Planetary Data System

Archive Integrity Requirements: Archive Data Integrity, Tracking, Availability, and Preservation Planning

DRAFT

June 1, 2008

Version 0.10080601

Jet Propulsion Laboratory
Pasadena, California

JPL D-xxxxx

CHANGE LOG

Revision	Date	Description	Author
Start Draft	2007-08-29	First Draft of Combined Documents	S. Hughes
0.10070829	2007-09-04	EN Review of document	D. Crichton, R. Joyner, S. Hughes
0.10070920	2007-09-20	WG Review of document	M. Gordon, E. Guninness, S. Hughes
0.10080505	2008-04-20	Incorporated Node comments into document	R.Joyner, S.Hughes, D.Crichton

Table of Contents

1 Introduction

At its meeting in November 2006, the PDS Management Council adopted the following policy to ensure the integrity of PDS archives:

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.

This document establishes the requirements, based on this policy, for the integrity, tracking, and availability of PDS data holdings and essential functions. There are four areas of concern:

1) File corruption during preservation and during transfer of data to, from, and across PDS.

2) End-to-end tracking of data from delivery by data providers through the PDS to the deep archive, NSSDC. Delivery Tracking is following data from the data provider to the PDS. Archive Tracking is maintaining an inventory of all files and ensuring that they are accounted for and accessible to the community.

3) Backup of both the data holdings and essential functions of the system —"disaster recovery" and "continuous operations," respectively.

4) Auditing the data holdings resident at the deep archive.

Sections 6 through 9 address requirements for each of the above four areas. They are based on policies, the PDS Level 3 Requirements, and use cases that were individually developed for each of the above areas.

1.1 Overview

The following diagram illustrates how the four processes detailed in this document, Data Integrity, Delivery and Archive Tracking, Data Availability, and Preservation Planning, play a role in the flow of data into and out of the PDS.

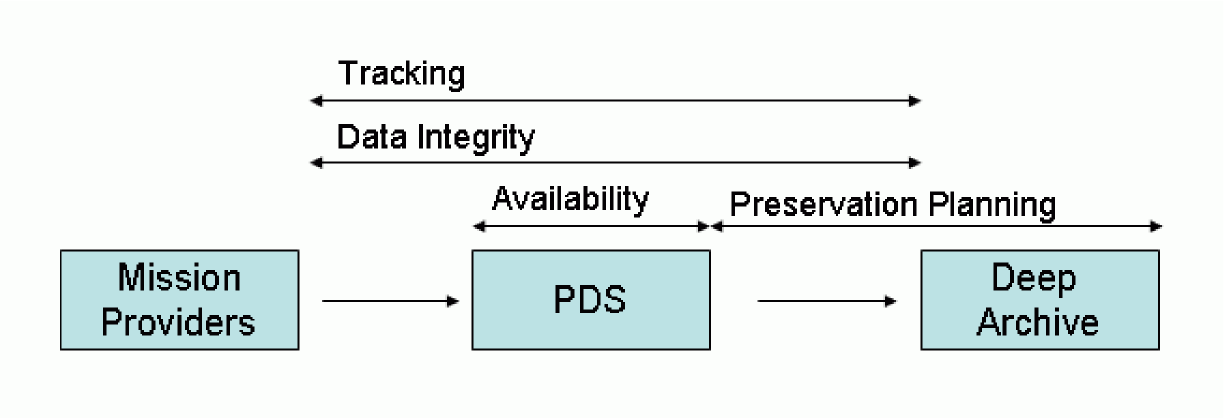


Figure 1-1 – Flow Diagram

Figure 1-1, Flow Diagram, depicts a "throughput" of files into the PDS and out of the PDS. Data Providers supply "data", in the form of files, to PDS. PDS supplies "data", in the form of files, to the Deep Archive. Files are "tracked" from the point when PDS receives the "data" from the Data Providers to the point when PDS supplies the "data" to the Deep Archive. File "integrity" is ensured from the point when PDS receives the "data" from the Data Providers to the point when PDS supplies the "data" to the Deep Archive. File "availability" is ensured throughout its life in the PDS. "Preservation Planning" ensures the file transfer interface between PDS and the Deep Archive is preserved.

1.2 Controlling Documents

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

1.3 Applicable Documents

[2] Planetary Data System Data Integrity Use Cases, DRAFT, October 11, 2006, Version 0.10060926.

[3] Planetary Data System Archive and Delivery Tracking Use Cases, DRAFT, January 29, 2007, Version 0.10070129.

[4] Planetary Data System Archive Data Availability Use Cases, DRAFT, July 31, 2007, Version 0.10070731.

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

1.4 Relevant “Documents” in Preparation

The following is a list of “documents” that may be written which are envisioned to be relevant to the requirements presented herein:

- Deep Archive Interface: Electronic Transfer of Data to the Deep Archive, NSSDC
- Backup Interface: Transfer of Data to the Secondary Archive, SDSC
- Operation of Data Nodes within the PDS

1.5 Document Maintenance

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

2 Definitions

The following definitions are used in the requirements.

1. **Acceptance Status –** A status indicating whether a data delivery has been accepted.
2. **Actors -** An actor is a person, organization, or external system that plays a role in one or more interactions with your system
3. **Archive Manifest –** An archive manifest is a list of files contained in a repository. The information for each file should include at least filename, file checksum, and the checksum type. The list of files should include at least those that comprise data sets that have a status indicating the data set has passed peer review or is being saved long-term.
4. **Accessible** - easily reached or located: easy to enter or reach physically.
5. **Authorized Actor –** An actor who has been approved to be informed of actions by the system. For tracking system notification, authorized actors are a subset of tracking system actors that have been specifically authorized to be informed of specific actions by the system.
6. **Availability** - The property of information being accessible and usable upon demand by an Authorized Actor or process.
7. **Corruption –** errors that occur during transmission or retrieval, introducing unintended changes to the original data.
8. **Data Consumer -** Entities that receive data from PDS.
9. **Data Product –** A data product label and one or more data objects.
10. **Data Product Label –** One or more data object descriptions.
11. **Data Set –** A data set is a set of data products collected for a specific purpose and all related ancillary data and documentation.
12. **Deep Archive Manifest –** The Deep Archive Manifest, a list of files contained in the archive repository of each Node, is used to identify the files transferred to the Deep Archive, NSSDC. The information contained in the Deep Archive Manifest is a subset of the information in the Archive Manifest. The information for each file should include at least filename, file checksum, and the checksum type. The list of files should include at least those that comprise archival quality data sets to be transferred to the deep archive.
13. **Delivery Status –** A status indicating whether a data delivery has been successfully transferred.
14. **File –** A collection of bytes stored as a single unit within a file system.
15. **Manifest** - A manifest is a list of files.. For the purpose of tracking, the manifest is a list of files and will include at a minimum the filenames, checksum, and checksum type. (See Archive Manifest and Delivery Manifest).
16. **PDS Delivery Manifest –** The delivery manifest is a list of files in a delivery. The information for each file should include at least the file name, the file checksum, and the checksum type.

