THE NATIONAL SPACE SCIENCE DATA CENTER
- AN OPERATIONAL PERSPECTIVE -

Ronald Blitstein, ST Systems Corporation/NSSDC
Dr. James L. Green, NSSDC

ABSTRACT

The National Space Science Data Center (NSSDC) manages over 110,000 data tapes with over 4,000 data sets. The size of the digital archive is approximately 6,000 GBytes and is expected to grow to more than 28,000 GBytes by 1995. The NSSDC is involved in several initiatives to better serve the scientific community and improve the management of current and future data holdings. These initiatives address the need to manage data to ensure ready access by the user and manage the media to ensure continuing accessibility and integrity of the data.

This paper will present an operational view of the NSSDC, outlining current policies and procedures that have been implemented to ensure the effective use of available resources to support service and mission goals, and maintain compliance with prescribed data management directives.

INTRODUCTION

The NSSDC is a heterogeneous data archive and distribution center operating in a dramatically changing scientific and technological environment. For most of its thirty year history, it has operated as a batch-oriented library providing custom support for the ingest and distribution of data. Its rate of growth, as measured in volume of data held and request activity have been steady but modest when compared with expected future activity. The NSSDC responds to approximately 3000 requests for data per year. Some requests are supported through on-line or near-line capabilities, but many are filled through the replication and distribution of data tapes or images. During the past five years, over 8500 individual requestors have been provided data, with over thirty percent of them repeat customers. The average volume of data distributed with each request has increased dramatically from 900 MB to 1500 MB during this period. As a data archive, the NSSDC has established policies for media and data management that strive to ensure the continued integrity and availability of its data holdings. These policies cover the ingest, archive, maintenance, and migration of data, as well as the management of the supporting documentation, software, and metadata necessary to meaningfully access and use the data.

INGEST AND ARCHIVE ENVIRONMENT

All data currently received at or generated by the NSSDC enter the archive through a data ingest process. This process requires that two copies of each data volume are made to the current "technology pair" (described below) of media identified by the data center. A copy is retained locally and the other sent off site to the backup archive currently maintained at the Washington National Records Center (WNRC). After the copies are made and validated, the original data volume is not retained. This procedure ensures that the data are written to new, archive-quality media and enables the NSSDC to accurately track each creation date.

As an integral part of the ingest process, entries are made in catalog and inventory data bases. These data bases track spacecraft, experiment, and data set attributes of value to researchers and browsers of the NSSDC's data holdings. Additionally, media-specific information is entered which enables the data center to locate and retrieve desired data files and manage media characteristics, usage, and maintenance actions necessary to ensure the continued integrity and technological currency of the media.

The concept of "technology pair" is one developed by the NSSDC in response to the accelerated obsolescence of recording media resulting from the rapid development and introduction of new storage technologies. The frequency with which new technologies are being introduced makes it difficult to identify and evaluate the archivability of any one media/format before another enters the marketplace. The concept defines the archivability of new media in terms of several factors. These criteria evaluate the appropriateness of any media through the extended lifecycle expected of an archive facility.

- Degree of standardization
- Availability of hardware/software
- Error detection/correction
- Integrity as a function of age
- Capacity
- Transfer rate
- Compatibility with robotic load devices

The NSSDC has identified two technologies, 9-track/6250 bpi and IBM 3480 cartridges, as its current media of choice for institutional archival purposes. Additionally, the data center has installed capability to support near-line archival of data on 12-inch WORM platters, each with a capacity of 2 GB or 6.2 GB. As a new medium is selected, acquired, and installed to support the full spectrum of operational requirements, it will replace the oldest technology then in place (eg. 9-track). Together with the other currently supported medium (eg.

IBM 3480 cartridge), will then comprise the new technology pair. This conservative approach ensures that unforseen problems with new media do not jeopardize the total archive holdings, and that orderly migration of data from one media to another is supported.

Historically, the NSSDC was often resource constrained and the ability to generate and maintain a backup copy of all data was often beyond its reach. This occurred during periods of relatively low rate of change in the technological environment, and the "push" from the commercial sector to adopt new media technologies did not exist. Most of the data ingested at the NSSDC during this period came from missions in progress, and the project scientists provided back up capability with their copies of the data. Today's policies reflect a new philosophy in stewardship, where the total responsibility for data management lies with the primary archive data center. To implement these policies, the NSSDC has taken actions to maintain the integrity of its current holdings and prepare for the massive amount of data from future missions. These actions include a comprehensive data restoration effort, migration of data to near-line accessible media, improvement in research tools, and proactive involvement in the data management planning of future missions.

DATA RESTORATION

Through its data restoration effort, the NSSDC is currently migrating its older data holdings to new technology pairs. Success in this effort has been outstanding. To date, data recorded on approximately 25 percent of the media volumes in the archive have been migrated with greater than 98 percent of the integrity preserved. These media volumes were in 7-track and low density 9-track formats, many 20 years old or more. The success of this effort was unexpected, and many feared that the data would be "lost on earth". But as a result of basically sound storage procedures and the development of an appropriate set of procedures, a more optimistic view is emerging.

