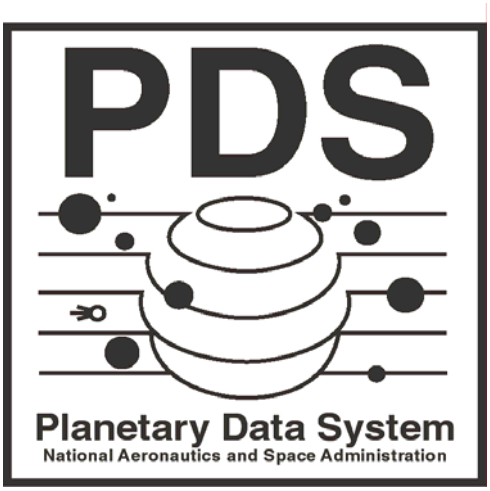
Planetary Data System

Ingest, Validate, and Track (IV&T) Concept of Operations

DRAFT

March 15, 2005

Version 0.050315



Jet Propulsion Laboratory
Pasadena, California

JPL D-xxxxx

Planetary Data System

Ingest, Validate, and Track (IV&T) Concept of Operations

March 15, 2005

Version 0.050315

Approved:

_____	_____
Bill Knopf PDS Program Executive	Date
_____	_____
Tom Yunck PDS Program Manager	Date
_____	_____
Reta Beebe PDS Project Scientist	Date
_____	_____
Steve Hughes PDS Project Engineer	Date

Jet Propulsion Laboratory
Pasadena, California

JPL D-xxxxx

CHANGE LOG

Revision	Date	Description	Author
Start Draft	2004-08-01	Initial outline	R. Joyner/S. Hughes
Draft Release 1	2005-01-07	DRAFT distributed to Nodes	R. Joyner/S. Hughes
Draft Release 2	2005-02-01	Submittals from Tiger Team	IV&T Tiger Team
Draft Release 3	2005-03-03	Distributed to Node Mgrs for review	R. Beebe
Draft Release 4	2005-03-10	Added Background section	R.Joyner / R.Beebe
Draft Release 5	2005-03-15	Updated Background section	IV&T Tiger Team

TABLE OF CONTENTS

1.0 INTRODUCTION

1.1 Purpose

At a high level, the PDS Ingest, Validate, and Track (IV&T) services will transfer, validate, and report on the status of data products being delivered to the Planetary Data System (PDS).

1.2 Scope

This document specifies a high level concept of operations required to fulfill the purpose. It describes the role and function of each service and how products and information move through the system.

1.3 Background

Given the predicted increase in both the VOLUME and NUMBER of data products in the coming 5-10 years, PDS was asked to identify alternatives for ensuring adherence to measures of quality assurance for the data being produced by all missions.