17. **Product Collection –** A product collection is a set of data products collected for a specific purpose.
18. **PDS Node –** Any PDS node including Discipline Nodes, Data Nodes, and the Engineering Node. The Discipline Nodes include both science and support nodes.
19. **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.
20. **Physical Media –** Any computer system device used for short or long term storage of data including but not limited to optical media, tape, and magnetic disk.
21. **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.
22. **Repository Inventory –** The repository inventory is a list of primary and secondary repositories that contain PDS data sets.
23. **Secondary Repository -** A secondary repository contains a copy of a Node's primary repository. Secondary Repositories may be online or offline.
24. **Tertiary Repository –** A tertiary repository is a deep archive which provides long term preservation of PDS data holdings.
25. **Use cases**. A use case describes a sequence of actions that provide something of measurable value to an actor.
26. **Volume –** Any organized collection of files that reside on physical media for the purpose of near term storage, online access, data submission, electronic distribution, or long-term archive. Note that this definition includes the PDS archive volume.

3 Applicable PDS Policies

1. Archive Integrity Policy (November 2006)

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.

2. Data Delivery Policy (October 2005)

The following policy, adopted by the Management Council (MC), addresses the data deliveries to PDS and their management:

Data producers shall deliver one copy of each archival volume to the appropriate Discipline Node using means/media that are mutually acceptable to the two parties. The Discipline Node shall declare the volume delivery complete when the contents have been validated against PDS Standards and the transfer has been certified error free.

The receiving Discipline Node is then responsible for ensuring that three copies of the volume are preserved within PDS. Several options for "local back-up" are allowed including use of RAID or other fault tolerant storage, a copy on separate backup media at the Discipline Node, or a separate copy elsewhere within PDS. The third copy is delivered to the deep archive at NSSDC by means/media that are mutually acceptable to the two parties.

3. Backup and Recovery Policy (November 2006)

Each node is responsible for defining and implementing a disaster recovery plan which covers loss of data and/or system functionality within guidelines provided by the Management Council. The plan shall be delivered to and approved by the PDS Program Manager.

4 PDS Level 3 Requirements

4.1 Driving Requirements

The following Level 3 requirements were used to derive the level 4 requirements in this document.

1.3.4 PDS will coordinate with the data providers to establish schedules for delivery of archival products to the PDS

Rationale: PDS will plan and track the delivery of data files from data providers to the PDS.

1.3.5 PDS will coordinate with data providers to establish schedules for public release of archival products

Rationale: PDS will plan and track the delivery of data files from data providers to the PDS that will include a schedule for release of the data to the public.

2.2.2 PDS will track the status of data deliveries from data providers through the PDS to the deep archive

Rationale: PDS will plan and track the delivery of data files from data providers, at PDS nodes, and to the NSSDC/deep archive.

2.5.1 PDS will develop and publish procedures for accepting archival data

Rationale: PDS needs to ensure that it defines a set of procedures for accepting archived data from missions to ensure that all files have been delivered, error-free.

2.5.2 PDS will implement procedures for accepting archival data

Rationale: PDS needs to ensure that it follows a common set of procedures when accepting archived data from missions to ensure that all files have been delivered, error-free and that all information is tracked**.**

2.5.3 PDS will inform a data provider why a rejected archival product does not meet archiving standards

Rationale: PDS needs to ensure that deliveries from data providers satisfy PDS criteria to protect against data integrity problems including either missing or corrupted files. If the criterion is not met, then the delivery needs to be rejected.

2.6.3 PDS will integrate the catalog with the system for tracking data throughout the PDS

Rationale: Part of tracking is cataloging the information of the artifacts being tracked. From an architecture perspective, it is important to ensure the catalog system can support tracking of the files that PDS manages.

2.7.1 PDS will develop and publish procedures for storing archival data

Rationale: A critical aspect of data integrity is ensuring that files are stored error-free. This requirement ensures that there are common procedures for storing files that requires PDS to have a mechanism for detecting corruption.

2.7.2 PDS will maintain appropriate storage for the PDS archive

Rationale: This requirement ensures that there will be a mechanism for detecting file corruption on the media in which it is stored and an infrastructure to ensure that files are available from multiple locations.

2.7.4 PDS will maintain appropriate storage for non-archived data managed by the PDS

Rationale: This requirement ensures that there will be a mechanism for detecting file corruption on the media in which the data is stored and an infrastructure to ensure that files are available from multiple locations.

2.8.1 PDS will maintain a distributed archive where holding are maintained by Discipline Nodes, specializing in subsets of planetary science

Rationale: This requirement ensures that files transferred between PDS nodes without data corruption and ensures that users can continue to get to PDS data in the event of a system failure or catastrophe.

2.8.2 PDS will maintain a distributed catalog system which describes the holdings of the archive

Rationale: This requirement ensures that data is transferred between PDS nodes without data corruption.

2.10.1 PDS will monitor the system and ensure continuous operation

Rationale: It is critical that PDS ensure that users can continue to get to PDS data in the event of a system failure or catastrophe.

3.2.1 PDS will develop and maintain online mechanisms allowing users to download portions of the archive

Rationale: It is critical that PDS ensures that data is transferred between PDS nodes without data corruption and that users can continue to get to PDS data online in the event of a system failure or catastrophe.

3.2.2 PDS will develop and maintain a mechanism for offline delivery of portions of the archive to users

Rationale: It is critical that PDS ensures that data is transferred to the user community without data corruption and that users can get copies of PDS data offline despite system failures or major catastrophes.

3.2.3 PDS will provide mechanisms to ensure that data have been transferred intact

Rationale: This requirement is critical to ensuring that files are not corrupted during data transfer.

