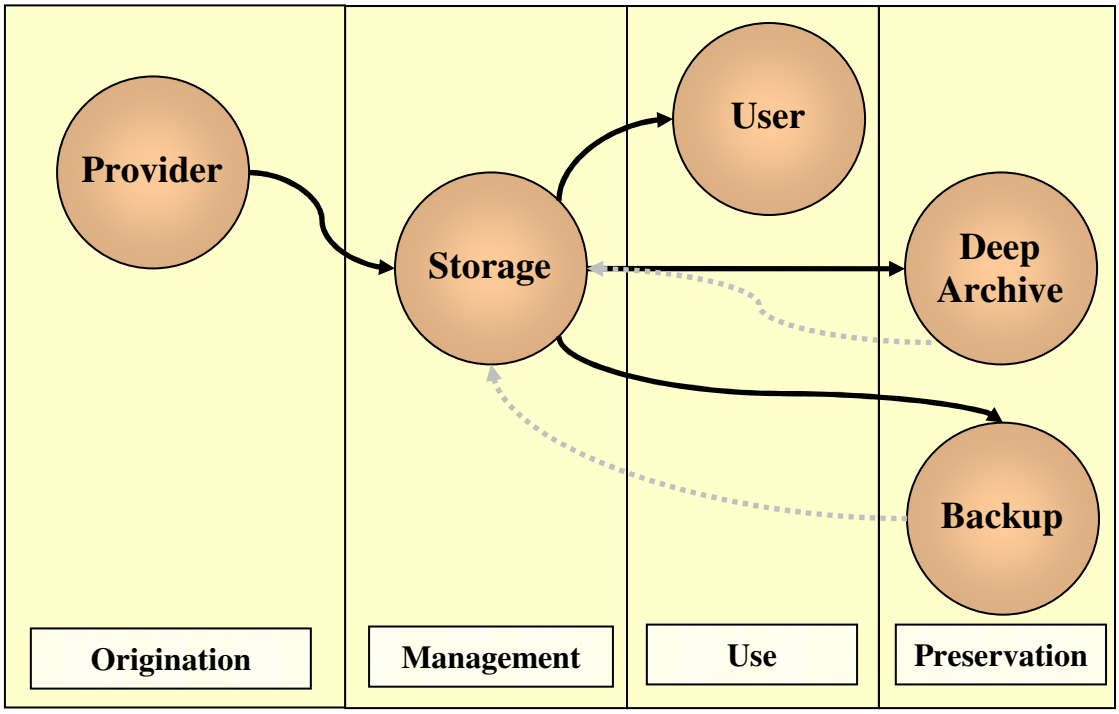
PDS PPI Node
Data Integrity Process
June 19, 2008

Overview

The PDS PPI Node data integrity process is a set of linear processes which check the integrity of the data that is delivered to and maintained by the Node. Each process ensures the integrity of the data at a single phase of the data life cycle. The data lifecycle consists of four stages:

Origination: The delivery of data from a provider to the Node for storage.
Management: Ensuring that all files in the holding accounted for and intact.
Use: The delivery of the data to a user.
Preservation: The transfer of all or part of the data holdings to an alternative storage location.

The lifecycle stages, roles and information flow in depicted in following figure.

