System Design:
Data Ingestion Update

PDS System Design Review II
Greenbelt, Maryland
June 21-22, 2011

Sean Hardman

Topics

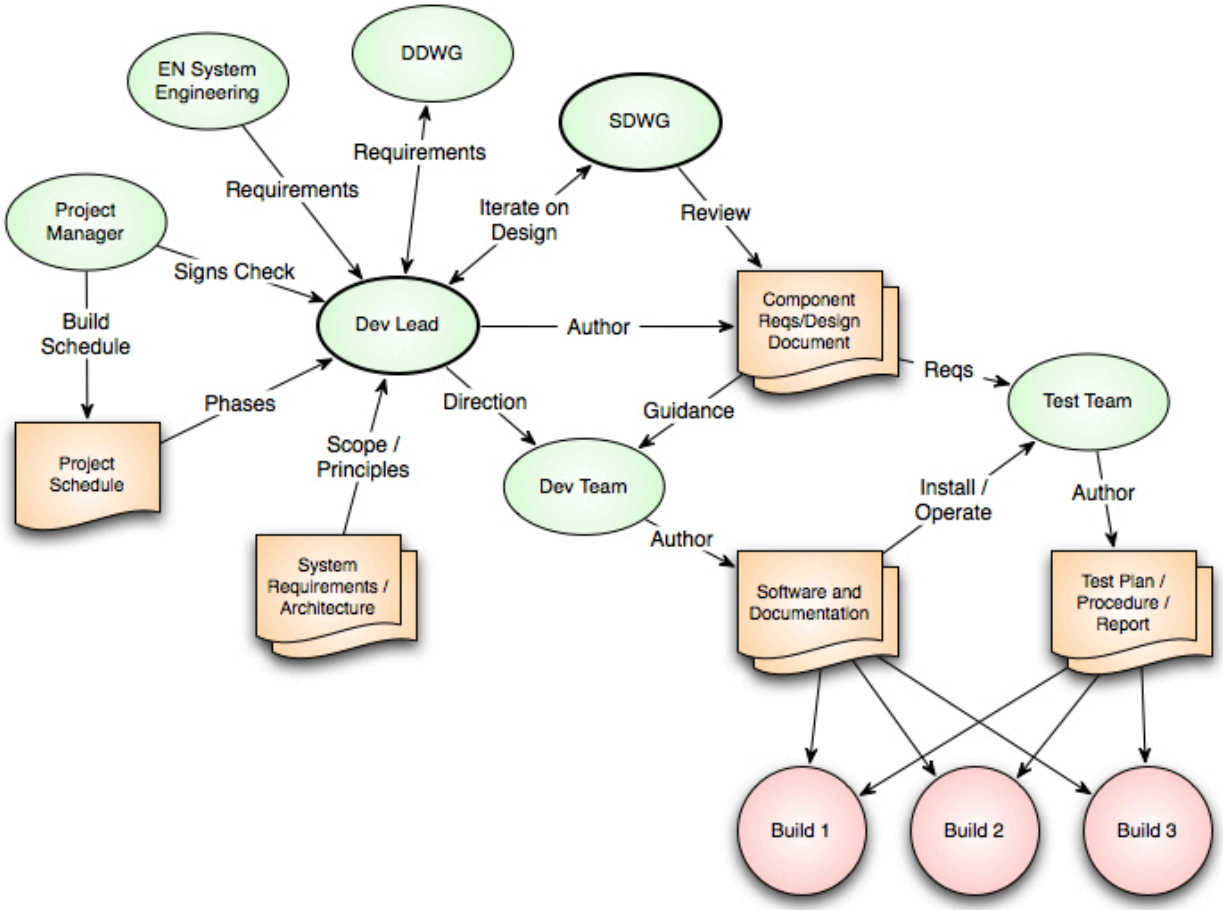
- Overview and Design Process
- Ingestion Related Components and Flow
- Registry Service
- Harvested Metadata
- Deployment and Plans
- Wrap Up

Overview

- Although covered in the first system review, this presentation provides an update on Data Ingestion.
  - Includes progress to date.
  - Provides more detail than was available the first time around.

- Data ingestion involves preparation, receipt and registration of PDS products.
  - This includes PDS3 and PDS4 data.

Design Process



Design Process cont.

