub-function called Archival Information Update as shown in Fig. 5. Archival Information Update, responding to the Manage System Configuration sub-function, requests

Dissemination Information Packages (DIPs) from the Access function. It then processes DIPs to create Submission Information Packages (SIPs) and feeds these to the Ingest function. As described earlier, the Ingest function turns these into new AIPs for storage by the Archival Storage function and it provides appropriate Descriptive Information for the Data Management function.

3 Legacy Tape Migration at NSSDC

The National Space Science Data Center has been an archive for digital data for over 40 years. During this period, a number of migrations have taken place as reported at PV-2004 [1]. By 1995, NSSDC had migrated 35,000 tapes into 6,100 pairs of 9-track and 3480 cartridges with a recovery rate in excess of 98%. About 1/3 of the tapes were 7-track, and when writing to 9-track tape the extra two bits per byte were padded with zeroes when in character mode. For binary data the bit stream was preserved with padding at the end of the word or record if needed. Some tapes, from the mid-1970s and from specific vendors, were classified as 'sticky tapes' because their recording film tended to flake off and stick to the tape drive read heads, necessitating frequent head cleaning, and often resulting in data loss. In some cases the tape could be read the first time, but not on a subsequent pass because of degradation of the recording medium. The policy for all tape copies was to do verification on the new tape's content and quality; old tapes were usually sent for re-certification or discard. Provenance Information addressing the 2% data loss was recorded on hardcopy.

A decision was made to migrate these 'legacy data'— including data previously migrated and data still residing on original tapes — to NSSDC's DLT-based storage system. This was motivated by concerns over media decay and by the need for improved cost-effectiveness. These data reflect nearly 1,600 distinct data sets, each with its own or shared documentation. The documentation forms range from digital, to microfilm, to paper, depending on the particular data set. During the previous migration of these data, multiple original tapes were written to ('stacked upon') higher capacity tapes to save space, as noted above. However the documentation largely reflects their previous existence on lower density tape, but with some notes to reflect this new media. Documenting the mappings from low density to higher density tape was performed by manual data entry in a tape inventory database known as Interactive Digital Archive (IDA).

3.1 Defining AIPs for Legacy Tape Data

NSSDC has adopted the AIP concept and has developed a single file implementation that is an AIP/AIU. The design of this implementation was accomplished, in the OAIS sense, under the Preservation Planning function shown in Fig. 3. The original AIP implementation combined a single science data file with a set of attributes and pointers to related documentation maintained by NSSDC [3]. For a number of reasons it was recognized that the AIP implementation needed to be upgraded to handle multiple science files, with attributes about each, and with attributes about the set of files in each AIP. Since the legacy tape documentation primarily reflected the original organization of data in files on low density tape, it was decided that the set of science files originally on a single low density tape would be put into a single file AIP. Appropriate attributes for each file, as well as for the entire tape image, would be captured or generated and included. This would minimize the extent of updates needed for the documentation.

A schematic of the resulting AIP, holding two science files, is shown in Fig. 6. All the attributes are included in one encapsulated object called the 'NSSDC Attribute Object', but partitioned into general attributes and attributes for each science file. Attributes for each science file include PDI associated with tape-to-disk and disk-to-AIP packaging, as well as limited Representation Information (format information) and pointers to full Representation Information. These pointers are called Authority and Description Identifiers (ADIDs) and are internationally standardized to support retrieval of the Representation Information from distributed registries called Control Authority Offices. NSSDC operates such a Control Authority Office and registers descriptions for each unique format in an AIP.

Each science file is grouped with an identifier and can be matched to its partitioned attributes. Since each science file is on ½-inch magnetic tape, the records of these files are separated by inter-record gaps. These files are transformed by inserting byte count fields before each record to maintain record identification. Each resulting file, referred to as a "canonical file", is stored as a continuous stream of bytes

in the AIP. Byte counts and checksums of a file are recorded before and after the transformation because experience has shown that data corruption can happen in many ways and is often difficult to detect. Clearly these are reversible-transforming migrations that result in new AIPs with new internal AIP Identifiers. Associated Provenance Information, from this routine processing, is incorporated to reflect the transformations.

