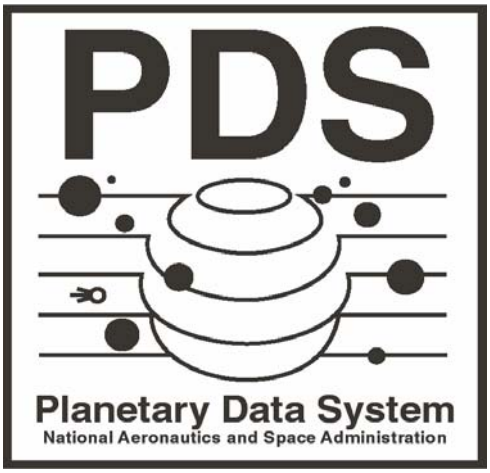
Planetary Data System

Archive Data Availability Use Cases

DRAFT

July 31, 2007

Version 0.10070731



JPL

Jet Propulsion Laboratory
Pasadena, California

JPL D-xxxxx

CHANGE LOG

Revision	Date	Description	Author
Start Draft	2007-04–01	First Draft	R. Joyner
0.10070504	2007-05-04	Edits	D. Crichton, S. Hughes
0.10070605	2007-06-05	Edits from first telecom	S. Hughes
0.10070623	2007-06-23	Edits from second telecom	S. Hughes
0.10070731	2007-07-31	Edits from first requirements telecom	S. Hughes

Table of Contents

1 Introduction

The purpose of this document is to capture use cases which will describe how the PDS addresses availability of its data holdings. This encompasses the need for backups in the case of "disaster recovery" (4.1.4) or simple data loss, the need for "continuous operations" (2.10.1), and overall system performance especially during peak load periods.

Note that PDS has developed a set of related use cases and requirements for *Data Integrity and Tracking*. While the Data Integrity and Tracking use cases / requirements focus *on delivery tracking, archive tracking, and file corruption,* the use cases specified herein focus on *continuity of operations to ensure on-going access to its data holdings.*

At the meeting on Nov/2006, the Management Council adopted a policy to ensure the integrity of the PDS archive as follows:

Each node is responsible for periodically verifying the integrity of its archival holdings based on a schedule approved by the Management Council. Verification includes confirming that all files are accounted for, are not corrupted, and can be accessed regardless of the medium on which they are stored. Each node will report on its verification to the PDS Program Manager, who will report the results to the Management Council.

From the perspective of providing both continuous operations (2.10.1) and disaster recovery (4.1.4) for the archive, the use cases described in this document are based on the following assumptions. These assumptions will ultimately need to be reviewed and approved by the PDS Management Council. Note that in any tradeoff between integrity of the archive and quality of service, the emphasis is on the integrity of the archive.

Disaster recovery ensures that PDS can recover data and services from an unforeseen event which might cause outages to services, facilities and hardware. For disaster recovery, the following assumptions are made:

1) There are three copies of the archived data. For this document these copies are called a) the primary repository, b) the secondary (aka backup, mirror) repository, and c) the tertiary (aka deep archive).

2) The primary repository is accessible online except in a few specific instances, such as infrequently used Radio Science data sets. The secondary repository can be off-line.

3) For disaster recovery such as a major earthquake in Southern California or St. Louis Missouri, at least two of the repositories must be at more than one geographically distinct location..

4) As per PDS policy, each PDS Node is to develop a disaster recovery plan to be submitted to and approved by the PDS management council. In this plan, the perceived risks and types of

disasters will be documented and solutions appropriate to the individual node, including the rationale for the choice of geographically distinct locations for each repository, will be provided

5) As per PDS requirements (4.1.5) the PDS places a copy of its data holdings into the NSSDC to meet U.S. Federal regulations for the preservation of data. It is assumed that NSSDC policies and procedures ensure the long-term preservation of data consistent with U.S. Federal regulations, allow for the recovery of data from its repositories, and are committed to supporting a recovery interface with the PDS. The NSSDC is to have a tertiary (deep archive) copy of the data. It is also assumed that the PDS Management Council will want the recovery interface tested.

Continuous operations ensures that PDS strive to achieve a minimum "quality of service" (QoS) rating for its users. For continuous operations, the following assumptions are made:

1) Optimal Operational Scenario - It is desirable that data and services of high interest to the PDS user community are available world-wide 24x7 and experience limited downtime.

2) Routine Operational Scenario

a) An on-line primary data repository at a node should never be unavailable for longer than 24 hours.

b) An off-line primary data repository at node should be able to make data available for distribution to a user within 72 hours.

c) Over weekends, holidays, or other situations where node staff are unavailable, additional delays in service may occur.