The development of data and media management guidelines is very important. A great deal of the success in the data restoration effort can be attributed to the environmental conditions in which the data were archived. NSSDC is continually reviewing its policies in these areas to gain increasing benefit from advances in technology and collective knowledge. Current areas of interest include:

- Pre-certification of archival tapes
- Use of specialized off-site archive facilities
- Increased use of robotic near-line storage for media maintenance
- Error detection and correction
- Data compression

NEAR-LINE DATA ACCESS

The NSSDC is responding to an ever increasing number of scientific inquiries by placing requested data in an public-access retrieval account on the NSI wide area network. Two strategies are being employed to provide this high level of data retrieval, near-line and on-line mass data storage. An example of the success obtainable from effectively managed near-line storage can be found in the data management for the International Ultraviolet Explorer (IUE) mission. The NSSDC has loaded all the IUE data, consisting of over 70,000 unique star images and spectra, in the IBM 3850 Mass Store device operated by the NASA Space and Earth Sciences Computer Center. An interactive system on the NSSDC VAX cluster allows remote users to order data from the electronic Merged Observer Log. This order is processed off-line, where the data of interest are located on the mass store, transferred over the network from the IBM to a public VAX account and a message sent to the requestor that the requested data are ready for retrieval. This process typically takes less than one work day to complete. The use of the mass store is being phased out, and the IUE data will soon be available through a automatic near-line retrieval capability on the NASA Data Archive and Delivery Service (NDADS) optical disk juke box. In its final configuration, NDADS will manage data, meta data, and documentation, all stored within the same system.

On-line access of NSSDC-held data is currently possible for smaller, often requested data sets. The NSSDC On-line Data and Information Services (NODIS) provides public access to data sets that can be researched, viewed, and retrieved by a requestor during a single interactive NSI-net session. Both Earth and space science data are currently available in this manner, including Nimbus Ozone and merged OMNI data, as well as access to the NASA Master Directory.

RESEARCH TOOLS

Proper data management is only part of the picture. To facilitate the research of data, the meta data needs to be afforded an equal level of support. The researcher needs to know of the existence of possible data of interest, and how to use it once located. Through tools developed by the NSSDC, the researcher is able to spend less time and effort on these actions, and more time doing scientific research on the data. NSSDC is involved in the development and dissemination of NASA-wide directory and catalog information, and has installed versions of its Master Directory in numerous data centers throughout the world.

Once located, the data and their formats must be understood if useful research is to be conducted. NSSDC has promoted the correlative use of data across missions through sponsorship of Coordinated Data Analysis Workshops

(CDAWs). In support of these workshops, Common Data Format (CDF) tools have been developed and implemented to allow the researcher to focus on the content of the data and develop meaningful relationships among data having different resolutions and areas of coverage. With CDAW 9.4 held recently, the latest in CDF and graphical display tools were demonstrated.

Another important research tool is data browse. Browse capabilities have been built into the data organization strategies used extensively for data available on CD-ROM. The NSSDC currently maintains approximately two dozen titles on this medium, supporting research in the Earth, space, and planetary sciences.

PROACTIVE DATA MANAGEMENT FOR FUTURE MISSIONS

Data archives have a responsibility to manage future as well as current data holdings. Through its experiences in data restoration, on-line access, and tool development, the NSSDC is sensitized to problem-avoidance strategies for future missions. The data center has developed a cost model to estimate the resource requirements of data ingest, archival, management, and distribution. It is using data from this model to identify future missions requirements for inclusion in the appropriate Project Data Management Plan (PDMP). The PDMP is a multi-lateral agreement that is executed for all future NASA missions. Data management issues addressed by this plan includes the level of service to be provided by the archive, the nature, volume, and frequency of the data to be ingested, the type of media, expected request activity, etc. This process is enabling the NSSDC to reliably estimate future costs for these missions; a critical element when one considers the very large volumes of data that missions such as EOS and the Space Station will generate. PDMPs have been developed for several of the newer missions, including Magellan and Gamma Ray Observatory.

SUMMARY

The course of future scientific research can not be predicted, nor can the data needs of this research. As a national data archive, the NSSDC must not only ensure the continued integrity of the data intrusted to it, but must also ensure the continuing evolution in its ability to provide the correct data to the user in the correct way. As the volume of its data holdings increases, the shift from specialized service to a uniform spectrum of generic services must continue. The NSSDC is pursuing this goal through various initiatives in mass storage, networks, media management, tool development, and standards advocacy.

As is often true, the hardware capabilities and the technological sophistication necessary for very large mass storage systems is rapidly being

developed. In short duration project environments, the selection, installation, and implementation of viable systems is relatively easy. But in the view of an archival data center, such as the NSSDC, the massive volume of non-homogeneous data from hundreds of missions for which it is responsible make this a very difficult procedure. The selection of high capacity storage media must be accompanied with corresponding strategies to ensure the integrity of the data for many years. The current media transfer rates and the requirement for the generation of backup copies effectively doubles the volume of data to be managed. Higher density storage without accompanying capabilities in robust error detection and correction that provide lossless recovery of data may be inappropriate for permanent archives. Frequent migrations of data from one media to another, especially if accomplished in a true automated fashion, are attractive alternatives to the manual processes widely used today, but pose enormous requirements of inventory and catalog data bases that have visibility across all the various archive systems in use in any facility (or even across facilities in a distributed archive environment).