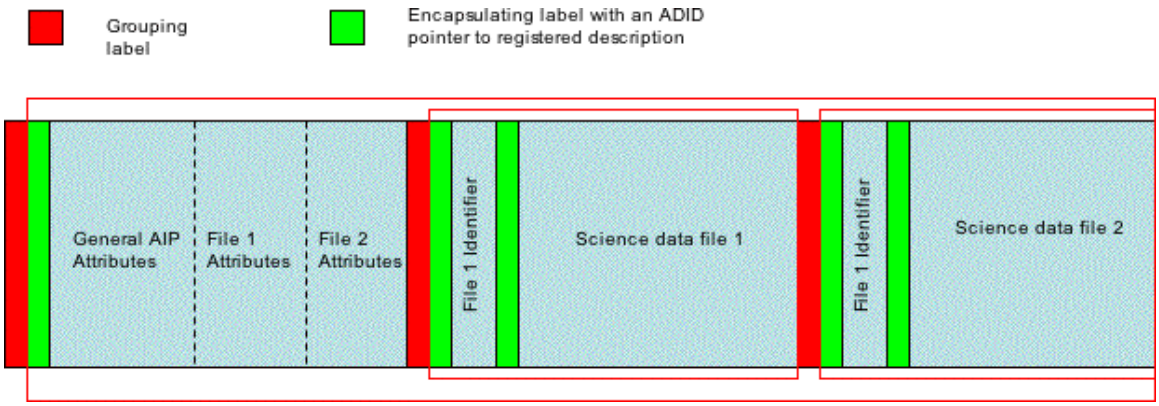

Fig. 6. Schematic of NSSDC's single file AIP implementation showing its ability to contain two (or more) science files and their associated attributes. The labels contain byte counts to delimit the following data and standard pointers (ADIDs) to registered descriptions (Representation Information) of the delimited data. Red labels indicate internationally standardized containers with standard descriptions of the content. NSSDC Attribute Objects contains routinely-generated PDI and some Representation Information, including duplication of the science file ADID pointers from their labels

3.2 Building AIPs from Current Tape Data and Metadata Stores

Previous migrations of NSSDC legacy data were performed using largely non-automated procedures that relied on manual recording of processing results. Methods used for capture and preservation of Provenance Information from those migrations were not consistent. Much of the processing information that was captured exists in hardcopy form only and is not readily accessible.

A different approach based on OAIS concepts is employed for the current legacy migration. NSSDC has developed the "NSSDC Legacy Tape Data Migration Process Description". This comprehensive data migration plan covers: collection of reference information from varied sources in a single database, use of a highly automated AIP generation process, use of a database to manage migration processing, and the capture and preservation of Provenance Information into a relational database. The plan has a long evolutionary history, starting out as a high level migration strategy that became progressively more detailed during the lengthy analysis of migration issues. The plan has allowed detailed review of the proposed approaches by the 'migration planning team' and will be updated to reflect operational experience. After the migration, the plan will be preserved to provide a historical context for the migration effort.

Over the course of the migration analysis and planning process it became apparent that identification, capture, and preservation of all supporting information objects (i.e., all PDI and Representation Information) as defined in the OAIS Reference Model would be crucial to the long-term usability of the legacy data. The fact that Provenance Information generated in previous migrations is available only on hardcopy complicates preservation of supporting information in the current migration. A more serious problem is the large number of inconsistencies observed in the mappings from low density to higher density tape. These errors in the IDA database appear to have been introduced during manual updates in previous migrations.

The legacy tape migration process is complex, requiring a number of separate operations to collect PDI and Representation Information, retrieve science files from media, correctly associate science files with PDI and Representation Information, generate AIPs, and store AIPs on permanent media. In order to manage these activities and capture Provenance Information generated while performing these activities NSSDC has created several new databases and applications.