3) Loss-of-data Scenario – In case of a loss-of-data event at a node but where operational capability is not impaired, restoration of the data from a backup should occur within 1 week.

4) Catastrophic Scenario - In the case of a catastrophic event at a node where there is a loss-of-data and all operational capability, the primary data repository should be available within one month at either the original or an alternate node. The level of service provided will include at least the retrieval of data files using file identifiers over simple internet and file system protocols.

1.1 Controlling Documents

[1] Planetary Data System (PDS) Level 1, 2 and 3 Requirements, May 26, 2006.

1.2 Applicable Documents

[1] Planetary Data System Data Integrity and Tracking Use Cases Document, September 26, 2006, DRAFT.

[2] Planetary Data System (PDS) Data Integrity and Tracking Level 4 Requirements, November 13, 2006, DRAFT

[3] Planetary Data System Engineering Node Mirror documentation. http://pds-engineering.jpl.nasa.gov

1.3 Document Maintenance

This document and the use cases specified herein will be kept under configuration control by the PDS Engineering Node.

2 Actors

An actor is a user who is involved in any step of the life cycle of a PDS data product from data ingestion to data usage. The following actors are referenced or implied in the use cases specified herein.

- **PDS Node**
    1. Discipline Nodes
    2. Data Nodes
    3. Engineering Node

- **Primary Repository**
    1. Online
    2. Offline

- **Secondary Repository**
    1. Online
    2. Offline

- **PDS Operations**

- **Tertiary Repository**
    1. National Space Science Data Center (NSSDC)

- **Data Consumer**
    1. Planetary Scientist
    2. Mission Flight Project members
    3. Mission Operations
    4. Educator
    5. General Public

3 Definitions

The following definitions are used in the use case Sequences.

1. **Actors** - An actor is a person, organization, or external system that plays a role in one or more interactions with your system
2. **Data Consumer -** Entities that receive data from PDS.
3. **Data Product** – A data product label and one or more data objects.
4. **Data Product Label** – One or more data object descriptions.
5. **Data Set** – A data set is a set of data products collected for a specific purpose and includes not only the product label file, but all data files that comprise each data product.
6. **PDS Node** – Any PDS node including Discipline Nodes, Data Nodes, and the Engineering Node. The Discipline Nodes include both science and support nodes.
7. **PDS Resource –** A PDS Resource is any web resource, accessible via http protocol, that has been ingested into the PDS main catalog Resource tables. Information required for each resource includes an identifier, name, type, description, URL, and associated data sets.
8. **Primary Repository -** A primary repository is the principal location for a Node's data holdings. Primary repositories are managed by the PDS Discipline Nodes and maybe online or offline.
9. **Repository Inventory –** The repository inventory is a complete list of the primary and secondary repositories for each PDS data set. Each repository is considered a PDS Resource.
10. **Secondary Repository -** A secondary repository contains a copy of a Node's primary repository. This may be online or offline.
11. **Tertiary Repository –** A deep archive provides long term preservation of a data holding
12. **Use cases**. A use case describes a sequence of actions that provide something of measurable value to an actor.

4 Requirements

The following are driving requirements for data accessibility addressed in the following Level 3 requirements:

> *2.8.1 PDS will maintain a distributed archive where holdings are maintained by Discipline Nodes, specializing in subsets of planetary science*
>
> *2.10.1 PDS will monitor the system and ensure continuous operation*
>
> *4.1.2 PDS will develop and implement procedures for periodically ensuring the integrity of the data*
>
> *4.1.4 PDS will develop and implement a disaster recovery plan for the archive*
>
> *4.1.5 PDS will meet U.S. federal regulations for preservation and management of the data through its Memorandum of Understanding (MOU) with the National Space Science Data Center (NSSDC)*

The following policy, adopted by the Management Council (MC), addresses how the PDS will protect the integrity of its holdings:

Each node is responsible for periodically verifying the integrity of its archival holdings based on a schedule approved by the Management Council. Verification includes confirming that all files are accounted for, are not corrupted, and can be accessed regardless of the medium on which they are stored. Each node will report on its verification to the PDS Program Manager, who will report the results to the Management Council.

November 2006

5 Data Availability Use Case Diagram

Figure 1 illustrates the data availability use case scenarios associated with the primary, secondary, and tertiary repositories. If the primary repository is online, then the data is transferred over the internet via an electronic download. If the primary repository is offline then the data may be brought on-line and transferred or distributed using some type of physical media.

