Policies and Procedures for Accessing Archived NASA Data via the Web

Nathan James
National Space Science Data Center
NASA/GSFC
Greenbelt, MD USA
Nate.James@nasa.gov

Abstract— **The National Space Science Data Center (NSSDC) was established by NASA to provide for the preservation and dissemination of scientific data from NASA missions. This white paper will address the NSSDC policies that govern data preservation and dissemination and the various methods of accessing NSSDC-archived data via the web.**

I. Introduction

The National Space Science Data Center (NSSDC) was established by NASA to provide for the preservation and dissemination of scientific data from NASA missions. Specifically, it serves as a multidisciplinary archive supporting the permanent storage of NASA's space science data and related metadata. NSSDC has established long-term storage agreements with other discipline-specific repositories for NASA's space science data; these repositories are called "active archives." In addition, with mediation from the active archives, NSSDC also collects data from NASA space flight missions, and individual principal investigators. Once acquired, NSSDC stands ready to make requested archived data accessible to both the research community and the general public, as resources allow. The preferred method of access to these long-term data stores is via the Internet. This white paper will address the NSSDC policies that govern data preservation and dissemination and the various methods of accessing NSSDC-archived data via the web.

II. NSSDC Archived Data Policy

As NSSDC acquires scientific data, its policy is to require that the data be accompanied by sufficient documentation to ensure that the scientific data are independently understandable, preservable, and usable into the indefinite future. This documentation should include detailed descriptions of the data, such as the experiment(s) that generated it, the spacecraft on which the experiment(s) flew, and the algorithms and formats used to store the data. The more detailed the data descriptions, the more efficiently the data can be preserved and the easier it will be to make the data more searchable and immediately accessible.

NSSDC policy also requires that archived data are preserved on media that is both reliable and accessible, because the integrity of the media affects the integrity of the data. Also, the software used to transfer data from one medium to another must provide error checking to ensure data integrity. NSSDC wants to use reliable high-capacity media that can be made more easily accessible electronically. Enforcing these policies helps to ensure the usefulness of the data, thereby preserving their value to the science community and the public into the future.

III. NSSDC Archived Data

NSSDC archives data from several disciplines: astrophysics, solar and space plasma physics, and lunar and planetary science. By the end of 2009, NSSDC was managing 4425 distinct data collections and accompanying documentation packages [1]. Table 1 indicates the disciplines from which these collections come and whether they are digital or not. Space physics is the dominant discipline by dataset counts, accounting for nearly half of NSSDC's holdings. This reflects two trends. First, in its early years, NASA launched a preponderance of space physics missions. Second, space physics spacecraft typically carry more independent experiments than do astrophysics missions.

TABLE I. COUNTS OF NSSDC DATASETS

Discipline	*Digital*	*Non-Digital*	*Total*
Astronomy	286	129	415
Space/Solar Phys	1,312	743	2,055
Planetary	1,137	765	1,902
Earth Science*	115	135	250
Other (including Ephemeris)	132	443	575
Total Datasets (excludes duplicate entries)	**2,777**	**2,073**	**4,850**

*Earth science data was passed to the Earth Observing System in 1995 and the remaining data collections are now being migrated to the Earth Science Data and Information System (ESDIS).

Data are actively being moved from NSSDC's traditional offline archive to near-line storage devices called super digital linear tape (SDLT) jukeboxes. These jukeboxes are attached to unix and linux servers. Much of these data are newly archived in Archive Information Packages (AIPs), which hold data files and companion attribute files and are media-independent and platform-independent. These are defined as per the AIP concept of the ISO/CCSDS Open Archival Information System reference model [2]. About half of the data stored in AIPs is made network-accessible via FTP for the convenience of a growing segment of the user community.

