Standards-based science data archiving at NASA's National Space Science Data Center

J.H. King, D.M. Sawyer, P.W. McCaslin*, R.W. Post*
National Space Science Data Center
NASA/Goddard Space Flight Center
Greenbelt, MD 20771
* also Raytheon/ITSS

Introduction

During the 35+ years of operations of the National Space Science Data Center (NSSDC) at NASA's Goddard Space Flight Center, NSSDC has accumulated an archive of over 4,000 distinct "data sets" from 1300+ instruments flown on 370+ mostly-NASA scientific spacecraft. For many years, NSSDC has held digital data (~25 TB as of 12/31/01) on magnetic disk for immediate user access (by ftp or higher functionality interfaces), on an optical disk jukebox for minutes-response-time user access, and/or on offline tapes, CD-ROMs and other media. The optical disk jukebox was retired last year, and a Digital Linear Tape (DLT) jukebox system was introduced to support the permanent archive function. NSSDC also has extensive holdings of imagery and other data on film of various form factors. During 2001, NSSDC added 3.3 TB of digital data in starts or extensions of 132 distinct data sets from 24 spacecraft. Virtually no non-digital data arrive at NSSDC these days.

In addition to its basic data stores, NSSDC has multiple information systems that facilitate NSSDC's management of the data archives and support the finding, understanding and use of the data mainly by the space science research community but also by the general public.

The twin foci of this meeting are the ensuring of long term data preservation and the adding of value to the data to ensure the convertibility of data to information. Accordingly, the twin foci of this presentation are the recently implemented standards-based approach to NSSDC permanent archiving and the management and evolution of the multiple information (metadata) systems at NSSDC that enable both NSSDC's data management and users' ability to convert data to information.

[NSSDC employs data standards in facilitating data access and display (e.g., the management and evolution of the Common Data Format that underlies many NASA and other spacecraft data systems (e.g., ISTP, Cluster) and multiple NSSDC systems (e.g., CDAWeb, OMNIWeb). NSSDC also provides "value added" service by creating cross-normalized multi-source data sets (e.g., OMNI) and such systems as CDAWeb and OMNIWeb. Readers may learn of these through the NSSDC web page at http://nssdc.gsfc.nasa.gov/, but they will not be the major foci of this paper.]

NSSDC Permanent Archiving

About three years ago, NSSDC committed to move from an older archiving environment with a number of "shortcomings" to a new environment exploiting newer data management standards and technologies. Among the shortcomings were:

- That NSSDC's major user-accessible nearline mass storage system (NDADS, a VMS-based optical disk jukebox system) was approaching the end of its 10-year life;

- That there were significant VMS dependencies in data structures and metadata in a world moving beyond VMS;

- That the large majority of NSSDC's digital data sets existed only on offline media, notably 9-track tapes, 3480 tape cartridges, 4-mm and 8-mm tape and CD-ROMs. The offline nature of the archive made it labor-intensive to maintain and meant these data were not network-accessible to users in an era where "accessible" and "network-accessible" have become virtually synonymous;

- That the offline data sets were in a wide range of formats including many largely obsolete vendor-specific binary formats. Also, they were accompanied by documentation of a broad quality spectrum (relative to completeness, clarity, etc.). The number of data sets from the early years of the space age is large. Many such data sets, although uniquely time-stamped, have had their scientific value largely negated by later higher-resolution data sets (and hence have little future access potential). Therefore, it was judged infeasible (unaffordable and not needed) to upgrade the usability of all legacy data sets equally.

- That the overall documentation of the media and files of a given data set was distributed across multiple NSSDC information systems that interacted only weakly if at all.

Three distinct thrusts were needed to convert from the state of three years ago to the future:

- First, a unix-based automated permanent archive was needed in which data files accompanied by at least minimal supporting material needed for their use had to be created and populated.
- Second, a conversion of formats from obsolete ones to modern ones, plus documentation reviews and upgrades, were needed for data sets having significant future use potential, while a documentation of obsolete formats was needed for those data sets of too little future-use potential to justify the expense of reformatting.
- Finally, NSSDC had to ensure an optimal organization of its metadata and that all relevant and available metadata would be easily accessible, along with corresponding data, both to NSSDC staffers managing the archive and to NSSDC's external users.

The last of these three is still a work in progress and is more fully described in the next section. The second has led to (1) the conversion of some formats from various binaries

to ASCII or to NSSDC's Common Data Format and to (2) the documentation of the bit-level structures of real and integer words of each of many vendor-specific binaries. This documentation should allow the reading of words/numbers from any old NSSDC data sets now judged of too-low future-access potential to warrant format conversion, although such usage will not be easy. This bit-level documentation is at http://nssdc.gsfc.nasa.gov/nssdc/formats/. The first of these is also a work in progress and is described in the remainder of this section.