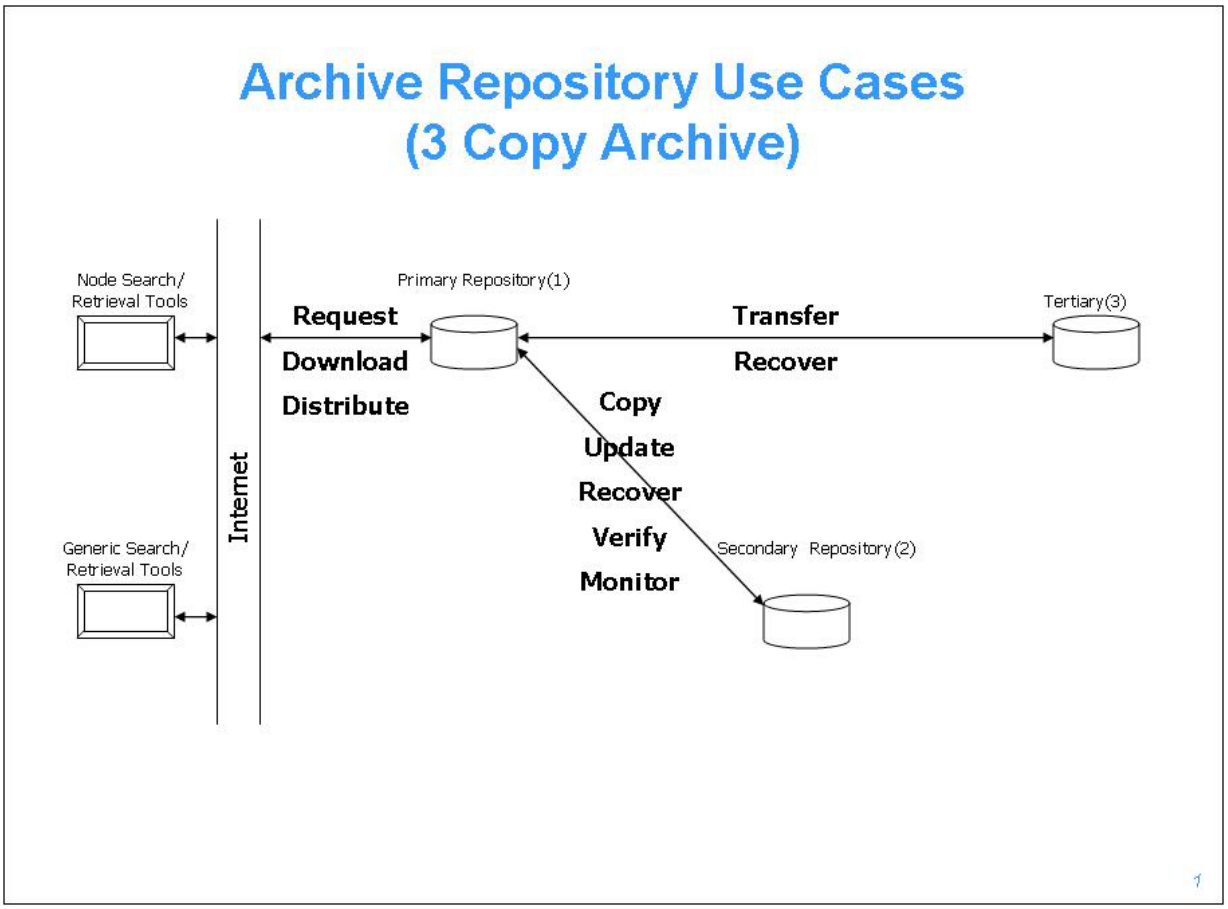


Figure 1 – Archive Repositories

6 Data Availability Use Cases

Each use case describes one or more high level scenarios where one actor or a group of actors has a need to access various PDS holdings. These use cases are suggested by several level one, two, and three PDS requirements. These use cases will subsequently aid in the formulation of level 4 and implementation level requirements. It is important to note that there are many ways to implement these use cases. The cases presented are intended to be general cases.