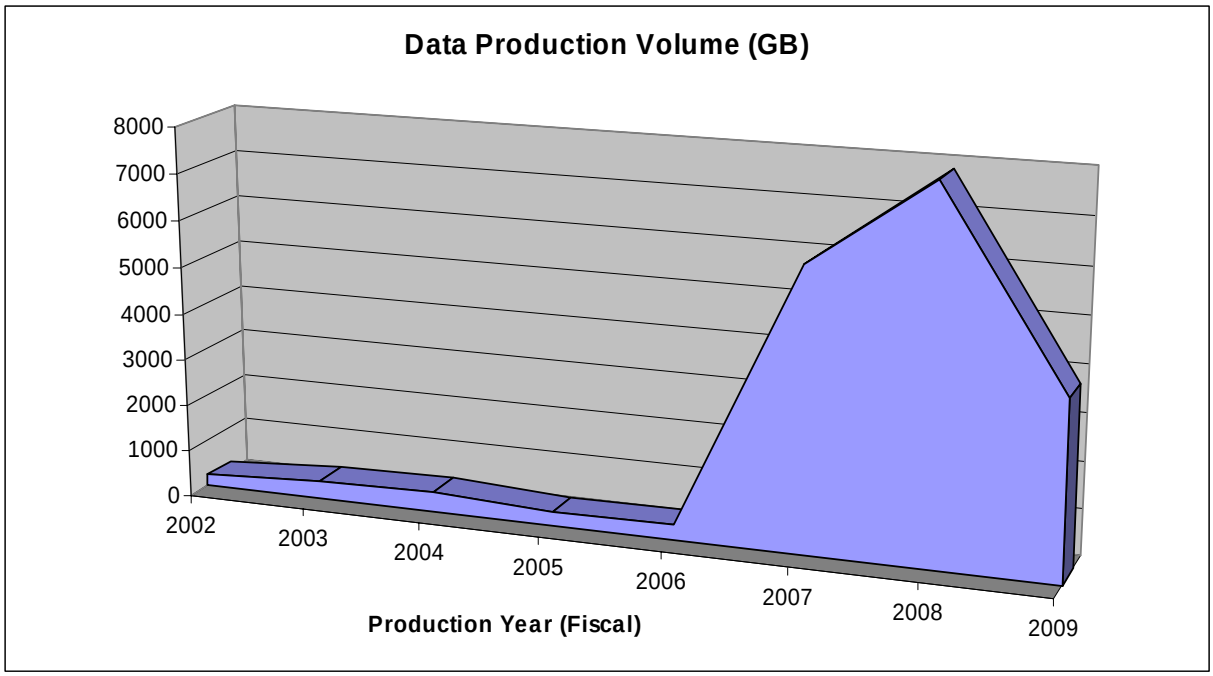


Figure 1.1 Predicted Volume of Data Products

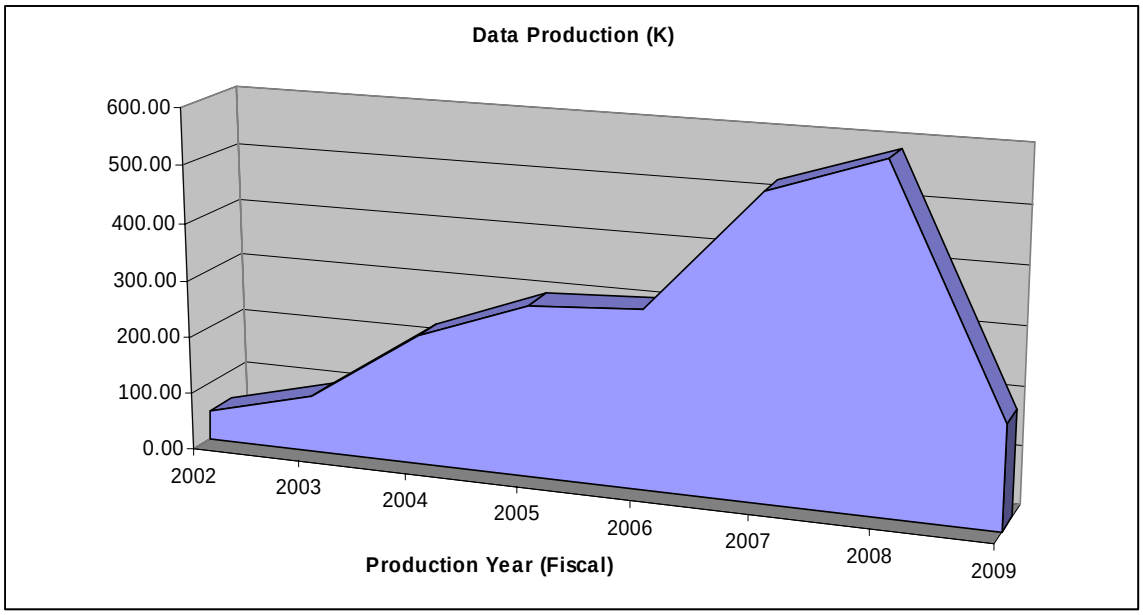


Figure 1.2 Predicted Number of Data Products

Note: The above figures indicate data production declining in the later years because data production estimates are not available for some of the future missions such as the Lunar Reconnaissance Orbiter mission (LRO), which is scheduled to launch in 2008.

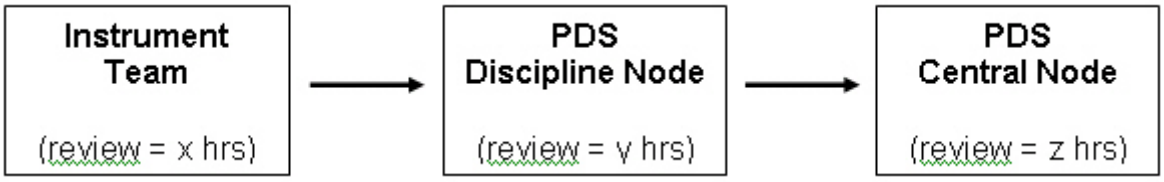
1.3.1 Current Method

Currently, ensuring adherence to measures of quality assurance is a largely manual effort using a mixed suite of tools:

- Approximately 85% of validation and ingestion is manual
- The remaining 15% is machine-assisted
- There are no automatic processes within the PDS

The entire suite of validation tools can be run against a typical archive volume within 10 minutes. Interpreting the output from those tools can consume multiple days and may require the assistance of someone who is an “expert” in PDS Standards. Different instrument teams from the same mission may encounter different validation and ingestion methods simply because they are interacting with different Nodes.