- Each component has a corresponding requirements and design specification.
  - Level 4 and 5 requirements traced back to PDS Level 1, 2 and 3 requirements.
  - Each specification undergoes multiple drafts with comments incorporated.
- Additionally there is a general system software requirements document for requirements that pertain to all or most of the components.
- Requirements traceability and mapping of requirements to the builds is captured in another document.

Design Status

- Controlling documents completed and reviewed:
  - System Architecture Specification
  - General System Software Requirements
- Documents completed and reviewed:
  - Registry, Harvest and Security Requirements and Design
- Documents in process:
  - Preparation (Tools) Requirements and Design
- Latest versions posted to Engineering Node site
  - http://pds-engineering.jpl.nasa.gov/index.cfm?pid=145&cid=134

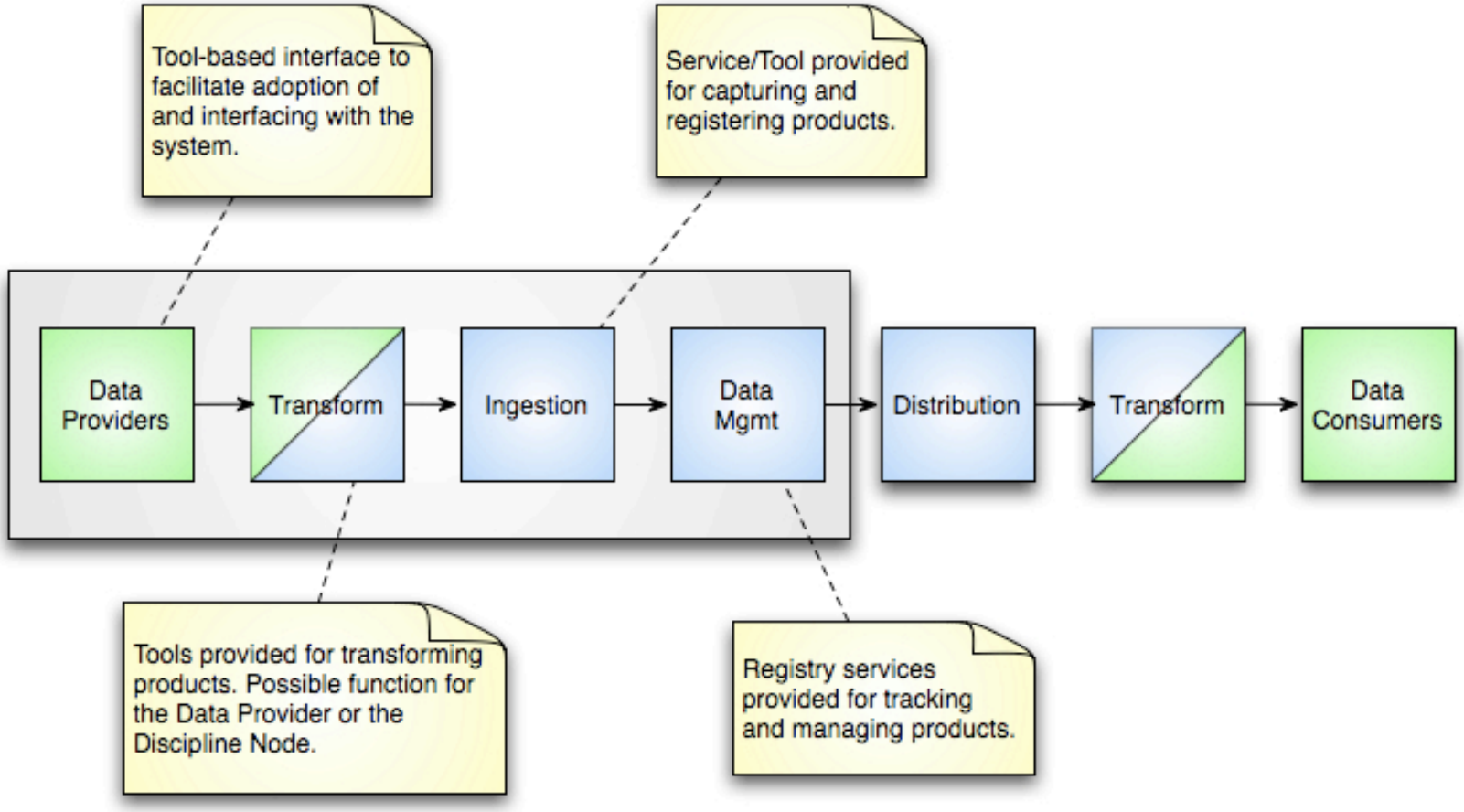
Key Level 3 Requirements

- **1.5.1** PDS will provide tools to assist data producers in generating PDS compliant products
- **1.5.2** PDS will provide tools to assist data producers in validating products against PDS standards
- **2.2.2** PDS will track the status of data deliveries from data providers through the PDS to the deep archive
- **2.6.2** PDS will design and implement a catalog system for managing information about the holdings of the PDS
- **2.6.3** PDS will integrate the catalog with the system for tracking data throughout the PDS
- **2.8.2** PDS will maintain a distributed catalog system which describes the holdings of the archive