- *Legacy Off-line Media Data Migration Master List:* This simple database application is used for the initial collection of attributes identifying all data sets to be migrated along with other related attributes addressing Context Information, Representation Information, Packaging Information, Archival Storage Mapping information, access control information and access location information. Using this application, NSSDC operations personnel assemble attributes from NSSDC documentation, from the NSSDC Information Management System (NIMS) database, and from the IDA database. The process of assembling these data set attributes is not automated but the Master List provides some validation mechanisms..

- *Offline Transition to Online (OTTO) Database:* This relational database is used to manage the legacy data migration. OTTO preserves a record of routine events, part of the Provenance Information associated with the migration, as well as:
  - Data set attributes imported from the Legacy Off-line Media Data Migration Master List.
  - Inventory of tape contents imported by operations personnel from NIMS
  - Information about the actual tape contents supplied by the Tape Read Software. OTTO stored procedures automatically compare actual tape contents with stored tape inventory information and flags discrepancies. Some automatic validation of data files is performed by comparison of data file characteristics with data set attributes imported from the Master List. (e.g., if stored attributes indicate a data set's file record format is fixed length but not all of the files are found to be the same length then a discrepancy is flagged)
  - Processing status
  - AIP generator processing parameters. OTTO produces files containing the processing instructions needed by the AIP generator software

- *Provenance Database:* The Provenance Database captures Provenance Information associated with non-routine events during the legacy tape data migration. Events and data anomalies and the relationships between them may be recorded in the provenance database.

- *Tape Read Software:* This specialized software is used to retrieve data files from magnetic tapes. The tape read software reads each file on a tape, calculates a CRC-32 value for the file, and writes the data into a variable length record structured file on a VMS magnetic disk. The software produces a tab delimited report file containing information which is imported into OTTO.

- *Multifile Package Generator and Analyzer (MPGA):* MPGA is the software utility that creates AIPs and accesses data objects and attribute objects contained within AIPs. MPGA was created from an existing NSSDC data packaging application that was extensively modified to be able package multiple data objects into a single AIP. There are five MPGA processing modules:
  - AIP generator – creates AIPs. Reads data files staged to VMS disk and packages them into AIPs for delivery to the NSSDC Data Ingest and Online Access System (DIOnAS). For the purposes of the legacy tape data migration, each AIP will contain data files from a single original tape as noted earlier.
  - Extractor – Extracts the NSSDC Attribute Object found inside an AIP The extractor is used by the DIOnAS to identify AIPs.
  - Splitter – Splits AIPs into canonical science files and attribute files containing NSSDC Attribute Objects. The splitter is used by DIOnAS to generate data files and attribute files for placement on the NSSDC FTP site.
  - Attgetter – Writes the NSSDC Attribute Object from an AIP into a text file. The text files are stored in a database in order to provide a permanent and searchable repository of attributes for all AIPs processed by DIOnAS.
  - Restorer – Reads AIPs and creates copies of the source data files in their original formats on a VMS machine. The restorer will be used as a quality assurance tool during the data migration.

The migration activities and the databases and applications that support them conform closely to the functional view of the OAIS Reference Model, as shown in Fig. 7. The one deviation from the functional view is that the DIOnAS (Data Ingest and Online Access System) performs some Ingest as well as Archival Storage functions. It is worth noting that the Archival Storage of AIPs at the NSSDC will soon be taken over by a separate system.