6.1 UC-1: A Node copies all node holdings from its Primary Repository to a Secondary Repository

Description: A copy of the primary repository is placed at the secondary repository

Actors: PDS Node, Primary Repository, Secondary Repository

Sequences:
1. A PDS Node develops an agreement with the secondary repository manager to support a secondary repository.
2. The PDS Node transfers a copy of all data holdings from its primary repository to the secondary repository using PDS recommended standards for data transfer and integrity checking.
3. The PDS node adds references to the secondary repository to the repository inventory. (A.T. UC-1)

6.2 UC-2: A Node makes incremental updates to a Secondary Repository

Description: Incremental changes to the primary repository are transferred to a secondary repository

Actors: PDS Node, Primary Repository, Secondary Repository

Sequences:
1. The PDS Node transfers new or changed data files from its primary repository to the secondary repository using PDS recommended standards for data transfer and integrity checking.

2. The PDS node updates references to the secondary repository in the repository inventory. (A.T. UC-1)

6.3 UC-3: A Node discovers a corrupted file in the Primary Repository

Description: A PDS Node discovers that a file in the primary repository needs to be restored from the secondary repository because it is corrupted or lost.

Actors: PDS Node, Primary Repository, Secondary Repository, Deep Archive

Sequences:
1. The PDS Node discovers or is notified of a problem with files in the primary repository.
2. If the file has not been transferred to a secondary repository and a local backup exists, then the file is recovered from the local backup.
3. Otherwise, the PDS Node recovers the file from the secondary repository using PDS recommended standards for data transfer and integrity checking.
4. Should the PDS Node discover that both the primary and secondary copies are corrupted or lost, the PDS Node recovers the file from the Deep Archive. (D.I. UC-7)

6.4 UC-4: A Primary Repository is unavailable as a result of a catastrophic event

Description: A PDS Node suffers a catastrophic event and the Node holdings are not available for access and download

Actors: PDS Node, Primary Repository, Secondary Repository, Deep Archive

Sequences:
1. A catastrophic event (hardware, facilities, natural disaster, etc) occurs and the PDS Node holdings are no longer available for access or download at the primary repository.
2. The PDS Node rebuilds the primary repository from a local backup or the secondary repository using the PDS standards for data transfer and integrity checking.
3. Should the PDS Node discover that neither a local backup nor secondary repository are available, the PDS Node recovers the file from the Deep Archive. (D.I. UC-7)

6.5 UC-5: A Node verifies the integrity of the its primary repository

Description: The PDS Node periodically verifies the integrity of the primary repository

Actors: PDS Node, Primary Repository, Secondary Repository, Deep Archive

Sequences:
1. A PDS Node periodically performs data integrity checks on the data holdings in the primary repositories. (D.I. UC-5)
2. If the PDS Node discovers an integrity check problem, the data is recovered from the secondary repository. (D.I. UC-7)
3. If data is not available from the secondary the data is recovered from the deep archive at NSSDC. (D.I. UC-7)
4. The results of integrity verification are presented to the PDS Management Council.

6.6 UC-6: A Node verifies the integrity of the its secondary repository

Description: The PDS Node periodically verifies the integrity of the secondary repository

Actors: PDS Node, Primary Repository, Secondary Repository, Deep Archive

Sequences:
1. A PDS Node periodically performs data integrity checks on the data holdings in the secondary repositories. (D.I. UC-5)
2. If the PDS Node discovers an integrity check problem, the data is transferred from the primary repository. (D.I. UC-7)
3. The results of integrity verification are presented to the PDS Management Council.

6.7 UC-7: A Node verifies the integrity of the its deep archive (tertiary) repository

Description: The PDS Node periodically verifies the integrity of the deep archive (tertiary) repository.

Actors: PDS Node, Primary Repository, Secondary Repository, Deep Archive

Sequences:
1. A PDS Node periodically performs data integrity checks on the data holdings in the deep archive (tertiary) repositories. (D.I. UC-5)
2. If the PDS Node discovers an integrity check problem, the data is transferred from the primary repository. (D.I. UC-7)
3. The results of integrity verification are presented to the PDS Management Council.

6.8 UC-8: A Node verifies the accessibility of its secondary repository

Description: The PDS Node periodically verifies the accessibility of its secondary repository

Actors: PDS Node, Secondary Repository
Sequences:
1. A PDS Node periodically checks the accessibility to its secondary repository
2. If the PDS Node discovers that the secondary repository is unavailable, then it works with the facility manager to restore it to an operational state.

6.9 UC-9: A Node verifies the accessibility of its deep archive (tertiary) repository

Description: The PDS Node periodically verifies the accessibility of the tertiary repository

Actors: PDS Node, Tertiary Repository

Sequences:
1. A PDS Node periodically checks the accessibility to its deep archive (tertiary) repository.
2. If the PDS Node discovers that the deep archive is unavailable, then it notifies the NSSDC and ensures that accessibility is restored.