3.3.6 PDS will develop and maintain a mechanism for notifying subscribed users when a data set is released or updated.

Rationale: Notification needs to be integrated with the tracking system to ensure proper operations.

4.1.1 PDS will define and maintain a set of quality, quantity, and continuity (QQC) requirements for ensuring long term preservation of the archive.

Rationale: This requirement is critical to identifying criteria for regularly auditing files maintained in the deep archive in order to verify the integrity of the PDS data holdings at NSSDC

4.1.2 PDS will develop and implement procedures for periodically ensuring the integrity of the data.

Rationale: This requirement is critical to ensuring that data maintained in the deep archive is regularly audited so as to verify the integrity of the PDS data holdings at NSSDC.

4.1.3 PDS will develop and implement procedures for periodically refreshing the data by updating the underlying storage technology.

Rationale: The integrity of files transferred from one media to another needs to be checked in order to verify that no corruption has been introduced at the bit level.

4.1.4 PDS will develop and implement a disaster recovery plan for the archive.

Rationale: PDS needs to ensure that three copies of data are maintained in order to ensure data can still be distributed in the event of a system failure or catastrophe.

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).

Rationale: This is not driving additional level 4 requirements for data integrity since the requirement is testable at this level.

4.2 Related Requirements

The following Level 3 requirements indirectly relate to the scope of this document but do not have any derived Level 4 requirements.

2.1.2 PDS will identify and maintain a list of proposed planetary science data sets to be added to the archive

Rationale: This requirement is related to tracking, however, it is proposed products and therefore there are no files to track. This function is currently relegated as a management function for PDS.

2.1.3 PDS will work with relevant NASA program officials to ensure that products resulting from data analysis programs are submitted to the archive

Rationale: This requirement is related to tracking, however, it is proposed products and therefore there are no files to track. This function is currently relegated as a management function for PDS.

2.6.1 PDS will develop and publish procedures for cataloging archival data

Rationale: Cataloging archival data is a generalized function and there are no driving requirements. The driving requirement for protecting the integrity of PDS is to ensure that files are tracked through the PDS Catalog System (2.6.3).

2.6.2 PDS will design and implement a catalog system for managing information about the holdings of the PDS

Rationale: The catalog system provides a general function for tracking information. The requirement to integrate the tracking system and catalog is captured through requirement 2.6.3.

5 Notation

The numbering of the requirements in this document will be formatted as **Ln.AA.BB.N**, where:

- **Ln** indicates the requirement level number.
- **AA** is an acronym representing the requirement category:
  - **DI –** Data Integrity
  - **DT –** Delivery Tracking
  - **AT –** Archive Tracking
  - **AV –** Data Availability
- **BB** is a two letter acronym for the requirement subcategory. (Optional)
- **N** is a unique number for the requirement.

Following the text of a requirement may be a reference to the requirement from which it was derived. The reference will be in parenthesis.

A paragraph following a requirement, which is indented and has a reduced font size, represents a comment that provides additional insight for the requirement. This comment should not be considered part of the requirement for development or testing purposes.

6 Data Integrity Level 4 Requirements

This section details the level 4 requirements for Data Integrity. These requirements are derived directly from the level 3 requirements [1] and the associated Data Integrity Use Cases [2].

The scope of these requirements is to address file corruption during the preservation and transfer of data to, from, and across PDS. It is also the intent that these data integrity requirements be media independent.

Within the context of this document, data integrity has the following broad meaning:

1. The condition in which data are identically maintained during any operation, such as transfer, storage, and retrieval.
2. The preservation of data for their intended use.
3. The a priori expectation of data quality.

The following diagram illustrates how Data Integrity plays a role in the flow of data into, across, and out of the PDS.

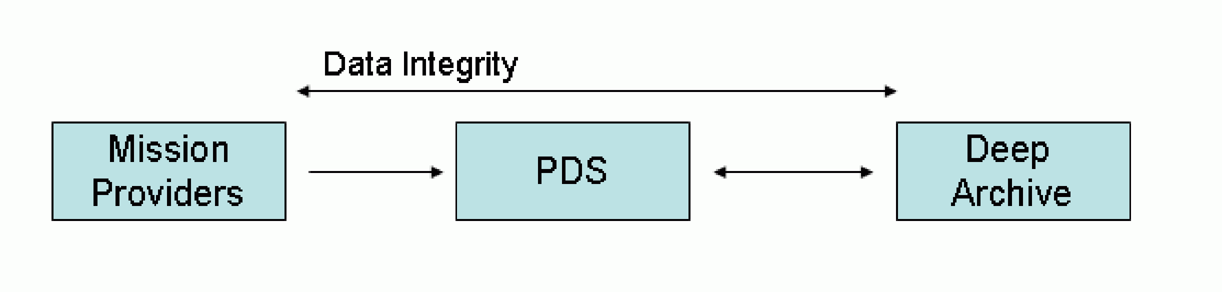


Figure 6-2 – Data Integrity Flow Diagram

Figure 6-1, Data Integrity Flow Diagram, depicts a "throughput" of files into the PDS and out of the PDS. Data Providers supply "data", in the form of files, to PDS. PDS supplies "data", in the form of files, to the Deep Archive. The Deep Archive supplies "data", in the form of files, to the PDS. File integrity is preserved from the point when PDS receives the "data" from the Data Providers to the point when PDS supplies the "data" to the Deep Archive. Concomitantly, file integrity is preserved when the Deep Archive transfers "data" back to the PDS.