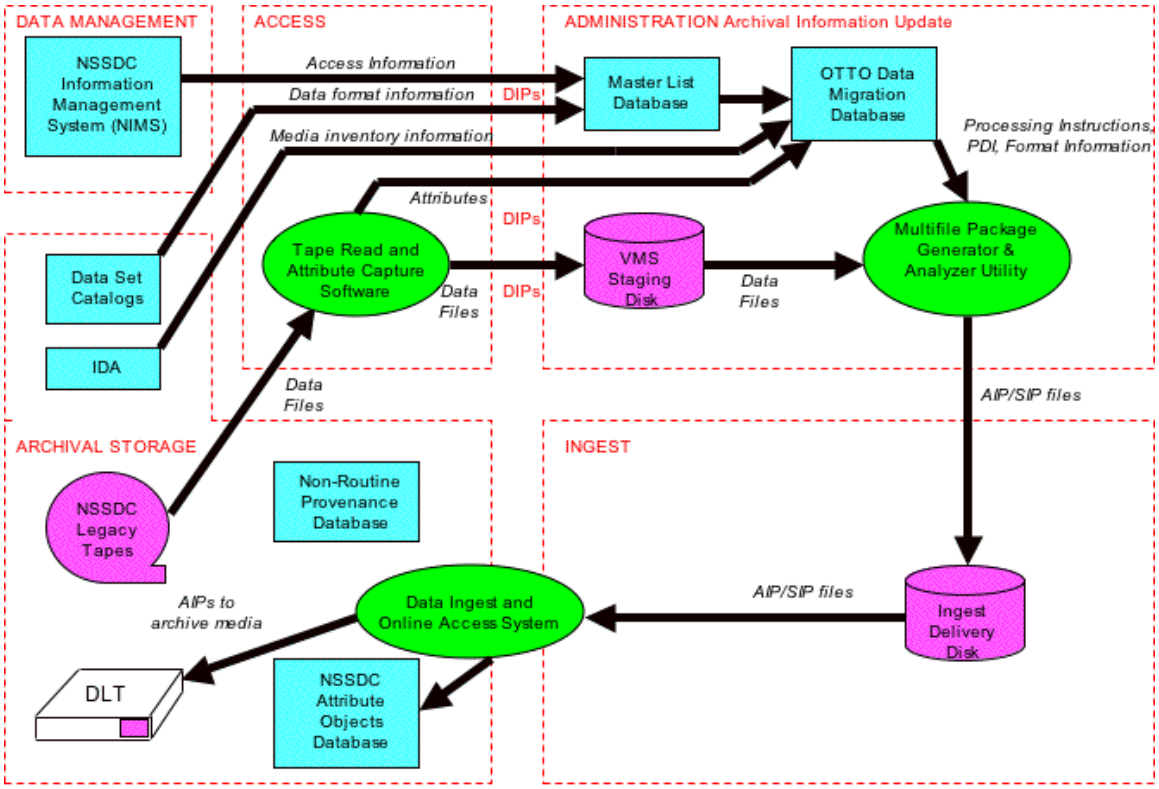

Fig. 7. Major processes and key components of the NSSDC Legacy Tape Migration are mapped to OAIS functions. NSSDC Information Systems and Data Set Catalog populate the migration-controlling OTTO database in Archival Information Update while specialized software in Access reads legacy tapes to add attributes to OTTO and provide raw data files for migration. MPGA transforms the data files into canonical form, creates AIPs, and sends them to Access where they are processed by DIOnAS and stored on DLTs in Archival Storage

Existing attributes of the legacy data sets reside in various NSSDC information repositories as shown in Fig. 7. The NIMS database, serving the Data Management function, contains access location information and access control information. Data set catalogs and the IDA database support Archival Storage. The data set catalogs, that may exist on paper, on microfilm, or in digital files contain data format information and Provenance Information from previous migrations. The IDA database holds archival storage information for the legacy media.

IDA relates data sets to tape volumes and relates existing tape volumes to original tape volumes. As such it is a combination of Archival Storage Mapping information and past migration Provenance Information. Media inventory information in IDA is imported into the OTTO database. This transfer of information was originally conceived to be a single transfer of inventory information for all NSSDC legacy media. However, the large numbers of errors found in the IDA information have obliged NSSDC operations personnel to review and correct information one data set at a time. The majority of these errors can be attributed to data entry mistakes made during earlier migrations when little automation was employed.

NSSDC operations personnel assemble access information, PDI and Representation Information for legacy data sets and record selected attributes, including pointers to Representation Information, in the migration master list database. Access location and access control information is provided to NSSDC acquisition scientists for review and possible modification. The NIMS database is updated to reflect any modifications. Assembled PDI and Representation Information are imported into the OTTO database.