NSSDC's new nearline permanent archive presently uses a 264-slot DLT (Digital Linear Tape) jukebox hosted by a Sun Enterprise 3000 computer. The data on the jukebox is directly accessible only to NSSDC staff. An external-user-accessible RAID magnetic disk array, currently attached to the same Sun computer, hosts much of the data written to the jukebox. As of this writing (early July, 2002), NSSDC has written about one terabyte of data to the RAID array and is awaiting delivery of a second unit with double the initial unit's capacity. A database knows of the location of data files on both the DLT jukebox and magnetic disk.

The DLT-archived data files are actually held as "Archival Information Packages" (AIP) where each AIP is a single file consisting of a data object (the bits from the data file), an attribute object and Consultative Committee for Space Data Systems (CCSDS)/ISO standard labels (http://www.ccsds.org/documents/pdf/CCSDS-620.0-B-2.1.pdf) . The labels include internationally unique CCSDS/ISO-recognized pointers (http://www.ccsds.org/documents/pdf/CCSDS-630.0-B-1.pdf) to additional information about the data. Currently this additional information base is not uniformly populated for NSSDC assigned pointers and is discussed in more detail in the following section. The AIPs are compatible with the Open Archival Information System (OAIS) reference model (http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-2.pdf) developed under CCSDS and ISO sponsorship with a leading role played by NSSDC's NOST (NASA/Science Office of Standards and Technologies).

The software enabling this new archive has been developed at NSSDC. On the one hand certain software modules accept a set of input data files as a given "job" and transform them to AIPs. The software ascertains files' attributes, employs processing algorithms to transform data file contents to canonically formatted data objects (see below), computes Cyclic Redundancy Check checksums, creates attribute objects and finally creates the AIPs. Other modules write a tar file, containing all the AIPs of a given job, to DLT tape. Yet other modules "split" most AIPs by writing their constituent data and attribute objects to externally accessible magnetic disk files. Finally entries are made to a database about the AIPs and their constituent files including their locations in the jukebox and on magnetic disk.

While various modules have different names at NSSDC, the name DIOnAS (Data Ingest and Online Access System) will be applied to the full software set for this paper. The database about the AIPs and files is called the DIOnAS database.

To date, the DIOnAS software has been used to create AIPs from the data files held by the retired VMS-based NSSDC Data Archive and Distribution System (NDADS). Much information from the NDADS Files 11 Extended Attribute Records (XAR) was captured into the AIP attribute objects. Certain data sets (e.g., those with files having variable length records) required conversions of the original data byte streams to conform to the appropriate NSSDC-defined canonical form. For example, for binary files with variable record lengths (as known to DIOnAS from Files 11 XARs), a 2-byte integer field, unsigned and formatted Big Endian, was inserted to prefix each record. That this was done to a data file is documented in its companion attribute object. This and many other aspects of NSSDC's new data management approaches are discussed more fully at http://nssdc.gsfc.nasa.gov/nssdc_news/dec00/dec00_toc.html.

Generalizations to the DIOnAS software are currently being developed to handle the case of AIPs with multiple data objects. This will be important for data sets of paired or otherwise mated "data" files. For example, NSSDC has several cases of "flat file" data sets from UCLA (University of California at Los Angeles) wherein an ASCII header file accompanies each binary data file. From a DIOnAS perspective, these are two data files that need to be held together.

NSSDC is currently preparing to ingest data from its legacy magnetic tape archive to DIOnAS. The extensions needed for the attribute objects are being assessed. Possible generalizations to the range of "canonical files" are also being assessed. The range of NSSDC tape data sets is being reviewed to exclude from DIOnAS-archiving any which contain no unique science data and also to identify what small subset might warrant reformatting or other upgrades to facilitate future usability. It is intended that the DIOnAS processing of tape data sets, which may involve about 10,000 tapes and 2,000 distinct data sets (with unique characteristics, documentation packages, etc.), will be as automated as possible and will require no more than 2 years to complete once in production.

Another very important aspect of NSSDC's transition to OAIS-specified archiving is NSSDC's working with data suppliers to enable them to create AIPs for submission to NSSDC. This significantly facilitates NSSDC's data ingest activity. To date, this has been pursued with the IMAGE (Imager for Magnetopause-to-Aurora Global Exploration) spacecraft managed at NASA/GSFC. NSSDC provided to the IMAGE team an upgraded version of its Package Generator Utility software (a module previously included under "DIOnAS"). With this software and a number of configuration files to set key values, the IMAGE Science and Missions Operations Center was able to compute needed attributes, create the attribute objects, compute other packaging information, create the CCSDS/ISO standardized labels, and finally bundle these with the data file content into AIPs for transmitting to NSSDC.