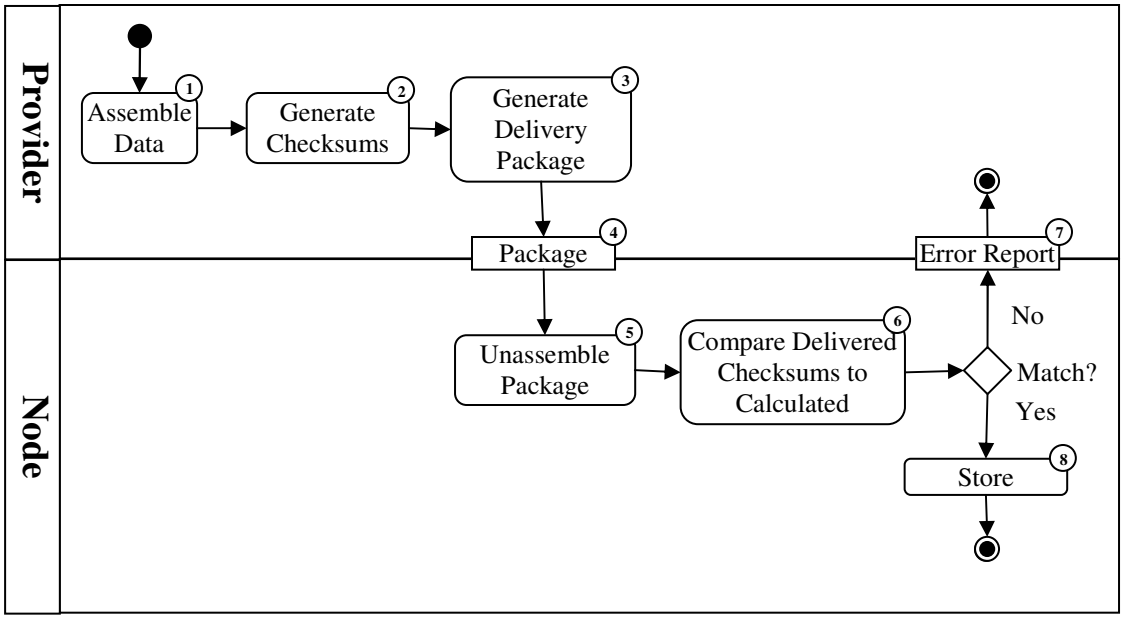


Process Details

The process of data movement, indicated by arrows in the lifecycle figure, is different between each lifecycle phase. This section describes each process.

Origination

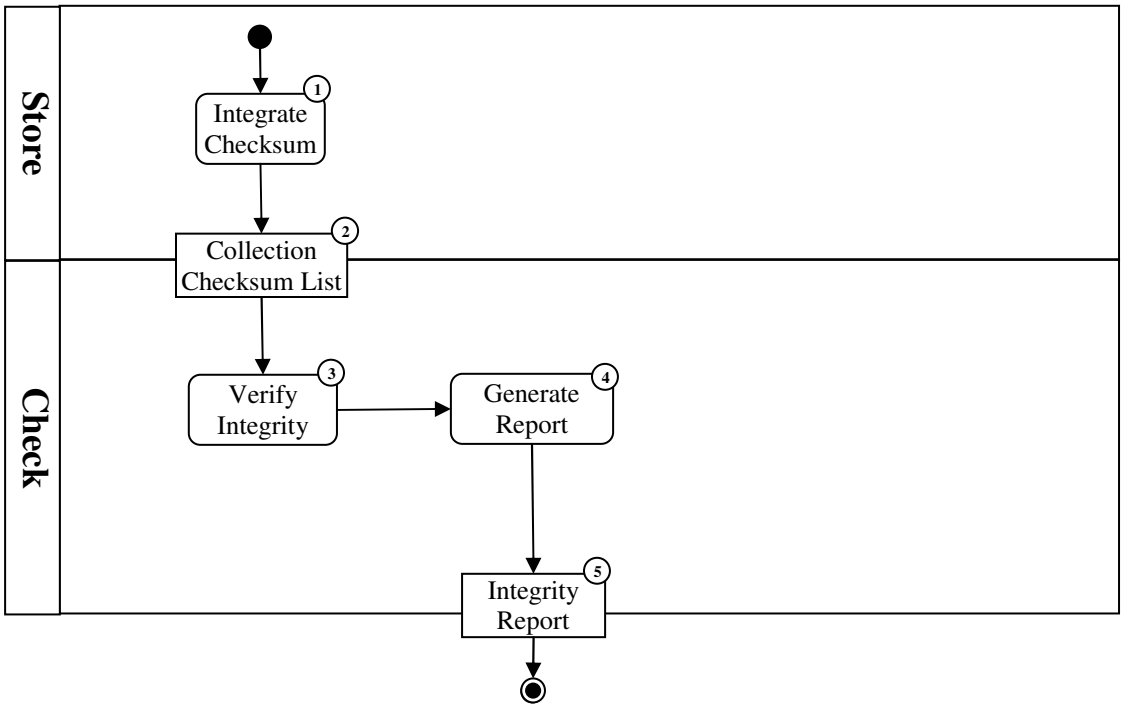
The Origination process is the delivery of data from a provider to the Node. Generated checksum are MD5 and stored in an md5sum compatible format. Packages are typically zip compatible and generated either using the tool preferred by the provider. Packages generated using the "tar" format is also acceptable.