Using the imported attributes the OTTO database automatically generates a Data Description Package (DDP) for each data set, assigning a format description identifier (ADID identifier) taken from a block of ADIDs that have been allocated for the legacy tape migration. These DDPs are considered provisional

until reviewed by the NSSDC Control Authority Office. They are preserved as a part of Archival Storage, although not shown explicitly.

NSSDC acquisition scientists review each data set's record in the OTTO database to validate PDI, Representation Information, and pointers to Representation Information before packaging data into AIPs. The OTTO database will not allow AIP creation for a given data set by MPGA until approval for that data set has been recorded.

Legacy tape volumes from Archival Storage are read on a computer running the OpenVMS® operating system. OpenVMS® is employed for this purpose because of its ability to recognize the various record formats found in the legacy data. Data files are retrieved from legacy tape volumes (often containing multiple original tape volume file sets) using specialized software, running in Access, that copies data files to magnetic disk and inserts information about each file retrieved into the OTTO database. Information items needed to identify and validate the data files are: the magnetic tape identification number, the position of the file on the tape, and the maximum and minimum record lengths encountered when reading the file. A CRC32 checksum value is calculated for each file as it is copied and is also stored in the OTTO database. This Fixity Information is used for validation during the AIP generation process and is part of the PDI preserved in the AIP. A short status message that the tape read software produces as each file is read is also stored in the OTTO database. This status message is the record of the data retrieval operation used by the OTTO database to determine the success or failure of the operation, and it is stored as Provenance Information in the OTTO database.

Once inserted into the OTTO database, the information returned by the tape read software is used for automated identification and validation of the data file. This information is compared with the media inventory information previously imported into the OTTO database. Based on each file's recorded sequential position from its legacy tape, the OTTO database determines the identity of the original tape volume that contained that file. To the greatest extent possible, the record format of each file is validated by comparison of the file's record structure reported by the tape read software with data set format information previously imported into the OTTO database. Any discrepancies that are encountered during the identification and validation process are flagged in the database and further processing of the affected original volumes file set is blocked until the problem is resolved.

The OTTO database provides a view of each original media volume file set that has been successfully retrieved, copied to disk, identified, and validated. Using a graphical interface, operations personnel select the file sets from such volumes to be staged for packaging. The OTTO database generates a set of instructions that run on the OpenVMS® computer to stage the science files from the selected original media volumes. For each selected original media volume a disk directory is created and populated with the appropriate data files. Processing status is captured and recorded in the OTTO database.

Another view in the OTTO database shows successfully staged original media volume file sets and allows operations personnel to select those to be packaged into AIPs. For each selected file set the OTTO database creates a "list file" containing the PDI and Representation Information attributes for that set that will be preserved in the AIP. The OTTO database places these list files in the staging directories with their associated data files. The appropriate commands to run the Multifile Package Generator and Analyzer Utility (MPGA) are constructed by the database and automatically invoked.

The AIP generator module of the MPGA reads the List File attributes and science files in the OpenVMS® AIP staging directories, transforms them to canonical form, and generates AIPs that are written to a delivery directory on the UNIX computer that hosts the DIOnAS. During the packaging process a CRC32 checksum for each data file (i.e., science file) is generated both before and after transformation, and the 'before value' is compared with Fixity Information generated when the data file was originally retrieved from tape. The MPGA reports processing status that is recorded in the OTTO database.

Yet another view in the OTTO database shows those AIPs that have been created but not yet processed by the DIOnAS. Operations personnel select AIPs to "deliver" to DIOnAS. The OTTO database generates DIOnAS processing instructions and the ingest process for the selected AIPs is automatically completed by the DIOnAS. The DIOnAS extracts the NSSDC Attribute Object contained in each AIP and the resultant object is written to the NSSDC Attributes Object Database as an aid to Archival Storage diagnostics, should this be needed. Each AIP is written to two separate DLTs in the DIOnAS jukebox. In addition, although not shown, a copy of each AIP is written to a magnetic disk in a remote location where it remains until it is confirmed that one of the two AIP copies on DLT has been moved to an off-site storage location.