L4.DI.1- PDS shall only accept, store, and will only deliver files that are not corrupted. (2.7.2, 2.7.4, UC-1)

> **Errors that occur during transmission or retrieval, introducing unintended changes to the original data, exemplify types of file corruption that would preclude PDS from accepting, storing, or delivering the corrupted files. File integrity is ensured during transfer of data to, from, and across PDS throught the interactions between PDS Discipline Nodes, PDS Data Nodes, the deep archive, and the PDS user community.**
>
> **Examples include:**
> 1. **PDS will ensure files transferred to the PDS are not corrupted.**
> 2. **PDS will ensure files transferred within the PDS are not corrupted.**
> 3. **PDS will ensure files transferred from the PDS are not corrupted.**
> 4. **PDS will ensure files retrieved from the deep archive are not corrupted.**

L4.DI.2 - PDS shall provide procedures for verifying that a data submission has not been corrupted. (2.5.1, 2.5.2, 2.7.1, UC-1, UC-4)

> **The procedures will cover the transmission and retrieval of data to, from, and across the PDS. The procedures will address the interactions between PDS Discipline Nodes, PDS Data Nodes, the deep archive, and the PDS user community to transfer files to, from, and across PDS.**
>
> **Examples include:**
> 1. **How the PDS ensures files transferred to the PDS are not corrupted.**
> 2. **How the PDS ensures files transferred within the PDS are not corrupted.**
> 3. **How the PDS ensures files transferred from the PDS are not corrupted.**
> 4. **How the PDS ensures files retrieved from the deep archive are not corrupted.**

L4.DI.3 - PDS shall require a data provider, *PDS Data Node*, or *PDS Discipline Node* to resubmit files that have been corrupted during transfer (2.5.2, UC-1, UC-4, UC-8)

L4.DI.4 - PDS shall periodically verify that data holdings have not been corrupted based on a schedule determined by the PDS Management Council (2.8.2, 4.1.2, UC-5)

L4.DI.5 - PDS shall verify that data holdings are not corrupted prior to migrating to another medium (4.1.3, UC-6)

L4.DI.6 - PDS shall ensure that data migrated from one medium to another have been successfully migrated and not corrupted during the transfer (4.1.3, UC-6)

L4.DI.7 - PDS shall request the deep archive (NSSDC) verify that the archival data holdings submitted to the NSSDC have been received intact (3.2.3, UC-3)

L4.DI.8 - PDS shall request the deep archive (NSSDC) verify that the archival data holdings recovered from the NSSDC are not corrupted (3.2.3, UC-7)

L4.DI.9 - PDS shall ensure that the archival data holdings received from the deep archive (NSSDC) are not corrupted (3.2.3, UC-7)

7 Tracking Level 4 Requirements

Data tracking focuses on tracking the status of data from its delivery by the data providers through the PDS to the deep archive, NSSDC. Tracking is differentiated into delivery tracking, primarily tracking data from the data provider to the PDS and archive tracking, the tracking of data within the PDS archive.

The following requirements are for data tracking within the PDS. They are derived directly from the PDS Level 3 requirements [1] and the Archive and Delivery Tracking Use Cases [3].

The following diagram illustrates how Delivery and Archive Tracking play a role in the flow of data into, across, and out of the PDS.

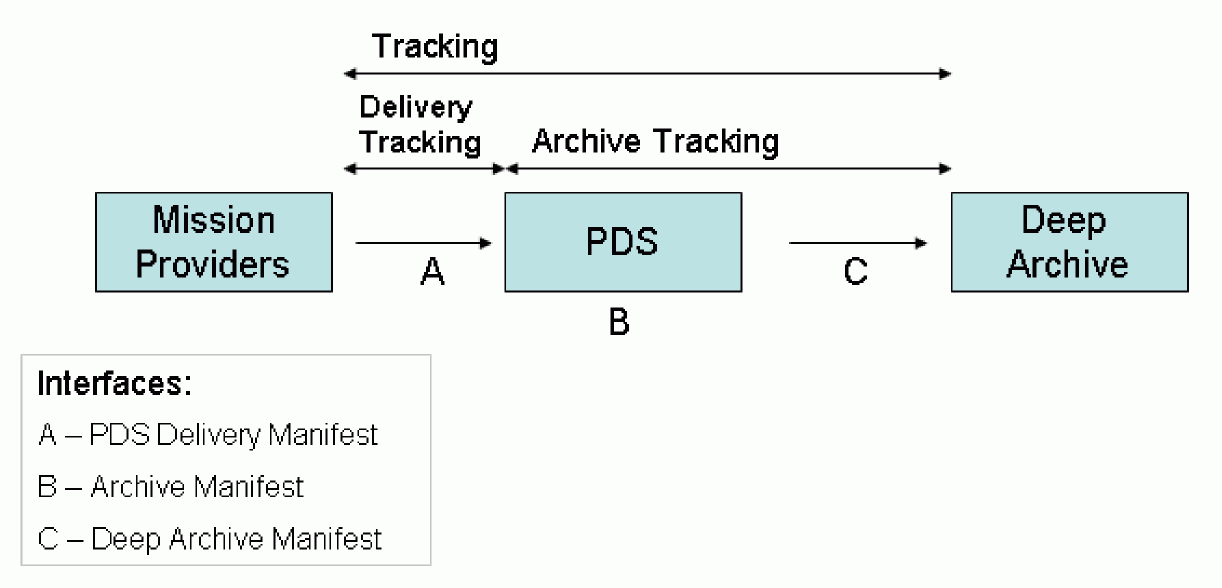


Figure 7-3 – Tracking Flow Diagram

Figure 7-1, Tracking Flow Diagram, depicts a “throughput” of files into the PDS and out of the PDS. Data Providers supply “data”, in the form of files, to PDS. PDS supplies “data”, in the form of files, to the Deep Archive. Files are “tracked” from the point when PDS receives the “data” from the Data Providers to the point when PDS supplies the “data” to the Deep Archive. Files are “delivery tracked”, using a Delivery Manifest, from the point when a Data Provider attempts to transfer “data” to the PDS up to the point where PDS accepts the “data”. Once PDS has accepted the “data”, the files are “archive tracked” by the PDS using an Archive Manifest”. Files are “archive tracked” from the point when PDS has accepted the “data” from the Data Provider to the point when PDS supplies the “data” to the Deep Archive. When transferring “data” to the Deep Archive, PDS will track the transfer of files using a Deep Archive Manifest.

7.1 Delivery Tracking Requirements

Delivery tracking encompasses the process of tracking files associated with data deliveries from any data provider through to the acceptance of the data by a PDS Node. These requirements focus on tracking data delivery events and their statuses. An important feature suggested by these requirements allows for data deliveries to be successfully transferred to but not necessarily accepted by a PDS node until reviewed. This “escrow” feature ensures that only data deliveries that meet certain criteria are actually accepted. These requirements are derived from the PDS Level 3 Requirements [1], PDS Tracking Use Cases [3], and PDS Policies.