1. **Assemble Data.** The provider locates the data which is stored in files. The files are organized in a branch of the file system. Folders are created according to PDS conventions. Documents in the "Document" folder, "data" in the "Data" folder, etc.
2. **Generate Checksums.** A list of checksums is generated for the files to be delivered. The tool used to generate the checksums is "md5deep" or the PPI tool "md5check" which generates a list consisting of one line for each file composed of an MD5 checksum, two spaces and the relative path to the file.
3. **Generate Delivery Package.** The branch of the file system to be delivered is packaged into a single file. Typically using a "zip" program (zip, gzip, etc). Packages using "tar" are also acceptable. The checksum list is included in the package. If the number of files is small individual files may be delivered without packaging.
4. **Package.** The package of files is delivered to the Node. The preferred method of transfer is using FTP, web uploads are also allowed and large deliveries may be transferred using data bricks.
5. **Unassemble Package.** The delivery package is unassembled in a temporary area.
6. Compare Checksums. The checksums delivered with the package are compared to the files to ensure intact delivery.
7. **Error Report.** Any mismatched are logged in an error report which is delivered back to the provider.
8. **Store.** Verified files are stored in the archive system.

Manage

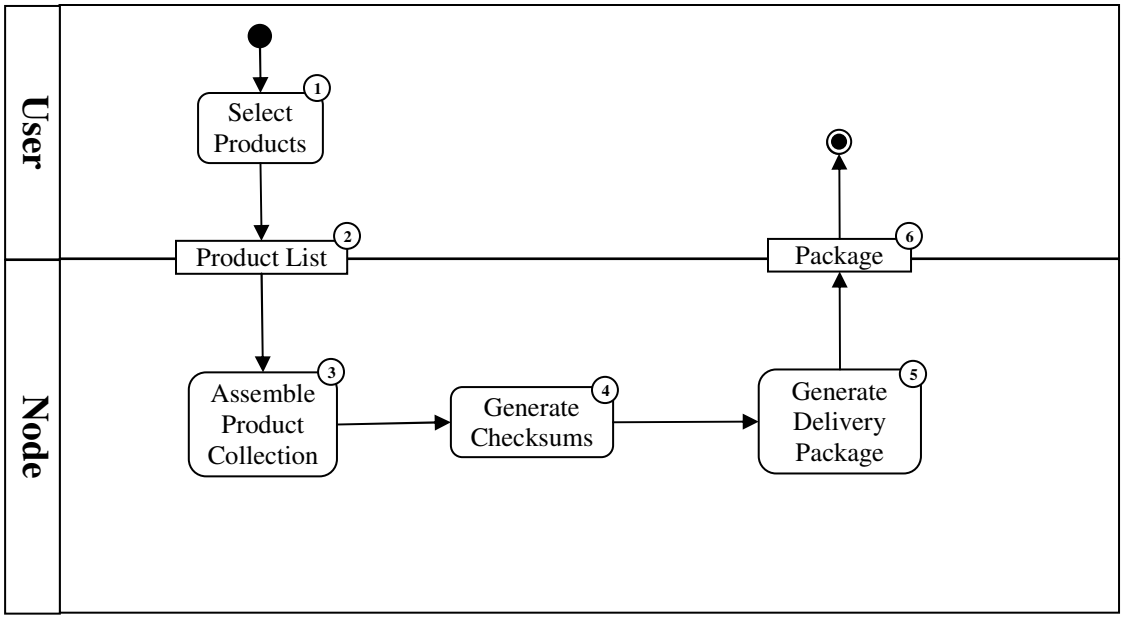
The Manage process involves checking the integrity of the data holdings and ensuring that all files in the holding are properly accounted. This involves ensuring that every file in the data holding has a checksum maintained in a checksum list. The data holdings are divided into collections (datasets, volumes, etc.) and each collection has a corresponding checksum list.



1. **Integrate Checksum.** For each file added to the archive a checksum is calculated and integrated with the checksum list for the collection. The archive is divided into logical collections (datasets, volumes, etc). Deliveries to the Node may be entire collections or portions of a collection (see Origination process).
2. **Collection Checksum List.** Each collection has a single checksum list. The checksum list consists of a checksum entry for every file in the collection.
3. **Verify Integrity.** Periodically the checksum list for the collection is compared to the calculated checksums for each file in the collection. The verification process determines if checksums match, if files are missing or new files exist in the collections which do not have entries in the checksum list. We use the "md5check" created at PPI to perform this function.
4. **Generate Report.** The results of the integrity check are collected into a report.
5. **Integrity Report.** The integrity report is delivered.