For completeness we note that data currently ftp-accessible to the world from NSSDC (from "nssdcftp"), but not yet processed into AIPs, will be processed by DIOnAS. This will further populate the new permanent archive and will give NSSDC the option of building a higher level interface to all its ftp-accessible data since their whereabouts on

magnetic disk will be captured in the DIOnAS database. We will similarly DIOnAS-process CDF-formatted data from CDAWeb whose science content are not otherwise managed under DIOnAS.

Value-added services and metadata

"Value-added" can mean many different things. NSSDC provides value-added data sets by cross-normalizing and merging single-source data sets, of which the prime example is the 1963-2002 OMNI data set of hourly solar wind data and some other data. NSSDC provides value-added services allowing concurrent access to many data sets with graphical browse and subsetting capabilities, of which CDAWeb and OMNIWeb are prime examples. NSSDC provides largely common but specific-dataset-tuned graphical browse and subset interfaces to a subset of its data holdings otherwise only ftp-accessible; this family of interfaces is known as "FTPBrowse."

For the remainder of this section we take "value-added" to mean only the management of data-accompanying metadata and other material needed to transform data into information.

It is useful to consider distinctions among data sets, files and media volumes. A data set is a set of like files holding data in one or more records per file. The files typically have common technical characteristics and have data at a given "processing level" acquired at successive points or intervals in the relevant independent variable space (e.g., time, or look direction to a remote object). The files and records of a given data set may have data from one or multiple sources (e.g., from multiple instruments on a given spacecraft).

[To digress, the concept of processing level is important to appreciate. Scientific instruments generate and provide currents or voltages or the like which are digitized (the raw data). The raw data must then be processed, sometimes through a series of steps, to yield the "physical parameters" (e.g., magnetic field values, particle fluxes, plasma densities, calibrated images, etc.) which are used in scientific analyses. For much of its existence, NSSDC has archived physical parameter data sets only, judging that raw data would be unusable by others and that most of the science potential of the raw data had been reliably transformed to the physical parameter level. With space science instruments getting increasingly sophisticated and multi-modal, it is becoming increasingly unlikely that most of the science potential of raw data will be reliably transformed to the physical parameter level. Hence, there is now an emphasis on archiving, in addition to physical parameter data sets, the best-annotated and organized data that are not yet irreversibly transformed. This is close to the "raw data" level (cf. http://nssdc.gsfc.nasa.gov/nssdc/data_retention.html). Ensuring the long-term correct and independent use of such low level data requires a great deal of documentation and supporting material, often more than is provided to archives with such data.]

Relative to data media, most legacy data sets were provided to NSSDC on tape media, although in recent years data have arrived at NSSDC electronically or on CD's and DVD's. Not including second copies, the files of a given data set may be held on one or

multiple media volumes, or may reside on a media volume along with the files of distinctly different data sets. A media volume "supports" (not "belongs to") a data set. In this media, file, data set framework, the following types of information must be captured and managed:

- Technical information about files and media (file organization, record lengths and separators, tape densities, etc.);
- Inventory information: what files and media exist for a given data set and where are they;
- Transaction and other provenance information about the files and media;
- Record structure and content information enabling the interpretation of the record's bits and bytes as a series of pixel intensities, values of named variables, or other values;
- Scientific information about the contents of the data records. Here we address such questions as:

    - How were the physical parameters of this data set produced from lower level data, or how can a user produce physical parameters from the low level data of this data set?
    - What are the assumptions made in the data processing?
    - What are the known errors in the data, under what conditions do they arise and how are they recognized?
    - What are the uncertainties in the various physical parameters and how do their magnitudes depend on values of other dependent or independent variables?
    - What range of science studies can the data reliably be applied to?
    - What scientific results have resulted from prior analyses of the data?

NSSDC has multiple databases and other information collections capturing much of the above information. The following paragraphs identify these databases and collections, and outline some of the issues facing NSSDC as it strives to define and implement an optimally integrated metadata and data environment. Except where noted, the databases are relational and use Oracle.

The JEDS (Java-based Experiments, Data sets and Spacecraft) database contains information about data sets as a whole, as well as information about the sources of the data, typically spacecraft and their instruments. Entries are typically created prior to first data arrival at NSSDC. From an archive management perspective, JEDS knows:

- data set identifiers;
- plans and status for archiving this data set;
- data arrival-at-NSSDC dates and types and numbers of media for each date;
- submitter name and other contacts;
- supporting material to send, with data, to requesters of data;
- numbers of requests for this data set, by year;
- etc.

