Delivery of Large Data Sets to the NSSDC
White Paper

D. Crichton, E. Grayzeck, S. Hughes, C. Isbell, E. Law, P. McCaslin, T. Stein

DRAFT
March 14, 2008

I. Introduction

This white paper captures the problem definition and proposed process for the Planetary Data System (PDS) to deliver large data sets (>300 GB) to the National Space Science Data Center (NSSDC) for deep archive as discussed by the NSSDC Delivery Working Group (NDWG) [Crichton, Grayzeck, Hughes, Isbell, Law, McCaslin, Stein]. It also includes recommendations of the next steps to be taken.

II. Background / Problem Definition

PDS had established a Memo of Understanding with the NSSDC to safeguard a copy of PDS data at the NSSDC for long-term preservation. In the past, PDS Nodes would deliver archived data sets to the NSSDC via tapes, CDs, and/or DVDs. The hard media were mailed to the NSSDC where the data sets would be incorporated into the NSSDC holdings and corresponding NSSDC ID's would be assigned and given to the originating PDS Node. However, there was no formal process/procedures to guide the Nodes in the delivery of archived data sets to the NSSDC. In addition, the mechanism by which the PDS/EN tracked the inventory of what had been delivered to the NSSDC slowly degraded after the advent of periodic data set releases for large missions.

In February 2006, the PDS Management Council finalized a set of PDS requirements. Specifically, Level 3 requirement 4.1.5 was included to address the long-term preservation of the PDS data via NSSDC as follows:

> "PDS will meet U.S. federal regulations for the preservation and management of the data through its Memorandum of Understanding (MOU) with the National Space Science Data Center (NSSDC)."

Recently, the NSSDC has re-architected their archive system to support long-term preservation of data using a tape-based archive system. The concept is that electronic archive packages are created and transferred to the NSSDC in order to be archived and preserved using the new system. This differs from the past when optical media from PDS were maintained on the media in which they were sent. PDS is interested in an electronic delivery mechanism whereby PDS data holdings are not bound to and managed on the media that is used to transfer them. Recent studies conducted by PDS regarding DVD and CD media have identified reliability problems, which may be problematic as a storage mechanism for preserving data long term.

In order to create archive packages, the NSSDC has provided the Multi-file Package Generator and Analyzer (MPGA) software for the PDS Nodes to package PDS volumes into Archive Information Packages (AIPs) that can be delivered to the NSSDC electronically via FTP. These AIPs are restricted to 300 GBs in size. Both the packaging and transfer mechanism present a problem for data sets that are several terabytes in size. It was decided at the December 2007 meeting of the PDS Management Council that it is necessary to develop a new process for sending large data sets to the NSSDC. As a result, the NDWG was formed to address this need.

The working group has identified the following critical objectives that need to be addressed:

- Objective 1: Identify a structure and mechanism for sending large data sets to the NSSDC
- Objective 2: Identify a process for requesting data from the NSSDC
- Objective 3: Identify a process for replacing data at the NSSDC

III. Proposed Processes

Concept

Conventional storage systems invariably have size limitations that need to be considered when creating a PDS data repository. Since large PDS data sets can exceed the capacity of storage devices, it is sometimes necessary that data sets be broken into multiple integral units. Within the PDS, these integral units have been the archive volume as defined within the PDS Standards Reference. In the case of a data set that does not exceed the capacity of a repository storage device, PDS nodes curating the data set will still create a single archive volume, at least for archiving purposes, using the logical structure as defined in the PDS Standards Reference.