To overcome the above limitations and maintain a measure of quality assurance, an approach was adopted where data products would pass through multiple (semi-redundant) reviews:



1.3.2 Proposed Method

The goal is to develop a set of modular tools to improve data validation and ingestion (i.e., develop functional components that can be configured into a highly automated ingestion pipeline). The end result is an ingestion workflow that is uniform across the entire PDS and that will reduce or eliminate the necessity for redundantly reviewing data products by moving the ingestion process up stream towards the instrument teams).

The tools will serve three functions; ingestion, validation and tracking of data.

Validation tools - the Validation tools will be designed so that they can be provided to individual instrument teams and integrated as early as possible into the mission instrument data pipe lines . The same tools should be used within the PDS during ingestion. These tools should automate validation of standard characteristics and facilitate delivery of PDS conformant data , leaving validation of scientific integrity to the discipline node.

Ingestion tools – the Ingestion tools will be consistent across nodes and modular to allow flexibility during development phases of data sets.

Tracking tools – the Tracking tools will provide the capability to allow both the PDS and the missions to ascertain when deliveries are due, when they are late, when PDS is late in ingestion/validation, etc. The Tracking tools will generate archiving status reports for missions and nodes, issue early warnings of expected data deliveries and notices of late deliveries.

IV&T Services will additionally:

- Provide a consistent ingestion method across the PDS (i.e., ensure different instruments from the same mission will not encounter different ingestion methods simply because they are interacting with different Nodes)
- Address complexity associated with increases in product types via scaleable tools / process flow
- Provide more rigorous control over validating data products
- Provide a portable and configurable generic validation workflow
  - Provide functional components that can be deployed individually for use in existing instrument data pipelines
- Ensure adherence to standards and documented agreements via multiple levels of compliance.
  - Improve ability to track changes
- Provide concise and more intuitive IV&T reports (i.e., reports which are useable by those who are not experts in PDS Standards).

1.4 Governing Documents / Standards / Specifications

The IV&T services will conform to the following standards / specifications:

- CCSDS 650.0-R-2 “Reference Model for an Open Archival Information System (OAIS),” Red Book, July 2001. [http://ccsds.org/documents/pdf/CCSDS-650.0-R-2.pdf]

2.0 DEFINITIONS OF TERMS

This section introduces key terms and concepts as they are used within this document. They are presented here in order to facilitate a clear and concise understanding of the material which follows.

- **Repository** – A place / location where products are stored.
- **Standard Data Repository** – A place / location, where data products are stored, that is in compliance with the PDS Standards for archive volume production and directory structures.
- **Product** or **Data Product** – A data product is a labeled grouping of data resulting from a scientific observation. Examples of a data product include planetary images, spectrum tables, and time series tables. A data product is a component of a data set. Each data product is comprised of a label and the one or more data objects that the label describes.
- **Metadata** – Structured data that contains a definition or description of a product.
- **Product Label or Data Product Label** – A data product label is metadata expressed in Object Definition Language (ODL) which identifies, describes, and defines the structure of the data.
- **Inventory** – A database containing an organized collection of attributes used to select and locate products.
- **Configuration File** – A file that provides information to be used in setting up a controlled sequence of processes. These processes will act on data products submitted for ingestion. For example, a configuration file may include the following:
  - **Product Class Name –** The name of the product class. – see Product_Class.
  - **Product Location** – source and destination product storage locations
  - **Cataloging Attributes** – Keywords in product labels to be included in the product inventory. Supports the search function and index file creation.
  - **Data Element** - the definition of a piece of metadata described using the attributes specified in the PSDD.
  - **Task** – one or more client defined processes that will act on a data product.
  - **Rule** – a specification for controlling the tasks and workflow that includes a condition that must be met and an action to be performed if the condition is met.

- **Label_Template** – A structure that acts as a specification for validating a specific type of product – see Product_Class.
- **Product_Class** – a class of a PDS product defined by a specific set of attributes.
- **Status_log** – a log of the status of all of the steps performed against all of the products specified in the configuration file. The user will be able to search the log and ascertain:
    - Products that were processed successfully,
    - Products that failed processing,
    - Detected anomalies / condition(s) for a particular product or products

The following are definitions for the actors referenced in this document:

- **Client** – Any entity requesting access to the service.. The entity may be a user accessing a web-based interface or another application communicating directly with the service.
- **Service** – The server application responding to events based.
- **User** – A person who uses or accesses products on the System.