NSSDC NATIONAL SPACE SCIENCE DATA CENTER

The National Space Science Data Center
- An Operational Perspective -

Ronald Blitstein, ST Systems Corporation/NSSDC
Dr. James L. Green, NASA/NSSDC

July 25, 1991

NSSDC NATIONAL SPACE SCIENCE DATA CENTER

An Operational Perspective

- Introduction
- Ingest and Archive Environment
- Technology Pair
- Data Stewardship
- Data Restoration
- Migration of Data to Network Accessible Media
- Improvement in Research Tools
- Proactive Involvement in Data Management Planning for Future Missions

NSSDC NATIONAL SPACE SCIENCE DATA CENTER

Introduction

The NSSDC is a heterogeneous data archive and distribution facility responsible for:

- Ingest
- Archive
- Distribution

Data Holdings:

- 110,000 data tapes
- 4000 data sets
- Six TBytes

NSSDC NATIONAL SPACE SCIENCE DATA CENTER

Ingest and Archive Environment

Data ingest Procedures

- Routine generation of two copies of all incoming data
- Use of precertified tapes
- Off site storage of backup

Metadata management to ensure useability of data

- Directory and catalog information
- Format information
- Inventory information

NSSDC NATIONAL SPACE SCIENCE DATA CENTER

Technology Pair

Purpose:

- The selection of two media technologies for data archival

Criteria:

- Degree of standardization
- Availability of hardware/software
- Error detection/correction
- Integrity as a function of age
- Capacity
- Transfer rate
- Compatibility with robotic load devices

NSSDC NATIONAL SPACE SCIENCE DATA CENTER

Technology Pair (cont.)

Current technology pair at NSSDC:

- IBM 3480 cartridge (primary)
- 9-track/6250 bpi (backup)

Migration strategy:

- Select new media technology
- Identify new technology pair
- Migrate data to new media
- Discontinue support for older media

NSSDC NATIONAL SPACE SCIENCE DATA CENTER

Data Stewardship

Charter:

- The total responsibility for data and media management lies with the primary archive data center

The NSSDC is addressing this responsibility through:

- Data restoration
- Migration of data to network accessible media
- Improvement in research tools
- Proactive involvement in data management planning for future missions

NSSDC NATIONAL SPACE SCIENCE DATA CENTER

Data Restoration

Purpose:

- The systematic migration of data from older media to current technology pair

Older media:

- 7-track
- 9-track/low density
- Many tapes are greater than 20 years old

Status:

- Twenty-five percent of media volumes completed
- Greater than 98 percent of data integrity preserved

NSSDC NATIONAL SPACE SCIENCE DATA CENTER

Data Restoration (cont.)

Experience:

- Storage environment of critical importance
- Frequent cleaning of drives necessary to avoid tape damage
- Problems have been manageable
- Use of precertified tapes recommended

NSSDC NATIONAL SPACE SCIENCE DATA CENTER

Migration of Data to Network Accessible Media

Increase public access retrieval on NSI wide area network

- Near-line storage - IUE example
  - Use of IBM 3850 mass store to manage over 70,000 images
  - Image size is typically one MByte
  - VAX interface to provide image ordering capability
  - Manual extraction and staging of data on FTP account
- On-line storage - NODIS
  - Used for small, often requested data
  - Captured VAX account provides browse and retrieval

NSSDC NATIONAL SPACE SCIENCE DATA CENTER

Migration of Data to Network Accessible Media (cont.)

Other Initiatives:

- NASA Data Archive and Delivery System (NDADS) optical disk juke box
- Mass Data Storage and Delivery System (MDSDS) six TByte system
- Increased network bandwidth

NSSDC NATIONAL SPACE SCIENCE DATA CENTER

Improvement in Research Tools

Through the use of tools the researcher is able to spend more time analyzing the data content, rather than the data formats

- Coordinated Data Analysis Workshops (CDAW)
- Common Data Format (CDF)
- Data browse

NSSDC NATIONAL SPACE SCIENCE DATA CENTER

Proactive Involvement In Data Management Planning for Future Missions

Adoption of problem avoidance strategies to anticipate future requirements:

- Cost model developed to estimate resource requirements for data archival
- Early participation in Project Data Management Plans (PDMP)
- Research into auto ingest of future data deliveries
- Involvement in EOS data system development/evaluation
- Sponsorship of data management conferences and committees
- Participation in the establishment of standards for data and media

NSSDC NATIONAL SPACE SCIENCE DATA CENTER

Summary

The NSSDC is a heterogeneous data archive and distribution center operating in a dramatically changing scientific and technological environment.

A conservative but forward-looking approach to data management is necessary to avoid situations that could jeopardize the integrity and availability of its data holdings.

Its 25 years experience in data and media management enable the NSSDC to proactively participate in the challenges of the EOS era.