In developing the concept, the working group identified five key principles for the transfer of data between PDS and NSSDC:

i) PDS should deliver a set of data to the NSSDC using a standard "PDS Submission Package"
ii) NSSDC should provide a submission identifier for each package
iii) More than one package can be delivered on the media
iv) Requests for recovery of data will be made at the submission package level
v) Updates to the contents of a submission package will require full replacement

In order to satisfy these principles, the working group recommends that the curating PDS node construct a submission package by copying an archive volume onto a delivery media such as a data brick. Included would be a subscription package manifest for the archive volume that would provide the necessary metadata (including checksum) to archive the data at the NSSDC. Nodes wishing to deliver more data on the data brick would simply copy additional archive volumes and the associated subscription package manifests onto the data brick. The node would also create a

delivery manifest for the data brick, which annotates the submission packages on the brick. NSSDC would provide a submission identifier for each submission package that effectively equates a submission identifier to a PDS archive volume for cataloging purposes. NSSDC will take the delivery media, ingest the submission packages into their system, and then return the media back to the PDS node.

Figure 1. Illustrates the medium, the submission package, archive volume, and data set and their relationships using UML class diagrams. Proposed submission package manifest, and delivery manifest are also shown.

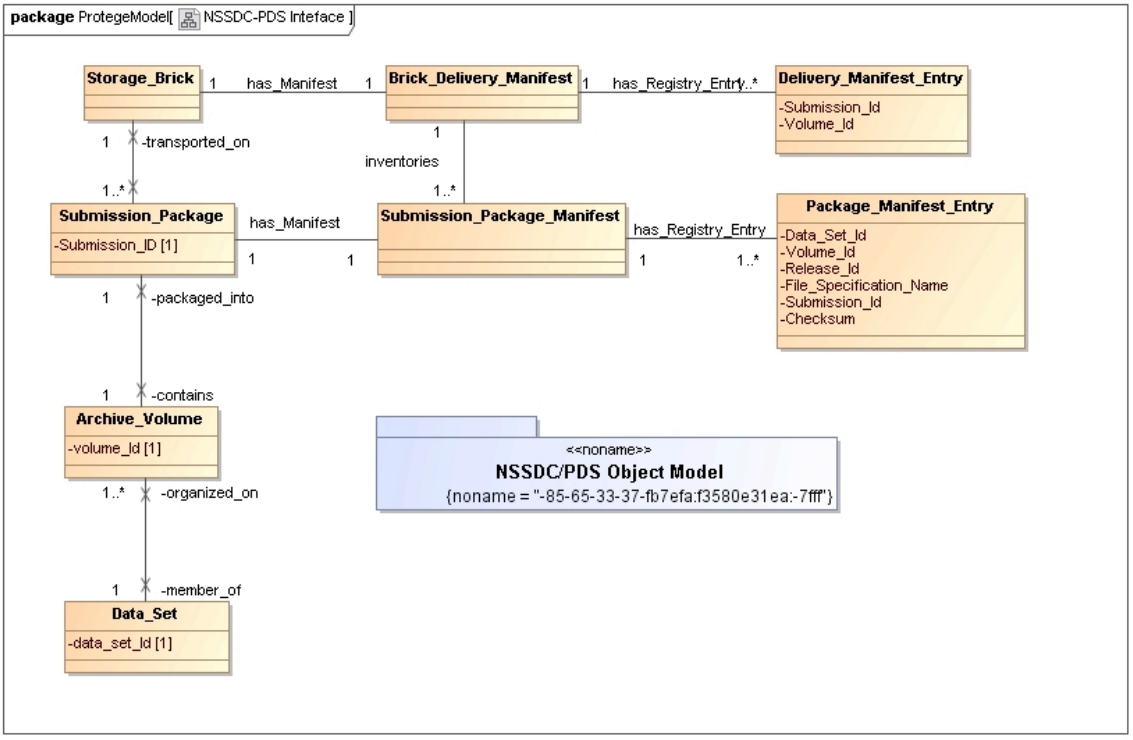

Figure 1 – Submission Package relationship UML Class Diagram

Definition

The following terms and key associations are used in the proposed process:

Archive volume – The integral unit described in the concept above adhering to the virtual directory structure defined in the VOLUME object included in the file "VOLDESC.CAT" per the PDS Standards Reference. The volume can be either an archive volume containing a complete data set or an archive volume containing a portion of a data set.