The following are definitions for the services referenced in this document:

- **Status** – This data flow provides the client with information regarding successful or unsuccessful completion of the event.
- **Ingest, to** (verb) – The process of registering, storing and cataloging a product in the repository.
- **Register, to** (verb) – To make or secure an “official” entry into the inventory/catalog.
- **Store, to** (verb) – To place a data product in the standard data repository for the purposes of preservation or use.
- **Catalog, to** (verb) – To save the metadata, such as that in a data product label, in the inventory for the purposes of classifying the metadata.
- **Archive, to** (verb) – To preserve data sets and ancillary files in a PDS compliant volume structure on long-term storage media, such as CD-ROM and DVD.
- **Track, to** (verb) – To report on the status of a resource -- the attributes representing the state of a resource at a given point in time, starting from the time when the resource is submitted to the PDS and continuing throughout its lifecycle.
- **Validate, to** (verb) – To determine whether a given resource or resource description is PDS compliant (or "valid").

3.0 CONCEPT OF OPERATIONS

3.1 Introduction

In the following sections we propose a concept of operations for the Ingest, Validate and Track (IV&T) functional components of the PDS data system. IV&T is intended to be a set of modules that can either be run independently or in a prescribed order. When an order of operations is specified, this is known as a workflow. Typically this workflow carries through all three components (i.e., Ingest, Validate and Track).

The process of ingestion involves the delivery to the system of a resource by a resource provider. The attributes of the resource are then parsed and validated for syntax and content. Once validated the resource is included in the inventory of resources in the system. Selected attributes for each resource are maintained as part of the inventory to aid in searching and selection by a user or for the generation of deep archive. As a resource is processed, a status may be maintained and reported so that its progress can be tracked.

3.2 The IV&T Workflow Concept

Workflow is a term used to describe the functions, procedural steps, components involved, required input and output information, and tools needed for each step in a business process. An example workflow for a PDS product ingestion scenario is illustrated in Figure 3.1.

The configuration of the workflow is prescribed by a configuration file. There is at least one configuration file for each product class. This information includes the components to be configured, rules for processing, product location, etc.

In the simplest terms, the user submits and registers a configuration file, then requests that a set of products be processed by the IV&T service. The configuration file defines the set of tasks to be performed against the products. The user is notified of the processing status of each task. If the products were successfully processed, the user is notified that the products were accepted. If any task fails to complete, further processing against the products is halted and the user is notified that the products were rejected. In the case of rejected products, the user has the option to correct the anomalies and resubmit the products, or to withdraw the products.