From an end-user perspective, JEDS knows:

- data set time span,

- discipline-oriented keywords enabling a narrowing of data searches,
- various attributes (resolutions, data processing level, parameters contained, format, etc.) as summarized in a free-text description,
- etc.

For most data sets, JEDS does not contain all the "scientific information" identified above as desirable in understanding the data processing already applied to, or needed by, the data. Such information, when available for legacy data sets, are typically found in publications or other reports of data generators/providers. These publications are identified in NSSDC's Technical Reference File (TRF) database that uniquely links publications to experiments or data sets. They may also be replicated in the data set catalogs discussed below. The TRF also links science papers uniquely to an experiment or data set thereby allowing data requesters to read some of the results of prior scientific analyses. A major challenge for scientists preparing low processing level data for archiving is the preparation of adequate documentation and other supporting material (e.g., software) that will be survivable over time.

For metadata about individual digital data media, NSSDC has the IDA (Interactive Digital Archive) database that knows:
- what data set(s) a given volume supports ,
- the location of the volume and its backup(s),
- the transaction history of the volume,
- technical attributes such as tape densities and block sizes, format (e.g., binary vs. ASCII) and, if binary, the vendor,
- time span of the data on the volume, etc.

The IDA file is currently being evolved to become Java-based and to better support a broader range of media types.

The DIOnAS database, mentioned earlier in connection with the new NSSDC approach to nearline permanent archiving, captures information about Archive Information Packages and their constituent data and attribute files. DIOnAS links these to NSSDC data sets, and captures AIP creation history information, technical information (file types, sizes, CRC checksums, etc.), location information (on DLT permanent archive and on customer-accessible RAID magnetic disk) and time span information.

In addition to the above three databases which characterize at a high level NSSDC data sets and their sources, the digital media and the DIOnAS-ingested data files of NSSDC, there are three major collections of further information about NSSDC data. These primarily give more detailed information needed to understand and correctly use the data. They are readme's in the ftp-accessible area, dataset catalogs, and records of the NSSDC-maintained CCSDS Control Authority database of registered data set descriptions.

The readme's are an ad hoc collection of dataset-specific descriptive texts giving record format information and variable levels of documentation on processing histories, uncertainties, etc. These are relevant to the subset of NSSDC data available from

ftp://nssdcftp.gsfc.nasa.gov/spacecraft_data/, some of which are known to the DIOnAS database and others not yet DIOnAS-known.

The data set catalogs are collections of pages per data set accumulated by NSSDC over its history. These catalogs include format statements, tape organization and inventory information, partial tape dumps, and miscellaneous reports and communications between NSSDC staff and data providers all contributing to data understanding and usability. About 4 years ago, NSSDC scanned the contents of these hard copy catalogs and wrote the resultant GIF files to CD-ROMs indexed by NSSDC data set ID's. These catalogs, while somewhat ad hoc and varying in detail, are NSSDC's most comprehensive collections of information about the contents of data sets in a detail greater than that found in the JEDS database described above.

The NASA/Science Office of Standards and Technologies (NOST) hosted by NSSDC plays several key roles relative to the international CCSDS. One such role is to be the "lead node" in an ensemble of CCSDS Control Authority Offices (CAO) which capture and "register" descriptions of data sets, and assign internationally unique CCSDS/ISO identifiers to these descriptions. The basic concept is that distributed data files can point, using these identifiers, to such descriptions held available and accessible by quasi-permanent entities. NOST's CAO database at http://ssdoo.gsfc.nasa.gov/nost/cao_nssd-adids.html holds descriptions (rich in format and technical information but non-uniform in terms of data processing information, etc.) for ~100 data sets, a small subset of the digital data sets at NSSDC.

One final virtual collection of information should be mentioned. Just as data availability is becoming increasingly distributed, so too is information about the distributed data. It is important to ensure data usability on a time scale long relative to the durations over which individual data creators/providers may make data available. Therefore the distributed information about data, in addition to the data themselves, need to be swept up by relevant archives.

Conclusion

Except for data made accessible by NSSDC's "active archive" partners within NASA, NSSDC is making virtually all its currently important space science data externally accessible via ftp and via a series of higher-functionality interfaces. However, NSSDC's very longevity has led to a highly heterogeneous environment relative to data formats and media and to metadata systems. NSSDC has taken important first steps to transform this legacy heterogeneous environment to a modern and integrated NSSDC Data and Information System encompassing its full archive reaching back to the dawn of the space age. Data preservation will be ensured. Convertibility of data to information will only be constrained by the adequacy of the supporting material provided with the data to NSSDC. That most legacy data are "physical parameter" data helps to minimize this problem. That more nearly raw data are being archived for recent missions places a major requirement on data providers to prepare supporting material to enable the long-term correct and provider-independent conversion of data to information.