The use of Cyclic Redundancy Checks (CRCs) is NSSDC's primary safeguard against data corruption during the legacy tape migration and over an indefinite period while data are preserved in the archive. CRC values are generated at several stages of the migration process and both the MPGA and the DIOnAS perform CRC comparisons. Nonetheless, the NSSDC plans to retrieve all AIPs from DLT, restore then to native format data files on the OpenVMS® system, and perform byte-for-byte comparisons with the original files.

The OTTO database identifies the entities involved in the migration (e.g., data sets, existing tapes, original tapes, data files) and records a processing status for all objects that it tracks. It represents the Provenance Information for routine legacy tape migration processing events. However, the OTTO database has no provision for preserving "special case" Provenance Information. There has been no way to record significant non-routine processing events associated with the migration and there has been no mechanism to identify noteworthy anomalies in the data that are being migrated. To preserve non-routine Provenance Information the NSSDC is developing a dedicated provenance database. This database allows significant processing events and observed data anomalies to be recorded. For example, a small number of legacy tapes have ANSI standard labels and must be handled in a non-routine manner. Relationships between events, anomalies, and objects tracked in the OTTO database may be established. The combination of the OTTO database and the provenance database allow a capture of a comprehensive set of migration Provenance Information. The question of how to preserve and link all the Provenance Information once the migration is completed and the OTTO database is no longer updated remains outstanding.

3.3 Initial Processing Experience

NSSDC has thus far selected 525 legacy data sets to be migrated. The assembly of PDI and Representation Information for these data sets in the Legacy Off-line Media Data Migration Master List is in progress. To date, 442 legacy tapes representing 166 data sets have been processed by the tape read software and their data files written to a staging disk. One legacy data set has been packaged into AIPs, although the Content Information in the six resulting AIPs has not yet been validated through byte-for-byte comparison with the source data files. Full processing in a production mode has been delayed while a number of software and hardware deficiencies are addressed as described below.

Collection of Master List attributes from paper, microfilm, digital files, and NIMS has proven to be extremely time-consuming. Given the unsystematic collection and preservation of this information during previous migrations and given the importance of this information to the long-term usability of the science data (Content Information), the collection process has received close scrutiny by the 'migration planning team' resulting in additional requirements levied on the Migration Master List. Such requirements to ensure accurate information capture include: enhanced validation of information items, automatic generation of some information items, and an improved mechanism for carrying out the review of data set PDI and Representation Information by NSSDC acquisition scientists. The Master List was originally conceived as a simple list of data set attributes maintained in a spreadsheet. Meeting the new requirements using a spreadsheet was not possible. Consequently, a new database with accompanying software for collection of data set attributes was designed. The new Migration Master List database is currently being tested by operations personnel and acquisition scientists prior to being deployed as a production system.

NSSDC has had difficulties reading data from older (> 10 years) 9-track tapes and has opted to retrieve data from 3480 cartridges whenever possible. It is a matter of concern that most of the remaining original tapes that were never migrated exist only as 9-track tapes. The read errors encountered while retrieving data from legacy tapes led to a redesign of the tape read software that allows a given tape, and its copy, to be read multiple times, increasing the possibility of complete data capture. This permits the data from separate data retrieval attempts to be combined in cases where no single attempt captured all the data. The tape read software was also modified to be able to process ANSI labeled tapes. The software was further modified such that the text of the processing status message more accurately reports the results of the data retrieval operation. This last issue is important because the text is preserved as Provenance Information in the AIP. All the modifications to the tape read software are complete.