Ingest, Validate, and Track Workflow

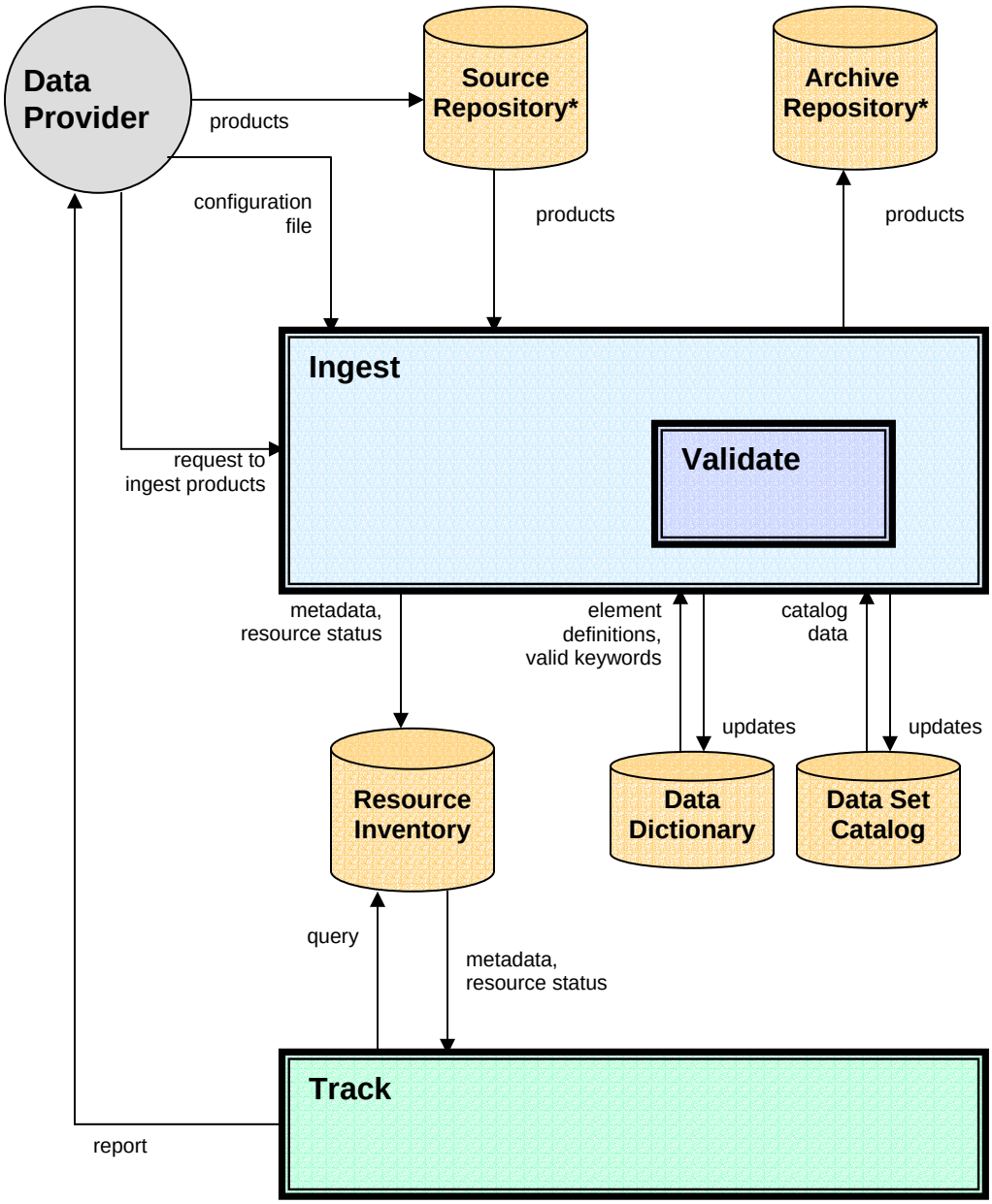


Figure 3.1 Ingest, Validate & Track Workflow

Figure 3.1 depicts the three major functional components of this design – Ingest, Validate, and Track (IV&T). Further details of these functions are depicted in Figures 3.2, 3.3 and 3.4. Figure 3.1 outlines a typical scenario where a request to ingest products is submitted to the Ingest function where a configuration file controls the workflow. Products from the source repository are ingested and validated, then stored in the archive repository. Product metadata are stored in a resource inventory which is used for the Track function. The Data Dictionary and Data Set Catalog provide information for the Ingest and Validate functions.

Figure 3.2 depicts a sample Ingest Function workflow. Again, this is intended to be a modular set of tasks which are separable. After each task is completed, the resource_status is updated which flows into the Track function. The ingest function includes all components necessary for registering, storing and cataloging data products.