Submission package – The smallest unit for NSSDC ingestion. It is equivalent to an archive volume with supporting delivery information in an associated submission package manifest. Each submission package has a unique submission package ID assigned by the NSSDC.

Submission package manifest – A set of metadata describing the package that is delivered. The information includes the following Volume ID, DATA_SET_ID, the associated RELEASE_ID, and a list of files on the volume. For each file, it includes a file specification name and its checksum.

Delivery Medium – The smallest unit of storage device to deliver to NSSDC. It contains one or more submission packages along with a delivery manifest.

Delivery manifest – A set of metadata describing what's on a storage device delivered to NSSDC. The information includes the number of submission packages to be ingested into NSSDC and the associated submission package IDs.

Submission Process

This process is used to submit one or more submission packages to the NSSDC using one or more media. The step-by-step procedures and protocols to be followed when submitting one or more packages are as follows:

a. PDS Node staff requests submission package IDs from NSSDC via email.
b. NSSDC assigns submission package IDs and returns them to the requestor.
c. PDS Node staff creates submission package manifests.
d. PDS Node staff creates delivery manifests accordingly.
e. PDS Node staff copies the archive volumes, submission package manifests and delivery manifests onto the media.
f. PDS Node staff sends the media to NSSDC.
g. PDS Node staff sends an email to NSSDC to let them know that a delivery has been initiated.
h. NSSDC acknowledges receipt of the media.
i. NSSDC ingests submission packages into NSSDC holdings based on the delivery and submission package manifests.
j. After successful ingestion of the submission packages, NSSDC notifies originated PDS Node staff.
k. In addition, NSSDC notifies and provides to PDS EN Operations the submission package manifests.
l. NSSDC sends the media back to the Node.
m. PDS EN Operations incorporates the submission package manifests into the PDS Catalog.
n. PDS Node acknowledges receipt of returned media.

Retrieval Process

This process is used to retrieve one or more submission packages from NSSDC. The step-by-step procedures and protocols to be followed when retrieving one or more submission packages are as follows:

a. PDS Node staff sends an email request to NSSDC providing specific submission packages IDs.
b. NSSDC acknowledges the request with a reply of estimated time of preparation and returning the packages.
c. NSSDC retrieves requested submission packages along with their submission package manifests.
d. NSSDC creates delivery manifests based on the request.
e. NSSDC copies the submission packages, submission package IDs, and delivery manifests onto media.
f. NSSDC sends media to the requestor and informs the requestor via email a return has been initiated.
g. PDS Node staff notifies NSSDC upon reception of the media.
h. PDS Node staff retrieves the submission packages from the media and then sends the media back to NSSDC.
i. NSSDC acknowledges receipt of returned media.

Replacement Process

This process is used to replace one or more submission packages at the NSSDC. The step-by-step procedures and protocols to be followed when replacing one or more submission packages at the NSSDC are as follows:

a. PDS Node staff creates submission package manifests to be replaced.
b. PDS Node staff creates delivery manifests accordingly.
c. PDS Node staff copies the replacement volumes, submission package manifests and delivery manifests onto media.
d. PDS Node staff sends an email to NSSDC to let them know that a replacement has been initiated.
e. PDS Node staff sends the media to NSSDC.
f. NSSDC acknowledges receipt of the media.
g. NSSDC ingests submission packages into NSSDC holdings based on the delivery and submission package manifests.
h. After successful ingestion of the submission packages, NSSDC notifies originated PDS Node staff.
i. In addition, NSSDC notifies and provides to PDS EN Operations the replacement submission package manifests.
j. NSSDC sends the media back to the Node.
k. PDS EN Operations incorporates the replacement submission package manifests into the PDS Catalog.
l. PDS Node acknowledges receipt of returned media.

IV. Recommendation

The NDWG recommends the following next steps to be taken:

a. The working group performs a test including submission, retrieval and replacement processes using a data set/volume(s) from both the Geosciences and Imaging Nodes.
b. The working group presents the results of the test back to the Management Council, if successful, along with the proposed process.
c. The working group works with the Management Council to make the process operational.