Use

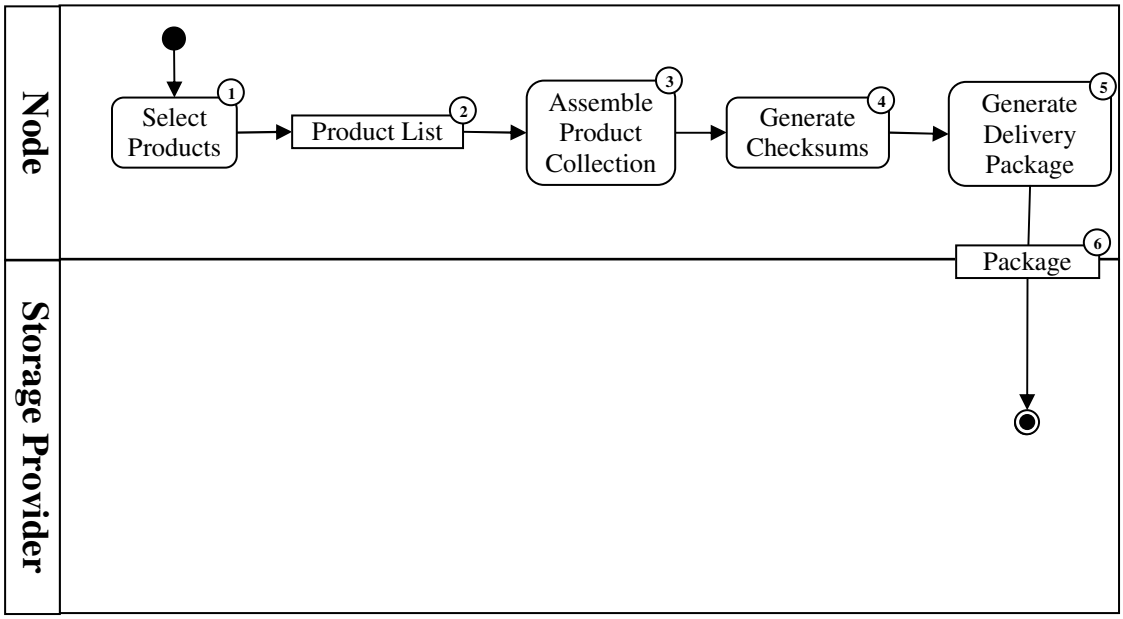
The Use process involves the delivery of the data to a user. This is symmetrical to the "Origination" process and uses many of the same tools and services. Generated checksum are MD5 and stored in an md5sum compatible format. Packages are zip compatible and generated either using "gzip" or by embedded zip generation in a service.



1. **Select Products.** The user selects the desired products typically through a web interface, but it can be through an e-mail or verbal communication. The selected products could be for an entire collection (dataset, volume, etc.) or for a set of individual products.
2. **Product List.** The products selected are specified in a product list.
3. **Assemble Product Collection.** Services or individuals assemble the product collection to match the product list.
4. **Generate Checksums.** A list of checksums is generated for the files to be delivered. The tool used to generate the checksums is "md5deep" or the PPI tool "md5check" which generates a list consisting of one line for each file composed of an MD5 checksum, two spaces and the relative path to the file. If the product collection is assembled by a service the checksum generation may be integrated into the service. This is the case with the web interface.
5. **Generate Delivery Package.** The branch of the file system to be delivered is packaged into a single file using a "zip" program (zip, gzip, etc). The checksum list is included in the package. If the number of files is small individual files may be delivered without packaging.
6. **Package.** The package of files is delivered to the User. The file may be streamed over a web collection or posted on an FTP server for subsequent retrieval by the user. In some cases, the package may be placed on a DVD-R and mailed to the user.

Preservation

The Preservation process involves the transfer of all or part of the data holdings to an alternative storage location. Each storage location employs local preservation rules Depending on the local preservation rules the alterative storage location may be deemed a backup or a deep (permanent) archive. The process is similar to the "Use" process with the Node selecting the products, but the package is delivered to the storage location. The storage provider may require specific packaging or the user of a particular service to transfer the products.



1. **Select Products.** The Node selects the desired products to send to the storage provider. The selected products could be for an entire collection (dataset, volume, etc.) or for a set of individual products.
2. **Product List.** The products selected are specified in a product list.
3. **Assemble Product Collection.** The product collection which matches the product list is assembled. For some storage providers this requires using a particular tool or service.
4. **Generate Checksums.** A list of checksums is generated for the files to be delivered. Some storage provider tools generate the checksums as part of the package generation or transfer process. If checksums are to be included as a separate file they are generated using md5deep or the PPI tool md5check.
5. **Generate Delivery Package.** The branch of the file system to be delivered is packaged into a single file using a "zip" program (zip, gzip, etc). The checksum list is included in the package. If the number of files is small individual files may be delivered without packaging.
6. **Package.** The package of files is delivered to the Storage Provider. The file may be delivered to an FTP server, streamed using a Storage Provider tool or delivered on a data brick.