The OpenVMS® computer that supported the initial processing is an older model with hardware and software limitations that affect the migration. The current system has insufficient disk space for continued staging of legacy data files. The system also has a slow network interface card that provides a relatively narrow network bandwidth. Ingest of new SIPs monopolizes the available network bandwidth making

concurrent new SIP ingest and legacy data migration impractical. The version of the network client software running under the older operating system is not completely compatible with the OTTO database. Lacking the ability to perform all required database updates from the OpenVMS® system, the software that performs automated staging of files has yet to be completed and tested. A new OpenVMS® computer has been acquired and is being configured to support NSSDC operations. The new system has sufficient hardware resources to address the hardware deficiencies and more recent database client software installed on it will address the database incompatibilities.

The initial migration processing experience revealed that inadequate Provenance Information has been captured in previous migrations and demonstrated that NSSDC was unprepared to capture non-routine Provenance Information during the current migration. The 'migration planning team' is working to define requirements for a provenance capture and management system. The resulting requirements, while not final, have resulted in the development of a prototype provenance database. The data migration may be resumed before a fully functional provenance database is deployed but only if interim procedures for consistent capture of Provenance Information in the absence of the final database can be put into place.

Although NSSDC's initial experience migrating legacy data has revealed a number of problems with the supporting applications and procedures, it has also proven the value of a comprehensive data migration plan incorporating OAIS concepts. This plan is providing a well documented framework addressing the various data sources, processes, and procedures being used and is facilitating a controlled process for addressing changes driven by operational experience.

3.4 Utility of OAIS Concepts

The NSSDC has adopted the OAIS Reference Model as the framework in which to describe its activities to its management and to others. To date it has been using the functional model at the level shown in Fig. 3, and for this migration it has begun looking at a partial breakout of the Administration function addressing Archival Information Update as shown in Fig. 5. Our ability to map our actual migration to these functions, as has been done in Fig. 7, greatly facilitates communication among NSSDC staff and is significantly improving the consistency of our system architecture. In particular, it is helping to clarify those data and metadata stores that need the most stringent preservation efforts from those that are less critical.

The OAIS Reference Model has extensive information modeling concepts. NSSDC has adopted the AIP, in the form of an AIU for actual implementation as a defined structure, as shown in Fig. 6. This has paid enormous dividends in clarifying the extent of the information to be preserved and in facilitating automation of both the ingest and the preservation processes. Future migrations of AIPs should be far less costly than the current migration. NSSDC has not yet focused on a specific implementation of the OAIS AIP's Archival Information Collection, although it does have collections of AIUs that are maintained by a database within the DIOnAS.

The NSSDC AIU contains Content Information in the form of science files and pointers (ADIDs) to their Representation Information. It also contains some PDI in the form of Fixity (checksums), Reference Information (AIU identifiers), and Provenance Information (states of data files before and after transformation to canonical forms). However there is no explicit Context Information within the single file AIP/AIU. Descriptions of space science experiments and spacecraft could be considered a part of the Provenance Information (see table 1), or they could be considered more generally as Context Information. Currently NSSDC may have such information in its Data Set Catalog and/or in other published documents. It is not explicitly tied to the AIP/AIU, but through the NIMS (Data Management) it can be tied to a collection of AIUs. This is not viewed as a deficiency of the OAIS PDI model but as a topic for a future NSSDC information architecture upgrade. In particular, NSSDC believes it needs to develop a preservation-focused document management system to accompany the current science file preservation system.

The legacy tape migration effort has provided a strong focus on capturing and preserving Provenance Information because its lack of consistency and depth from previous migrations has hindered much automation for this migration. The need for Provenance Information addressing archival processing is a valuable OAIS contribution. Its full implementation for an existing system such as NSSDC's is not trivial and is an issue that NSSDC will continue to address.

The NSSDC legacy tape migration has been described as a reversible-transformation migration because the transformation of science files into canonical form is fully reversible by application of an algorithm. It is certainly a migration because the intent is to preserve the information content. The result is a new version (first version) AIP/AIU for this information. It is not a new Edition because the intent is not to enhance the information content that NSSDC holds. This terminology does address the concepts NSSDC needs for this migration, but it can be confusing to those not well versed in the distinctions partly because the term 'version' is widely used with many meanings. It may be useful to consider modifiers for 'version' and possibly for 'edition' to help distinguish the special meanings within the migration context.