A. Searching NSSDC's Data Archive via the Web

The NSSDC Information Management System (NIMS) encompasses most of the separate databases that NSSDC has used to track data and information through the years [3]. The NSSDC has a long-term goal of incorporating its offline data inventory system into NIMS, and a major effort for this is underway. NIMS identifies virtually all launched spacecraft, the experiments carried by many of these spacecraft, and the resultant datasets primarily as archived at NSSDC.

The ability to search for NSSDC's data holdings based on spacecraft name and discipline has been in place since the mid-1980s, with the current web-accessible incarnation being available since the mid-1990s. This is supported by the metadata maintained in a database.

The database infrastructure, referred to as NIMS, contains information primarily of use for describing and searching for data holdings. In addition, NIMS also has the capability of handling some administrative functions that pertain to the tracking of archiving status of various flight projects.

A core piece of NIMS is the information about the data collections archived at the NSSDC. This segment describes the data so that interested parties can determine what would be of most use to their particular needs. Ancillary information about the data, including descriptions of the experiment(s) that generated it and the spacecraft on which the experiment(s) flew, are also available. Other ancillary information includes the publications generated from or describing additional details about the data collections, experiments, and spacecraft. These parts of NIMS are primarily for use as discriminators and for searching.

Other partitions in NIMS (as well as some parts in the core set of partitions mentioned above) are used primarily for administrative purposes. This includes, for example, information on the number of requests for a given data collection, information about requestors, and the total volume of data in a collection.

The NIMS database is moderately sized in terms of the number of records it holds: roughly 6600 spacecraft, 5400 experiments, and 5700 data collections with about 48,000 publications, 61,000 people, and 68,000 current digital media holdings. Some parts of the database are quite simple in nature, with people information spanning only five tables and a few dozen fields. Others are very complex. Spacecraft information is among the most complex in the database, comprised of 24 main tables with roughly 50 supporting (validation) tables and dozens of accompanying stored procedures to ensure consistent and reliable data.

This portion of the database is the source of information for many of NSSDC's information-rich web pages. The NSSDC Master Catalog (NMC) dynamically generates such web pages so that the latest information is presented to the user. A number of discipline and project pages are based on information derived from NIMS or utilize the NMC to generate such information.

B. Accessing NSSDC's Archived Data via the Web

The primary means of external access to the information about NSSDC's data holdings is via a web interface, the NMC. It consists of a suite of Java code that queries the database in various ways to produce dynamically generated web pages. This interface was completely re-engineered in 2006. Version 4.0.0 was delivered in November 2007 with a recent upgrade to V4.0.9 in May 2009. In addition to a web form which supports queries by spacecraft name and so on, there are now additional forms to query for experiments and data collections directly, thereby relieving the need to drill down to data information via the spacecraft.

Users can search through the NMC and locate data of interest, in many cases finding links to online sources of data holdings, both at NSSDC and elsewhere. This linking ability is supported mainly for highly sought-after collections, but the number of collections that are accessible in this manner has continued to grow at a steady pace.

The main portal to the NMC is via http://nssdc.gsfc.nasa.gov/nmc/. This page provides access to query forms for spacecraft, experiments, data collections, people, publications, maps, and lunar/planetary events. Once a given record is displayed, there are often links that allow cross-queries of the information from the database (e.g., between

publications and data collections, data collections and spacecraft, etc.) or which provide access to related resources, both at NSSDC and elsewhere.

Another means of external access to NSSDC information is via the Space Physics Archive Search and Extract (SPASE) Registry [2][3]. NASA's space physics research community continues to demand improved methods and procedures to facilitate finding, retrieving, formatting, and obtaining basic information about data essential for their research. The SPASE Data Model provides a structure and a basic set of terms organized in a simple and homogeneous way, to facilitate access to Solar and Space Physics resources.

NSSDC's SPASE registry currently contains access to information about the spacecraft (observatories), experiments (instruments), and people tracked in NIMS in a SPASE-compliant (XML) format. Version 2.0.0 of this interface was released shortly following the release of the V2.0.0 SPASE Data Model in April 2009. The SPASE registry software is based on the NMC software with some special code to support the generation of special information for SPASE compliance and to general XML rather than XHTML.

