



Update from the Data Integrity & Tracking WG

Management Council F2F

UCLA

Los Angeles, CA

August 13-14, 2007

<http://pds.nasa.gov>



Scope and Background



- **Management Council Action:**
 - “Crichton agreed to chair an MC working group on data integrity. New suggested starting with Level 4 requirements before dealing with any implementation issues. The DIWG should have recommendations before the Tech Session tackles the technical issues in data integrity at its proposed face-to-face meeting. The Tech Session can then decide where to go with the existing SCR on checksums, including sub-issues of specificity and process.” - Aug 2006 Action
 - “The DIWG will continue on the following: recommending an implementation for data integrity to the tech staff and then the MC, and developing requirements for disaster recovery and archive tracking (including the tracking system).” (November 2006)
- **Members**
 - Dan Crichton, EN
 - Mitch Gordon, Rings
 - Ed Guinness, Geosciences
 - Bill Harris, PPI
 - Todd King, PPI
 - Steve Hughes, EN
 - Chris Isbell, Imaging
 - Myche McAulley, Imaging
 - Al Schultz, GSFC
 - Mark Showalter, Rings
 - Tom Stein, Geosciences



Scope and Background

- PDS Policy on Data Delivery and Backup
 - 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.
- Archive Integrity Policy
 - 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. (Adopted by MC November 2006)



Scope Cont...



- **Integrity**
 - Scope: File corruption during transfer or in repository
 - PDS Integrity Use Cases
 - PDS Integrity Level 4 Requirements
- **Tracking**
 - Scope: Tracking of files and collections from data producer through to NSSDC
 - PDS Tracking Use Cases
 - PDS Tracking Level 4 Requirements
 - PDS Tracking Level 5, and 6 Supporting Requirements
- **Availability**
 - Scope: Availability of data holdings from a Primary Repository with contingencies to recover data from secondary (Backup) and tertiary (Deep Archive) repositories
 - PDS Availability Use Cases
 - PDS Availability Level 4 Requirements (In progress)



Availability Use Cases



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

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

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

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

UC-5: PDS verifies the integrity of the PDS secondary repository



Availability L4 Requirements

1. PDS shall ensure that data of high interest to the world-wide Planetary Science community has online access with minimal downtime. [2.10.1, 2.8.1]
2. 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] [UC-1]
3. 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]
 - Note: Level 5 requirements will address the tests necessary to ensure successful recovery including data integrity and access mechanisms.
4. PDS shall ensure that all secondary copies of data are synchronized with their primary copies. [2.10.1, UC-2]
5. PDS shall maintain operational procedures for recovering files for the primary repository from the secondary copies. [4.1.4, UC-3, UC-4]
6. 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. [4.1.5]
7. PDS shall verify that a tertiary copy of data is available for the successful recovery of data in a primary repository. [2.10.1, 4.1.2, UC-5]
 - Note: Level 5 requirements will address the types of testing necessary to ensure that the recovery interface works.
8. PDS shall ensure that all tertiary copies of data are synchronized with their primary copies. [2.10.1, UC-2]
9. PDS shall maintain operational procedures for recovering files for the primary repository from the tertiary copies. [4.1.4, UC-3, UC-4]



Availability Assumptions (1 of 3)



- From the perspective of providing both continuous operations (2.10.1) and disaster recovery (4.1.4) for the archive, the availability use cases and requirements 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.



Availability Assumptions (2 of 3)



- 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) at the NSSDC.
 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, any single data set needs to be in more than one repository at more than one geographically distinct location. If assumption 1 above is acceptable then there is always a copy of any single data set at three geographically distinct locations.
 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 a geographically distinct location for the secondary 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. It is also assumed that the PDS Management Council will want the recovery interface tested.



Availability Assumptions (3 of 3)

- 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
 - An on-line primary data repository at a node should never be unavailable for longer than 24 hours.
 - An off-line primary data repository at node should be able to make data available for distribution to a user within 72 hours.
 - 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 a least the retrieval of data files using file identifiers over simple internet and file system protocols.



Timeline

- Present WG progress status on requirements to MC (Aug 2007)
- Address final comments and suggestions (Aug 2007)
- Working Group to request review of the level 4 requirements by the PDS (Aug 2007)
 - Comments subsequently addressed by the WG
- Working Group to recommend acceptance of level 4 requirements for data integrity and tracking by the MC (Sep 2007)
- Implementation plan developed and presented to MC (next F2F)