Referring to Fig. 4, the concepts of Packaging Information, Archival Storage Mapping, and AIP Identifier have been very useful. Our AIP single file implementation, shown in Fig. 6, uses the labels and pointers as part of the Packaging Information as they bind together the components of the information to be preserved. Actual implementations of storage systems must keep track of the locations of the AIPs, however they are defined, and in our case they are identified by an Archival Storage ID (ASID) that plays the role of the AIP Identifier shown in Fig. 4.

In summary, we find the OAIS Reference Model has provided very useful terms and concepts for which NSSDC's implementations have not yet taken full advantage. No significant problems applying them have been encountered. While this could change as our implementations proceed, we have no evidence to this affect currently.

3.5 Lessons Learned

1. *Importance of using Fixity Information (e.g., Checksums):* To paraphrase Murphy's Law: if something can possibly go wrong, it will. This applies to supposedly failure-tolerant systems such as RAID disk arrays, and to logical schemes/schedules that make backup copies. A new backup copy of a file that has already been corrupted (which can happen without the operators being aware of the problem) will just preserve the corrupted version of the file. The generation, storage, and comparison of checksums at every step is an invaluable tool for assuring that AIPs are correctly stored and retrieved.
2. *Importance of capturing Provenance Information:* The lack of sufficient, reliable and accessible Provenance Information from previous migrations has significantly complicated the current NSSDC migration. It has meant extensive manual efforts to track down apparent anomalies and has greatly increased the complexity of the migration plan.
3. *Importance of preserving supporting documentation:* Preservation of metadata is every bit as important as preservation of data. The need for nearly continuous migration of Content Information to new storage media is widely acknowledged. With the rapid development of storage technologies the short intervals between data migrations is driven as much by concerns about technology obsolescence as it is by concerns about media degradation. The frequent migration approach should be applied to Preservation Description Information and Representation Information, continuously identifying and employing appropriate new technologies for its storage. NSSDC's implementation of the OAIS Reference model addresses this need to the extent that some associated Preservation Description Information and Reference Information is incorporated into the single file AIP that is written to archive media. However, much associated PDI and Reference Information resides in multiple forms in various information systems. At the start of this latest data migration the Context, provenance, and Representation Information was stored in a variety of ways: on paper, on microfilm, in digital files on various media, and in numerous databases with designs stretching back decades. Collection of this information proved to be tremendous task. Additional resources applied over the years to upgrading the technologies used for PDI and Representation Information preservation and migration would have paid dividends in this migration.
4. *Importance of maintaining a detailed migration plan:* Unlike previous NSSDC migrations, this migration effort has included a detailed migration plan. We have found that it provides several benefits including getting review and consensus from multiple parties that have detailed experience to offer, identifying steps in the process where additional efforts to guard against information loss or corruption may be beneficial, and providing a roadmap for operations personnel that can be updated as experience dictates. It also becomes a top level Provenance Information source for all the science data and documentation involved.

References

1. Sawyer, D., Hills, H. K., McCaslin, P.: Preserving Access to Legacy Information Through Data Migrations at NSSDC: Experiences and Lessons Learned. Proceedings of PV-2004: Ensuring the Long Term Preservation and Adding Value to the Scientific and Technical Data, WPP-232, ESA/ESRIN, Frascati, Italy, 5-7 October (2004)
2. Reference Model for an Open Archival Information System (OAIS). Consultative Committee for Space Data Systems, Recommendation for Space Data System Standards, CCSDS 650.0-B-1, Blue Book, Issue 1. Washington, D.C.: CCSDS, (2002). [Equivalent to ISO 14721:2003.] http://public.ccsds.org/publications/archive/650x0b1.pdf
3. Sawyer, D.: The Open Archival Information System and the NSSDC: NSSDC News, December (2000): http://nssdc.gsfc.nasa.gov/nssdc_news/dec00/oais.html