7.1.1 General (GR)

L4.DT.GR.1 – PDS shall track the contents of a data delivery identifying each file associated with a delivery (2.2.2, UC-1, UC-1.2)

L4.DT.GR.2 – PDS shall allow verification that the contents of a data delivery package matches its delivery manifest by checking that every file listed in the manifest is in the package and vice versa. (2.2.2)

L4.DT.GR.3 – PDS shall notify actors about events including scheduled events associated with deliveries and any change in delivery or acceptance statuses. (3.3.6, UC-6)

> **The PDS Notification/Subscription system could be extended to fulfill this requirement. The notification of an actor will be triggered by an event. Various requirements in the following sections will signal events. The delivery plan will provide a list of actors to be notified for specific events. The subset of “authorized” actors will also be provided in the delivery plan.**

L4.DT.GR.4 – PDS shall allow authorized actors to set the tracking system statuses. (UC-1.2, UC-1.3)

L4.DT.GR.5 – PDS shall generate reports detailing the statuses of past and projected future data delivery events. (2.2.2, UC-4, UC-5)

7.1.2 Data Provider Delivers Data (DP)

L4.DT.DP.1 – PDS shall track data delivery events and their statuses based on a negotiated delivery plan between a data provider and a PDS Node. (2.2.2, 2.8.2, UC-1.3, UC-6)

Possible data delivery events include Delivery to PDS, and Acceptance by PDS. Each delivery event can have one or more event statuses which are captured by respectively, Delivery_Status and Acceptance_Status. On integration with the PDS Notification system, these delivery events will be added to the already existing Release and Archive events.

L4.DT.DP.2 – PDS shall maintain the delivery status of a data delivery from a Data Provider to a PDS Node. (1.3.4, 1.3.5, 2.2.2, UC-1.2)

L4.DT.DP.3 – PDS shall notify authorized actors of a change in the delivery status. (2.2.2, 2.5.2, 2.5.3, UC-6)

L4.DT.DP.4 – PDS shall allow a PDS Node to accept, reject, withdraw, or mark as incomplete a data delivery depending on whether the delivery is acceptable according to negotiated delivery parameters. (2.2.2, 2.5.2, 2.5.3, UC-1.3, UC-2)

If the PDS Node determines that the delivery is incomplete (i..e., not all of the files detailed in the delivery plan were present in the delivery), the PDS Node can indicate a status of “incomplete” until such time that the “missing” files are delivered to the Node. If the Node determines that the delivery is not “compliant”, the Node can indicate a status of “rejected” until such time that the delivery is brought into compliance or waivers are negotiated. The assumption is that the data provider will re-submit a withdrawn data delivery at a later time.

7.1.3 Data Node Termination (NT)

L4.DT. NT.1 – PDS shall support the dissolution of a Data Node and the transfer of the Data Note’s holdings to a permanent PDS Node. (2.2.2, UC-3)

7.2 Archive Tracking Requirements

The following requirements are for archive tracking. In this document, "archive tracking" encompasses the tracking of data holdings in the PDS archive and are levied on the PDS to track data files, data products, data volumes, and data sets from the acceptance of the data at a PDS Node through to the submission of the data to the NSSDC. These requirements are on the PDS to track the files in the archive as a whole, spanning all nodes. An important feature suggested by these requirements allows for data deliveries to be successfully transferred to but not necessarily accepted by the NSSDC until reviewed. This “escrow” feature is implemented implicitly by differentiating between the transfer to and acceptance of a package by the NSSDC. In other words, a delivery can be successfully transferred but not necessarily accepted until the NSSDC has ensured that the delivery meets certain criteria. These requirements are derived from the PDS Level 3 requirements [1], PDS Tracking Use Cases [3], and PDS Policies.

7.2.1 General (GR)

L4.AT.GR.1 – The PDS shall be able to provide an accounting of all files in the archive (2.2.2, 2.6.3, 2.8.2, PDS Policy, 2.2.2 - see Policies Section)

L4.AT.GR.2 – The PDS shall be able to verify that all files in the archive are accessible. (2.2.2, 2.6.3, 2.8.2)

> **Accessible is used in the context of the above requirement to ensure the file can be "located and accessed" in the archive repository.**

L4.AT.GR.3 – PDS shall preserve information required to verify the original file delivered to PDS node has not been corrupted. (2.2.2, 2.6.3, 2.8.2)

> **A possible implementation would be maintaining the original checksum value for each file and using that value throughout the end-to-end system.**

L4.AT.GR.4 – The PDS shall maintain status and location information for each data set and volume in the archive in a common catalog. (2.2.2, 2.6.3, 2.8.2).

> **The PDS Catalog tracks data collections including data sets, volumes, and data deliveries. Specific information beyond data collection information to be tracked includes the curating_node_id, distribution_node_id, backup_node_id, archive_status, acceptance_status, delivery_status, nssdc_id, etc.**

L4.AT.GR.5 – The PDS shall notify actors about events including scheduled events associated with deliveries within or from the PDS archive and any change in delivery or acceptance / rejection statuses. (3.3.6)

> **This requirement focuses primarily on the NSSDC interface. The PDS Notification/Subscription system could be extended to fulfill this requirement. The notification of an actor will be triggered by an event. Various requirements in the following sections will signal events. The delivery plan will provide a list of actors to be notified for specific events. The subset of "authorized" actors will also be provided in the delivery plan.**

7.2.2 Data Transfer between PDS Nodes (TR)

L4.AT.TR.1 – The PDS shall be able to track the transfer of files from one PDS Node archive repository to another (2.2.2, 2.6.3, 2.8.2).

7.2.3 Preservation of Data at NSSDC (NS)

L4.AT.NS.1 – The PDS shall be able to track the transfer of a data set from a PDS Node to the NSSDC for long term preservation. (2.2.2, 2.6.3, 2.8.2, UC-3)

> **Possible data delivery events include Transfer to NSSDC. Each delivery event can have one or more event statuses which are captured by the NSSDC_Transfer_Status. On integration with the PDS Notification system, these delivery events will be added to the already existing Release and Archive events.**
>
> **A related Data Integrity requirement is L4.DI.7.**

7.2.4 Data Inventory Reporting (IR)

L4.AT.IR.1 – PDS shall generate reports on its data holdings including the location, status, and integrity of data deliveries, data volumes, data sets, data products and their associated files. These reports shall be generated periodically according to applicable PDS policies and also be available for generation on demand. (2.2.2, 2.6.3, 2.8.2, UC-6, PDS Policy, 2.2.2 - see Policies Section)

> **The above requirement is inclusive of files that are descriptive of the volume or data set but are considered ancillary (e.g., errata.txt and files located in extras directory).**

8 Data Availability Level 4 Requirements

The following requirements for data availability are derived from the PDS Level 3 Requirements [1] and the Data Availability Use Cases [4].