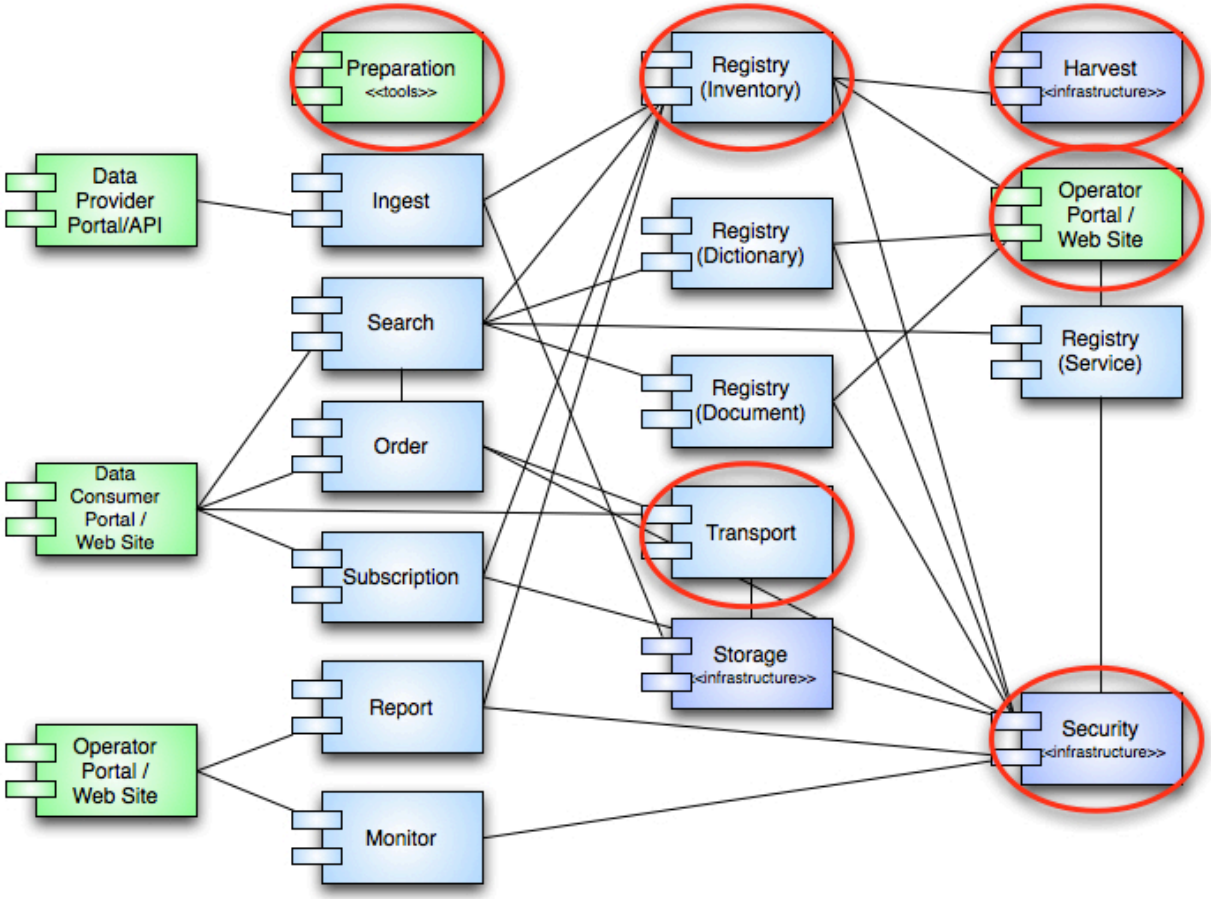
Topics

- Overview and Design Process
- Ingestion Related Components and Flow
- Registry Service
- Harvested Metadata
- Deployment and Plans
- Wrap Up