Ingest Workflow

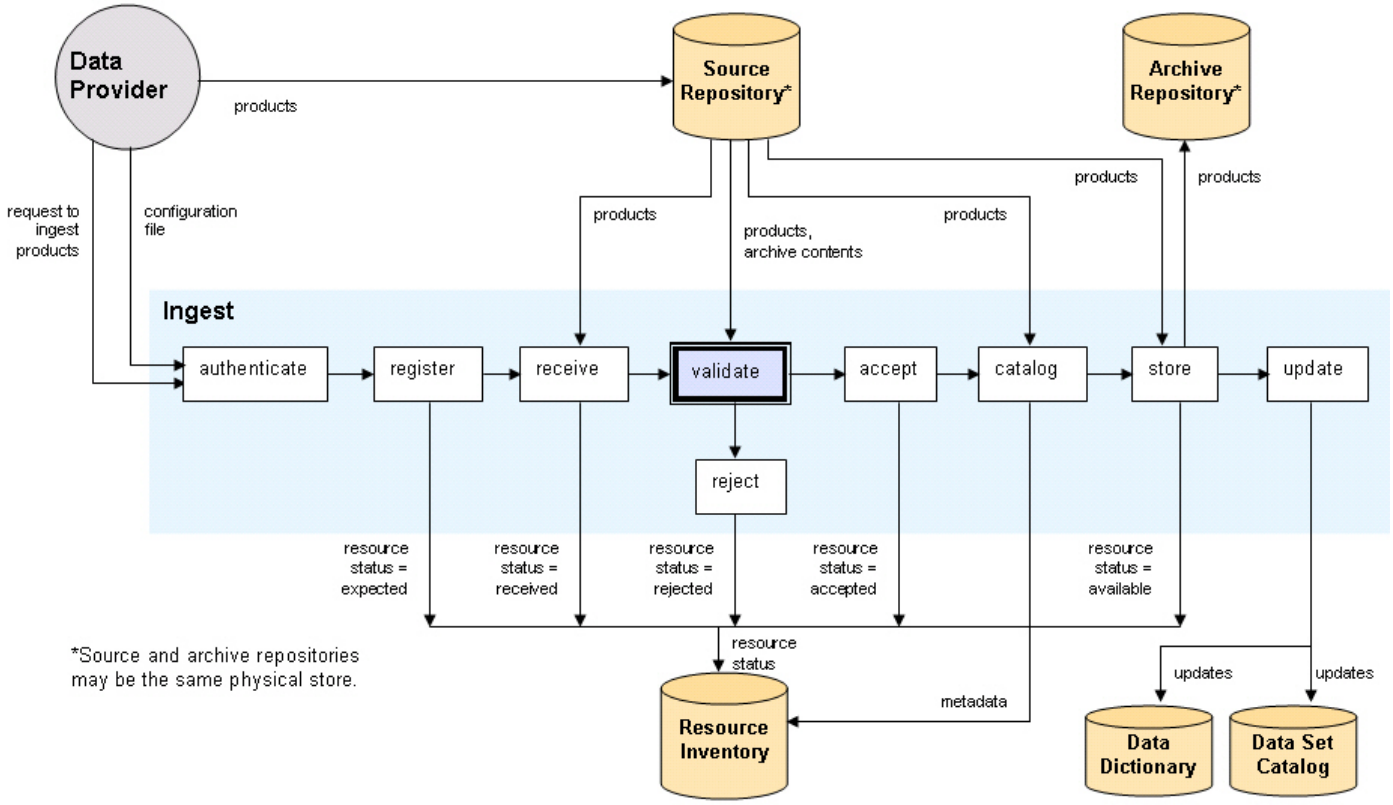


Figure 3.2 Ingest Function Workflow Example

Figure 3.3 depicts a sample Validate Function workflow. The validate function, a module configured for this scenario within the ingest function, determines whether a given product description is PDS compliant. The various steps of product and archive validation are shown here and may be run in sequence or as separate modules.