Data availability addresses only PDS archive repositories and not the underlying software systems provided by the nodes for distributing the data from the repositories. These repositories include the primary, secondary and tertiary repositories and involve all nodes, the NSSDC, and sites used for secondary repositories or backups. This supports the “three copy” rule that PDS imposed with the October 2005 policy on data delivery.

Availability of data is supported by two critical PDS Level 3 requirements that identify the need for “continuous operations” (2.10.1) and “disaster recovery” (4.1.4). From the perspective of providing both continuous operations and disaster recovery for the archive, the use cases described in [4] were based on several assumptions. These assumptions were reviewed by the Management Council at the August 2007 meeting. In some instances, a trade off between integrity of the archive and quality of service needs to be made. In those instances, the emphasis is on the integrity of the archive.

For the purposes of data availability, 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) (source: PDS Policy on Data Delivery, October 2005).

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 (source: PDS Management Council discussion, August 2007).

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 (source: various PDS Management Council discussions)

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 (source: PDS policy on backup and recovery).

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.

6) The concept of "continuous operations" requires that the PDS have operating requirements that strive to achieve a maximum "quality of service" (QoS) rating for its users. The following operational scenarios are provided together with their assumptions in order to provide a basis for the availability requirements.

a) 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.

b) Routine Operational Scenario[1]

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

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

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

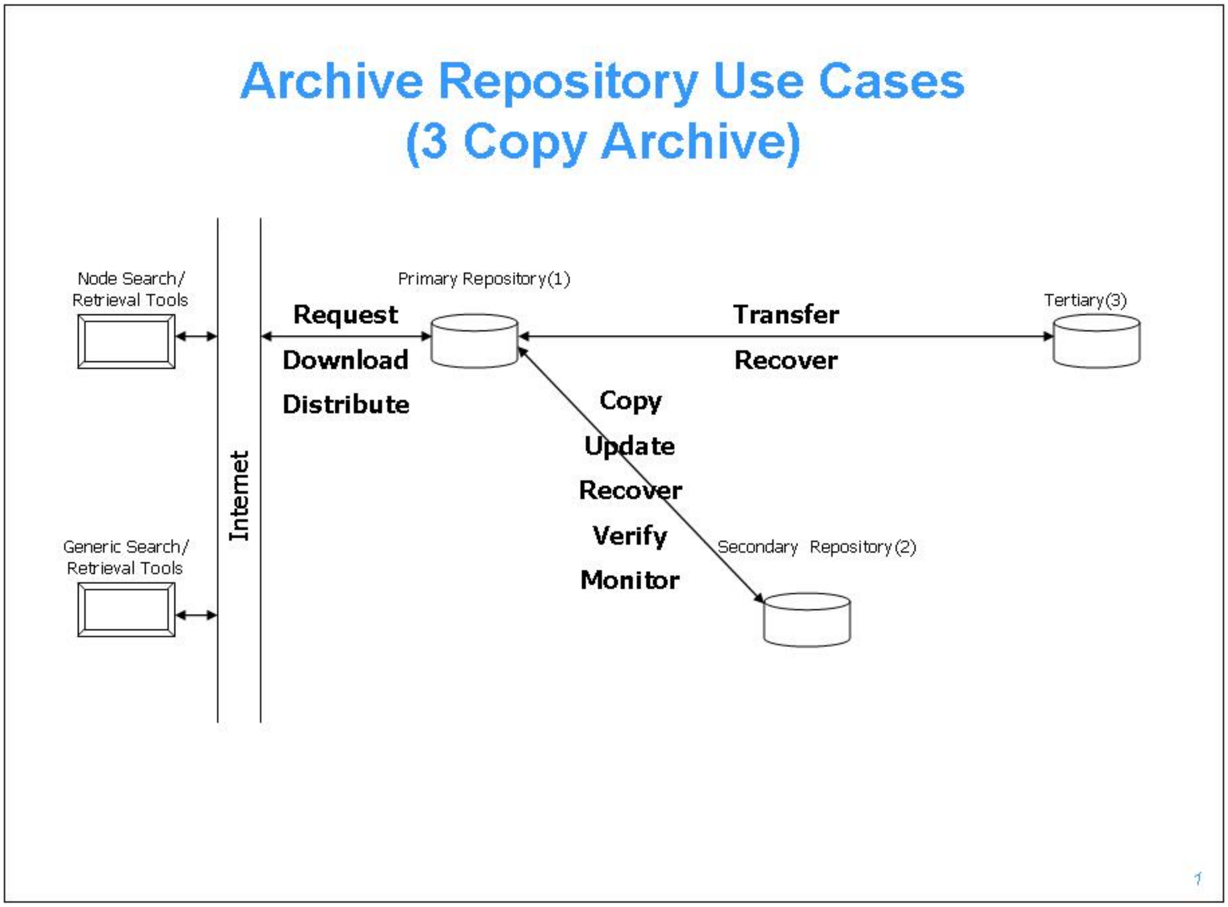
c) Loss-of-data Scenario – In case of a loss-of-data event at a node but where operational capability is not impaired, restoration of the data from a backup should occur within 1 week.

d) Catastrophic Scenario[2] - 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] An MC member suggested the restoration of off-line data could be reduced to 24 hours and that delays from weekends/holidays could be made implicit.

[2] An MC member argued that the catastrophic recovery time should be phrased in terms of availability of funding; It was countered that funding is really a management issue and shouldn't necessarily be folded into requirements.

Figure 8-1 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.



Archive Repository Use Cases
(3 Copy Archive)

Figure 8-1 – Archive Repositories

The following diagram illustrates how Data Availability plays a role in the flow of data into, across, and out of the PDS.