Ingestion
(Capture and Registration of Products into the System)



Ingestion Related Components



Ingestion Related Components
Preparation Tools and Transport Service

- Preparation Tools
  - Suite of tools for preparing data for ingestion into PDS focusing on design, generation, transformation and validation.
  - Allows for existing Node processes and procedures to be utilized for ingestion of data products.
  - Minimizes up-front interface changes for Data Providers.
- Transport Service
  - Represents continued support for FTP (push/pull) and Data Brick delivery mechanisms.
  - Currently looking into other mechanisms:
    - bbFTP, FDT (Fast Data Transfer) and bbFTP

Ingestion Related Components
Harvest Tool

- Crawler-based tool for capturing and registering product metadata.
- Allows for periodic or on-demand registration of products.
- Configurable to support registration of products residing in PDS3 and PDS4 archives.
- Designed to integrate well with existing Node operations.
- Provides the first line of metadata harvesting within the system in order to facilitate tracking of and access to products.

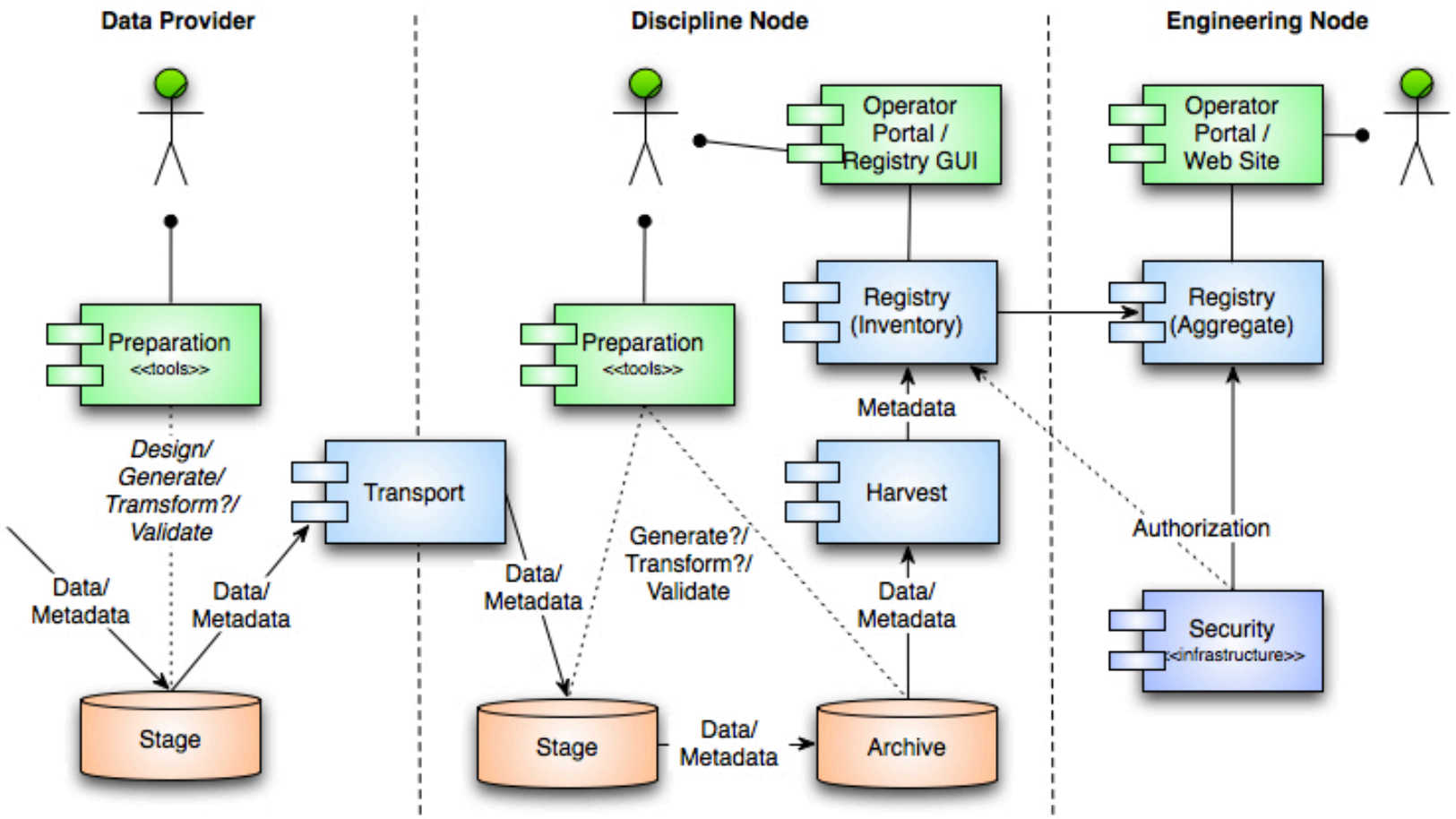
Ingestion Related Components
Registry Service