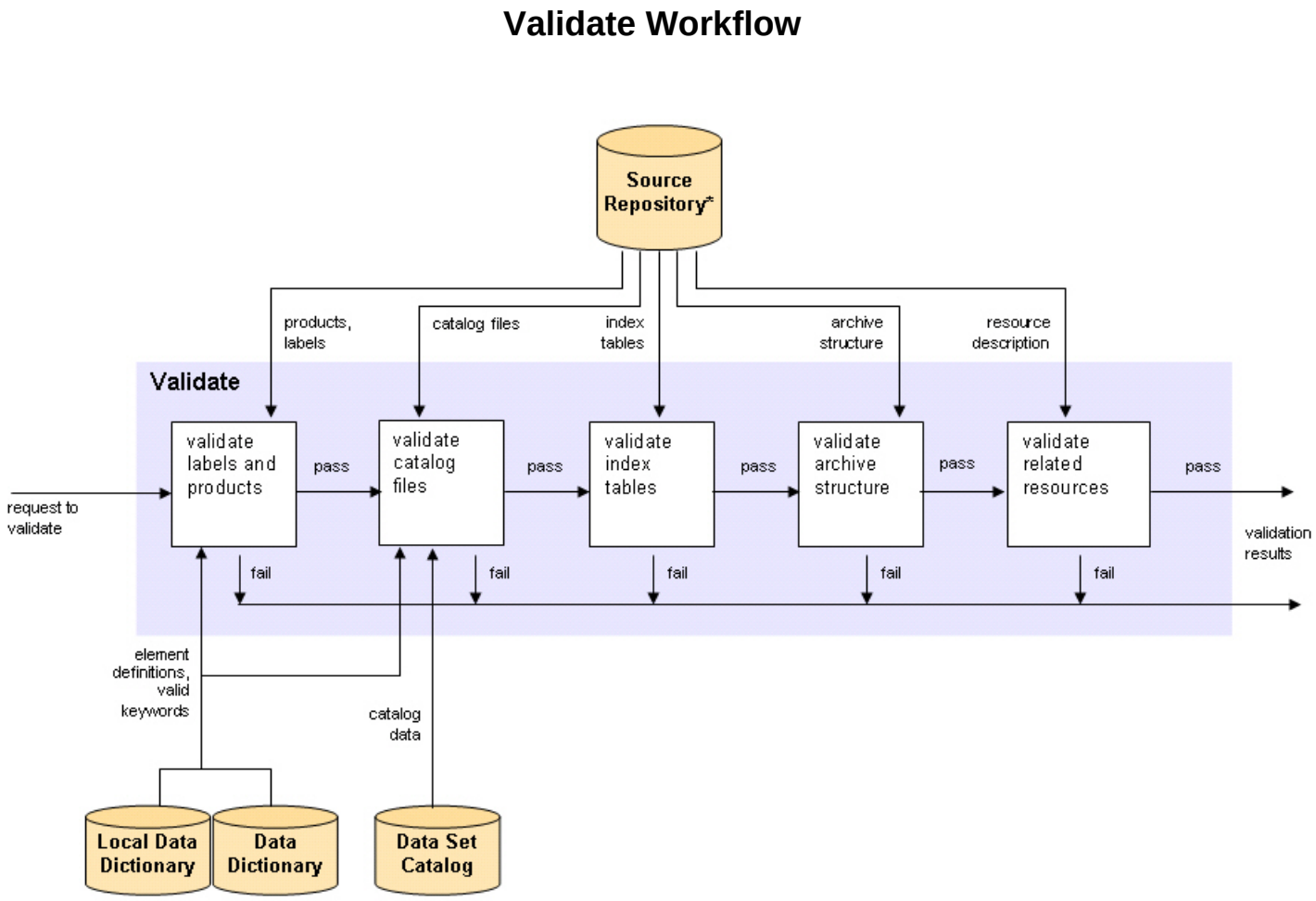


Figure 3.3 Validate Function Workflow Example

Figure 3.4 depicts a sample Track Function workflow. Attributes representing the state of a product are collected in the resource inventory and an interface is provided for user access. User notification is provided by a Notify component which triggers upon a change in the resource status.

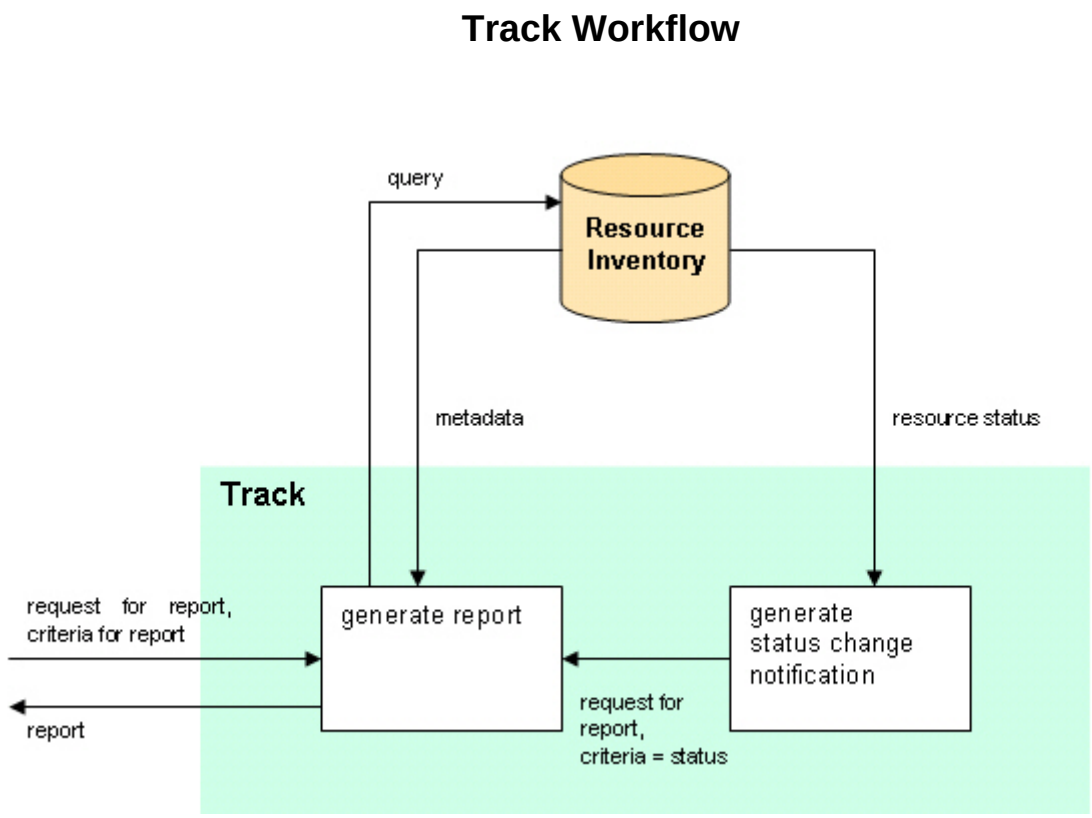


Figure 3.4 Track Function Workflow Example

3.3 A Detailed Description of the Concept of Operations

This section provides a more detailed description of how a user would deliver products to the PDS in the workflow context, either via the internet or by supplying physical media (e.g., data brick, DVD, etc), and how the IV&T service will process the users request – the end result being an ingested, validated, and online accessible set of science data. (See Figure 3.1.)

3.3.1 Service Access

1. The user, having a set of data to be archived with the PDS, has the option of either hosting the data on a server from which the data can be accessed by the IV&T service, or storing the data onto physical media which is delivered to the PDS.