The portal to the NSSDC SPASE registry is via http://nssdc.gsfc.nasa.gov/spase/. This page provides access to query forms for the three SPASE resources currently supported. Options are available to display the results in raw XML (which can then be used in a virtual observatory) or in a more human-readable format.

C. *Staging and Distributing NSSDC's Archived Data via the Web*

NSSDC manages data both in on-line mode and in off-line stores of tapes, films and other media. The preferred method for accessing NSSDC's 1 TB online data archive is via FTP. The NSSDC home page provides FTP pathways to a range of data files maintained permanently on a dedicated NSSDC FTP server machine (NSSDCFTP). Special web form interfaces to specific datasets (or groups thereof) are also available.

The Staging Process. The experienced NSSDC user may choose to navigate the FTP directory structure to find data that he or she knows are stored there. But the typical researcher looking for NSSDC archived data would search online via the NMC. If the user finds that the data of interest are available via FTP, then NMC would provide a link to the appropriate directory where the data are "staged" on NSSDCFTP. If the data are not available online, a request is submitted either by phone or email to NSSDC's Coordinated Requests and User Support Office (CRUSO). This office is staffed by a contractor. After assessing the task, CRUSO provides a cost estimate, which could range from $35 to $75 per hour. Though the contractor sets the processing rate, it is sometimes waived by the government.

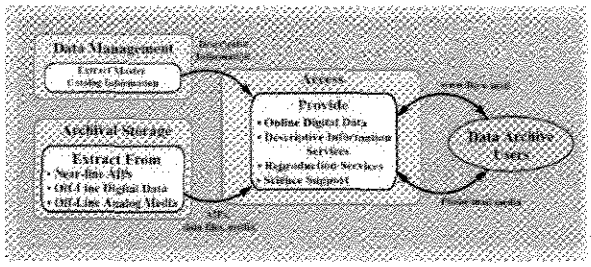

Figure 1 - NSSDC Data and Information Flow Diagram

Upon receipt of the requestor's approval to continue processing, CRUSO personnel would then work together with curation scientists to identify the requested data, locate them, and determine their storage format. The diagram in Figure 1 depicts the data flow environment. If the data collection resides in near-line storage, it would then be copied to the staging area on NSSDCFTP. Requested data not found in near-line storage would be digitized and also staged for access via FTP. The most popular data collections are prioritized for being stored in AIPs.

The Distribution Process. As NSSDC's digital archive grows, data distribution via FTP is preferred due to its immediacy and to NSSDC budget limitations. Though various modes of media are used for data distribution, FTP is the dominant mode.

IV. Conclusion

NSSDC is committed to managing long-term data stores that are both usable and searchable. To be considered usable, archived data must be both accessible and understandable. NSSDC implements processes that seek to ensure both. To any researcher, unfamiliar archived data are only as good as their documentation. Enforcing policies that require thorough documentation enables NSSDC to more adequately preserve searchable and usable data. And the practice of moving more archived data to near-line and online stores allows NSSDC to make more of its archives available and increasingly accessible to a growing community of present and future researchers.

Acknowledgment

The author wishes to thank Edwin Grayzeck, James Thieman, Ed Bell, Patrick McCaslin, Jane Russell, and Elizabeth Zubritsky for their assistance and contributions to this document.

References

[1] E. J. Grayzeck, "NSSDC Annual Report for 2009", March, 2010. Available on the WWW, http://nssdc.gsfc.nasa.gov/nssdc/annual/2009/2009AnnRep.pdf

[2] E. J. Grayzeck, J. R. Thieman, B. E. Jacobs, and Nathan L. James, "NSSDC Proposal to the Heliophysics Senior Review," July 2009, pp. 16-20.

[3] "NASA Heliophysics Data Management Policy," Appendix C April 2009".