- Provides functionality for tracking, auditing, locating, and maintaining artifacts within the system.
  - Artifacts include data products, data dictionary element definitions, service descriptions and project documents.
- Provides a common implementation for registry service instances based on the Registry Reference Model effort which in turn is based on ebXML.

Ingestion Related Components
Operator Portal and Security Service

- Operator Portal
  - A general web-based interface for managing registry policy, content and end-to-end tracking.
  - The interface is deployable for local instances of the Registry service at the Nodes.
- Security Service
  - Provides the authentication and authorization functions for the system.
  - Satisfied with an Open Source product supporting the Lightweight Directory Access Protocol (LDAP).

Ingestion Flow



Ingestion Flow Details

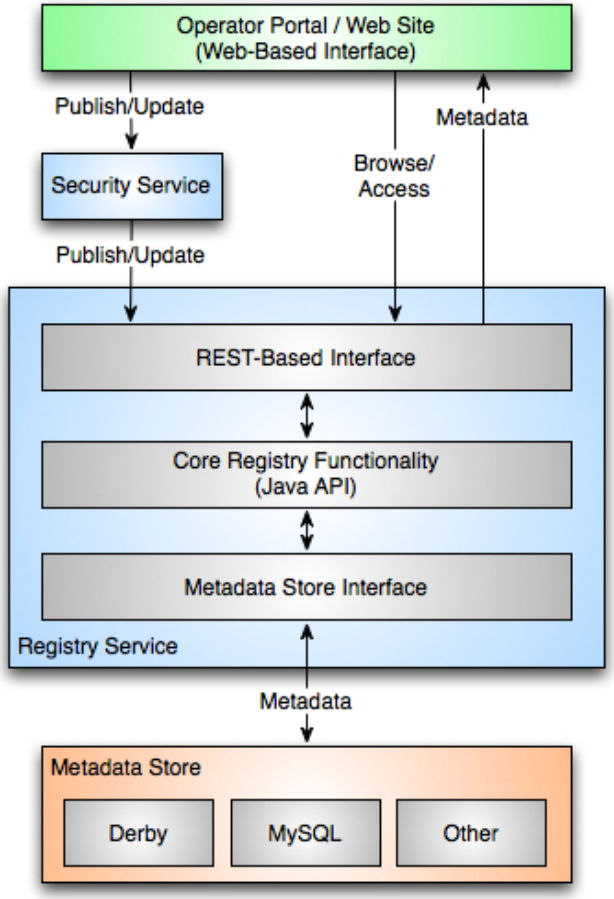
1. Data Provider receives data from the source (e.g., Project, Instrument Team, etc.).
2. Data Provider utilizes PDS provided tools to prepare the data for submission.
3. Data Provider submits transformed/labeled data to the Discipline Node via an agreed interface (e.g., FTP, Data Brick, etc.).
4. Discipline Node receives data/metadata from the Data Provider and stages it in local storage.
5. Discipline Node utilizes PDS tools or tools based on a common library to prepare the data for archive.
6. Discipline Node initiates harvesting of the archive, which registers product metadata with the Registry service. Metadata registrations are authorized by the Security service.
7. Metadata is replicated with the Registry service hosted at the EN.
8. Discipline Node manages housekeeping information and/or augments metadata for search enhancement via the Operator Portal.

Topics

- Overview and Design Process
- Ingestion Related Components and Flow
- Registry Service
- Harvested Metadata
- Deployment and Plans
- Wrap Up

Registry Architecture

- REST-based API over HTTP for registration and retrieval of metadata.
- Internals developed in Java with an API for manipulating registry objects.
- Metadata store interface allows for multiple database solutions.