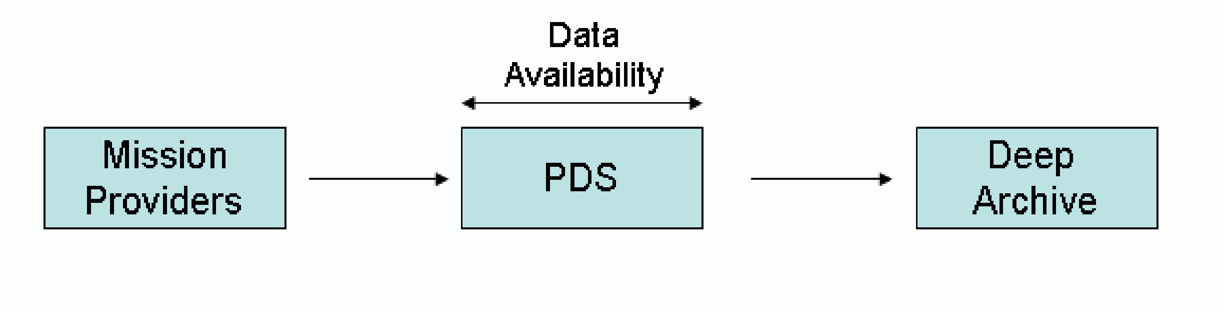


Figure 8-2 – Data Availability Flow Diagram

Figure 8-2, Data Availability Flow Diagram, depicts a “throughput” of files into the PDS and out of the PDS. Data Providers supply “data”, in the form of files, to PDS. PDS supplies “data”, in the form of files, to the Deep Archive. Files are “available” from the point when PDS receives the “data” from the Data Providers throughout the life of the data residing within the PDS.

The following identify the Level 4 requirements for data availability. The specific subcategories of these requirements include (General Requirement (GR), Secondary Requirement (SR), and Tertiary Requirement (TR)).

L4.AV.GR.1 – PDS shall ensure that data of high interest to the world-wide Planetary Science community has online access with minimal downtime. (2.7.2, 2.7.4, 2.8.1, 2.10.1, 3.2.1, UC-4)

L4.AV.SR.1 – PDS shall have a secondary copy of all archived data at one or more facilities at geographically distinct locations in order to support continuous operations.
(2.10.1, 3.2.2, UC-1, UC-6)

L4.AV.SR.2 – PDS shall verify that a secondary copy of data is available for the successful recovery of data in a primary repository. (2.10.1, 4.1.2, UC-5, UC-6)

> **Note: Level 5 requirements will address the tests necessary to ensure successful recovery including data integrity and access mechanisms.**

L4.AV.SR.3 – PDS shall ensure that all secondary copies of data are synchronized[3] with their primary copies. (2.7.2, 2.10.1, UC-2, UC-6)

[3] MC consensus is that "synchronized" copies means periodically doing spot check comparisons, including the *single* NSSDC copy, to ensure that the copies are the same and accessible. This does not require that every data set be checked within any specific time frame nor does it require that PDS delve into NSSDC backup systems. MC Meeting 8/6/07, D. Simpson.

L4.AV.SR.4 – PDS shall maintain operational procedures for recovering files for the primary repository from the secondary copies. (4.1.4, UC-3, UC-4)

L4.AV.TR.1 – PDS shall deliver a tertiary copy of all archived data to an offsite location that meets U.S. federal regulations for preservation and management of the data. (UC-7, UC-9)

L4.AV.TR.2 – PDS shall verify that a tertiary copy of data is available for the successful recovery of data in a primary repository. (2.7.2, 2.10.1,UC-5)

> **Note: Level 5 requirements will address the types of testing necessary to ensure that the recovery interface works.**

L4.AV.TR.3 – PDS shall ensure that tertiary copies of data are synchronized[3] with their primary copies. (2.10.1, UC-2, UC-7)

L4.AV.TR.4 – PDS shall maintain operational procedures for recovering files for the primary repository from the tertiary copies. (4.1.4, UC-3, UC-4)

9 Preservation Planning Level 4 Requirements

The following requirements for preservation planning are derived from the PDS Level 3 Requirements [1].

Preservation Planning principally addresses the process whereby PDS will audit the deep archive, NSSDC, to ensure the integrity of the PDS data holdings at NSSDC. The audit will focus on:

1. Ensuring no file corruption during preservation
2. Ensuring accountability for all files transferred
3. Ensuring files are accessible

The following diagram illustrates how Preservation Planning plays a role in the flow of data into, across, and out of the PDS.

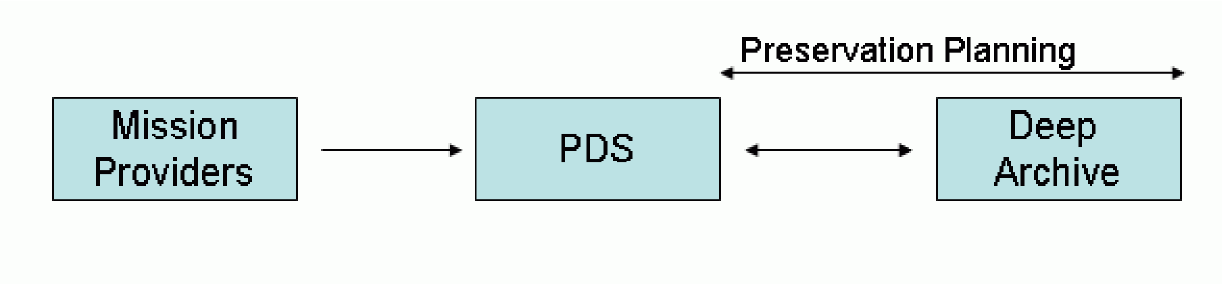


Figure 9-4 – Preservation Planning Flow Diagram

Figure 9-1, Preservation Planning Flow Diagram, depicts a “throughput” of files into the PDS and out of the PDS. Data Providers supply “data”, in the form of files, to PDS. PDS supplies “data”, in the form of files, to the Deep Archive. The Deep Archive supplies “data”, in the form of files, to the PDS. “Preservation Planning” ensures the file transfer interface between PDS and the Deep Archive is preserved (i.e., the PDS and the Deep Archive can reconcile PDS holdings submitted to the deep archive for corruption, accountability, and availability).