2. The user creates a configuration file for the set of products being delivered. A configuration file is a definition that describes each product to be registered with the IV&T service (i.e., specific instructions on how to process the products.)

3. The user and/or PDS creates a Label_Template for each product_class (i.e., a specification for validating each specific type of product). This piece is included in or is pointed to by the configuration file.

4. The user can only initiate two types of processing - configuration file registration and product registration.

3.3.2 Configuration File Registration

A configuration file provides information to be used in setting up a controlled sequence of processes. These processes will act on data products being submitted for ingestion. The following describes how the configuration file is used.

1. The IV&T service, having received a request to initiate processing, will parse the contents of the configuration file into an instruction set for each identified product.

2. If anomalies are not detected in the above process, the IV&T service will either add the configuration file and product metadata to the catalog, or make updates to the catalog as appropriate:

3. The IV&T service will then notify the user confirming receipt of the configuration file and note any detected anomalies, either in the configuration file or product metadata specified in the configuration file.

Note: Configuration file registration is mandatory and must be successful before the IV&T service will be able to register products. In other words, tasks defined in the configuration file (e.g., product validation, product cataloging) can only be performed once a well formed configuration file has been successfully registered.

3.3.3 Product Registration

1. The IV&T service, having received a request to initiate processing, will identify the product_class associated with each product and will ascertain that a configuration file and a Label_Template exist for the product_class. The absence of either will preclude any further processing of those specific products.

2. The IV&T service will either add the product metadata to the catalog, or make updates to the catalog as appropriate.

3. The IV&T service will process the products in accordance with the tasks and rules defined in the configuration file specific to the product_class (see Task / Rule processing below).

4. Anomalies will be annotated in a status_log which the user will receive at the completion of all processing.

3.3.4 Task / Rule Processing

1. The IV&T service will evaluate and execute all tasks and all rules associated with each task as defined in the configuration file specific to the product_class of the product to be registered. A task will not execute until all associated rules are satisfied or a defined wait-time has expired. There are three types of rules: mandatory, optional and retry. Both rules and tasks may have individual wait times associated with them.

A fictitious example for processing tasks / rules:

a) If the configuration file defines a Product Validation task, the IV&T service will validate the syntactic content of each product_label against the Label_Template (i.e., ensure that the structure and metadata in the product_label match the structure and metadata specified in the Label_Template). Anomalies will be annotated in a status-log.

b) If the configuration file defines a Product Catalog task, the IV&T service will catalog the product metadata with the next available version number for the product_class. Anomalies will be annotated in a status-log.

c) If the configuration file defines a Product Store task, the IV&T service will store the data products into the standard data repository. Anomalies will be annotated in a status-log.

3.3.5 User Notification

1. At the completion of all processing of all products, the IV&T service will notify the user identifying any anomalies detected in processing the products.

2. The user will be able to view any notification e-mail via a standard email client as the email will be formatted as ASCII text without markup.

3. The body of the notification e-mail will indicate the overall status (pass/fail) of the processing request. Optionally, the user may request and receive as part of the notification a product-by-product report listing in chronological order each step used in processing each product, a “pass/fail” status, and a brief message indicative of any processing anomalies.

4. The user can check the status of a processing request by contacting and requesting a report from the Track service.

APPENDIX A ACRONYMS

The table below lists acronyms and abbreviations used in this document.

ACRONYM	DEFINITION
ASCII	American Standard Code for Information Interchange
CD-ROM	Compact Disc—Read-only Memory
DVD	Digital Video Disc
IV&T	Ingestion, Validation, and Task
ODL	Object Definition Language
PDS	Planetary Data System
PSDD	Planetary Science Data Dictionary

Table A-1. Acronyms and abbreviations

APPENDIX B DEFINITION OF PDS TERMS

The table below lists terms, acronyms and abbreviations used in this document.

Archive	An archive consists of one or more data sets along with all the documentation and ancillary information needed to understand and use the data. An archive is a logical construct independent of the medium on which it is stored.
Catalog Data	Descriptive information about a data set (e.g., mission description, spacecraft information, instrument information) expressed in ODL which is suitable for loading into a catalog
Data Set	The accumulation of data products, supplemental data, software, and documentation, that will completely document and support the use of those data products. A data set can be part of a data set collection.
Object Description Language (ODL)	A specification language that is used to encode data labels for the PDS and other NASA data systems.
Product label	A label expressed in ODL which identifies, describes, and defines the structure of the data.
Resource	Anything that can be described by a resource description. Resources in the PDS include data, software, documents, web services, spacecraft, and personnel.
Resource description	The set of attributes describing a resource.
Standard data repository	A place / location, where products are stored, that is compliant with PDS Standards for archive volume production and directory structures.

Table B-1. Definition of Terms