L4.PP.1 – PDS shall periodically audit the deep archive to ensure that the archival data holdings, submitted to the deep archive for preservation, are not corrupted. (4.1.1)

> **Note: Level 5 requirements will further address how file level corruption will be identified and reported.**

L4.PP.2 – PDS shall periodically audit the deep archive to ensure that the archival data holdings, submitted to the deep archive for preservation, are accounted for. (4.1.1)

> **Note: Level 5 requirements will further address how files, presumed to have been transferred to the deep archive, will be identified and reported as "missing from the inventory".**

L4.PP.3 – PDS shall periodically audit the deep archive to ensure that the archival data holdings, submitted to the deep archive for preservation, are accessible. (4.1.1)

> **Note: Level 5 requirements will further address how files, presumed to have been transferred to the deep archive, will be identified and reported as "not accessible / locatable".**

L4.PP.4 – PDS shall periodically reconcile PDS holdings submitted to the deep archive against the holdings stored in the deep archive. (4.1.1)

> **The PDS will generate an inventory of PDS holdings submitted to NSSDC. NSSDC will generate an inventory of holdings received from the PDS. PDS and NSSDC will reconcile the lists to ensure there are no discrepancies.**

L4.PP.5 - PDS shall provide procedures for auditing the deep archive. (4.1.1)

APPENDIX A L4 TO L3 TRACEABILITY MATRIX

The traceability matrix maps the Level 4 requirements in this document to the Level 3 requirements from which the Level 4 requirements were derived.

Level 4 Requirement	Derived From	Use Case
L4.DI.1	2.7.2, 2.7.4	UC-1
L4.DI.2	2.5.1, 2.5.2, 2.7.1	UC-1
L4.DI.3	2.5.2	UC-1, UC-4, UC-8
L4.DI.4	2.8.2, 4.1.2	UC-5
L4.DI.5	4.1.3	UC-6
L4.DI.6	4.1.3	UC-6
L4.DI.7	3.2.3	UC-3
L4.DI.8	3.2.3	UC-7
L4.DI.9	3.2.3	UC-7
LR.DT.GR.1	2.2.2	UC-1, UC-1.2
LR.DT.GR.2	2.2.2	
LR.DT.GR.3	3.3.6	UC-6
LR.DT.GR.4		UC-1.2, UC-1.3
LR.DT.GR.5	2.2.2	UC-4, UC-5
LR.DT.DP.1	2.2.2, 2.8.2	UC-1.3, UC-6
LR.DT.DP.2	1.3.4, 1.3.5, 2.2.2	UC-1.2
LR.DT.DP.3	2.2.2, 2.5.2, 2.5.3	UC-6
LR.DT.DP.4	2.2.2, 2.5.2, 2.5.3	UC-1.3, UC-2
LR.DT.NT.1	2.2.2	UC-3
L4.AT.GR.1	2.2.2, 2.6.3, 2.8.2	PDS Policy
L4.AT.GR.2	2.2.2, 2.6.3, 2.8.2	
L4.AT.GR.3	2.2.2., 2.6.3, 2.8.2	
L4.AT.GR.4	2.2.2, 2.6.3, 2.8.2	
L4.AT.GR.5	3.3.6	
L4.AT.TR.1	2.2.2, 2.6.3, 2.8.2	
L4.AT.NS.1	2.2.2, 2.6.3, 2.8.2	UC-3
L4.AT.IR.1	2.2.2, 2.6.3, 2.8.2	UC-6; PDS Policy
L4.AV.GR.1	2.7.2, 2.7.4, 2.8.1, 2.10.1, 3.2.1	UC-4
L4.AV.SR.1	2.10.1, 3.2.2	UC-1, UC-6
L4.AV.SR.2	2.10.1, 4.1.2	UC-5, UC-6
L4.AV.SR.3	2.7.2, 2.10.1	UC-2, UC-6
L4.AV.SR.4	4.1.4	UC-3, UC-4
L4.AV.TR.1		UC-7, UC-9
L4.AV.TR.2	2.7.2, 2.10.1	UC-5
L4.AV.TR.3	2.10.1	UC-2, UC-7
L4.AV.TR.4	4.1.4	UC-3, UC-4
L4.PP.1	4.1.1	

Level 4 Requirement	Derived From	Use Case
L4.PP.2	4.1.1	
L4.PP.3	4.1.1	
L4.PP.4	4.1.1	
L4.PP.5	4.1.1	

APPENDIX B L3 TO L4 TRACEABILITY MATRIX

The traceability matrix maps the Level 4 requirements in this document to the Level 3 requirements from which the Level 4 requirements were derived.

Level 3 Requirement	Derives Level 4
1.3.4	L4.DT.DP.2
1.3.5	L4.DT.DP.2
2.2.2	L4.DT.GR.1, L4.DT.GR.2, L4.DT.GR.5, L4.DT.DP.1, L4.DT.DP.2, L4.DT.DP.3, L4.DT.DP.4, L4.DT.NT.1, L4.AT.GR.1, L4.AT.GR.2, L4.AT.GR.3, L4.AT.GR.4, L4.AT.TR.1, L4.AT.NS.1, L4.AT.IR.1
2.5.1	L4.DI.2
2.5.2	L4.DI.2, L4.DI.3, L4.DT.DP.3, L4.DT.DP.4
2.5.3	L4.DT.DP.3, .DT.DP.4
2.6.3	L4.AT.GR.1, L4.AT.GR.2, L4.AT.GR.3, L4.AT.GR.4, L4.AT.TR.1, L4.AT.NS.1
2.7.1	L4.DI.2
2.7.2	L4.DI.1, L4.AV.GR.1, L4.AV.SR.3, L4.AV.TR.2
2.7.4	L4.DI.1, L4.AV.GR.1
2.8.1	L4.AV.GR.1
2.8.2	L4.DI.4, L4.DT.DP.1, L4.AT.GR.1, L4.AT.GR.2, L4.AT.GR.3, L4.AT.GR.4, L4.AT.TR.1, L4.AT.NS.1, L4.AT.IR.1
2.10.1	L4.AV.GR.1, L4.AV.SR.1, L4.AV.SR.2, L4.AV.SR.3, L4.AV.TR.2, L4.AV.TR.3
3.2.1	L4.AV.GR.1
3.2.2	L4.AV.SR.1
3.2.3	L4.DI.7, L4.DI.8, L4.DI.9
3.3.6	L4.DT.GR.3, L4.AT.GR.5
4.1.1	L4.PP.1, L4.PP.2, L4.PP.3, L4.PP.4, L4.PP.5
4.1.2	L4.DI.4, L4.AV.SR.2
4.1.3	L4.DI.5, L4.DI.6
4.1.4	L4.AV.SR.4, L4.AV.TR.4
4.1.5	