Registry REST-Based API

- This interface delegates all functions involving a product:
  - http://pds.nasa.gov/services/registry/extrinsics/
    - GET: Retrieves a paged list of products from the registry.
    - POST: Publishes a product to the registry.
- This interface acts on a specific product (lid stands for logical identifier):
  - http://pds.nasa.gov/services/registry/extrinsics/logicals/{lid}/
    - GET: Retrieves the product from the registry.
    - POST: Updates the product in the registry.
    - DELETE: Removes the product from the registry.

Registry Data Model
Key Classes

- Association
  - Specifies a relationship between two registered objects.
- AuditableEvent
  - Records the actions taken against a registered object.
- Classification
  - Facilitates incorporation of taxonomies.
- ExtrinsicObject
  - PDS products are derived from this class.
- Service
  - Captures service descriptions.
- Slot
  - Captures additional attributes describing a registered object.

Registry Configuration

- The data model identifies the PDS product types.
  - Identifies the common metadata elements (slots) for each of the product types.
  - Identifies the associations for each of the product types.
- The data model also captures existing PDS taxonomies (classifications).
- This information is exported in a form to facilitate Registry Service configuration.
  - Will also be utilized by the Harvest Tool and GUI.
- Allows the data model to exert some semblance of control over the contents of the registries.

Topics

- Overview and Design Process
- Ingestion Related Components and Flow
- Registry Service
- Harvested Metadata
- Deployment and Plans
- Wrap Up

Harvested Metadata
Identification Area

- The logical id and version id become the unique identifier for the product.
- Product class is used to classify the object type.
- Title becomes the display name.

```
<Identification_Area_Product>
    <logical_identifier>urn:nasa:
       pds:data_set.A12A-L-SWS-3-
       SOLAR-WIND-28S-RES-V1.0
    </logical_identifier>
    <version_id>v1.0
    </version_id>
    <product_class>
       Product_Data_Set_PDS3
    </product_class>
    <title>APOLLO 12 ALSEP/SWS
       SOLAR WIND 28-SEC
       RESOLUTION TABLES V1.0
    </title>
    …
</Identification_Area_Product>
```

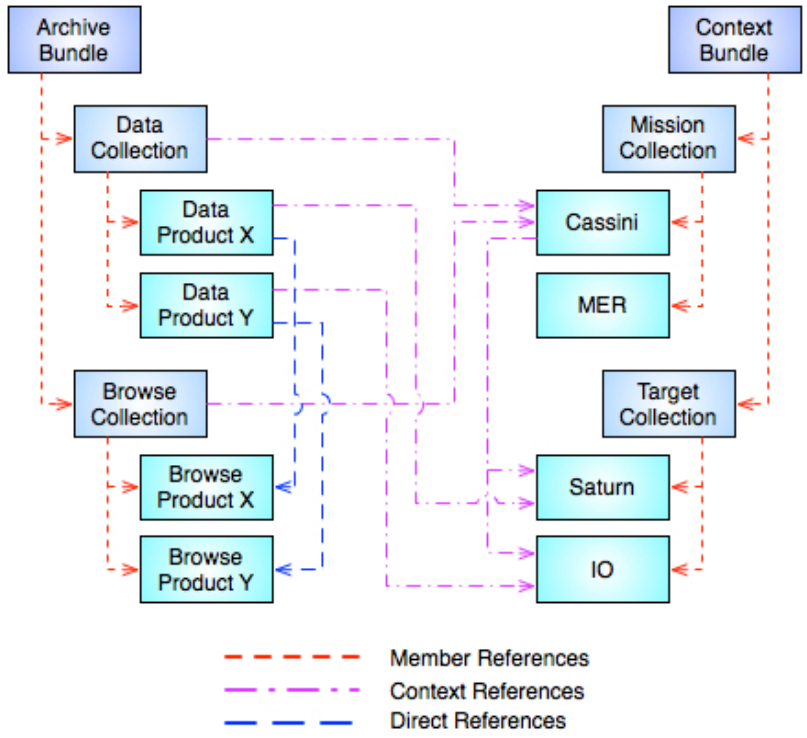
Harvested Metadata
Cross Reference Area

- A reference entry becomes an association in the registry.
- Relates two registered objects with an association type.

```
<Cross_Reference_Area_Context>
<Context_Reference_Entry>
   <lidvid_reference>
         urn:nasa:pds:
         investigation.APOLLO_12::1.0
   </lidvid_reference>
   <reference_association_type>
         has_investigation
   </reference_association_type>
</Context_Reference_Entry>
…
</Cross_Reference_Area_Context>
```

Harvested Metadata
Reference Types

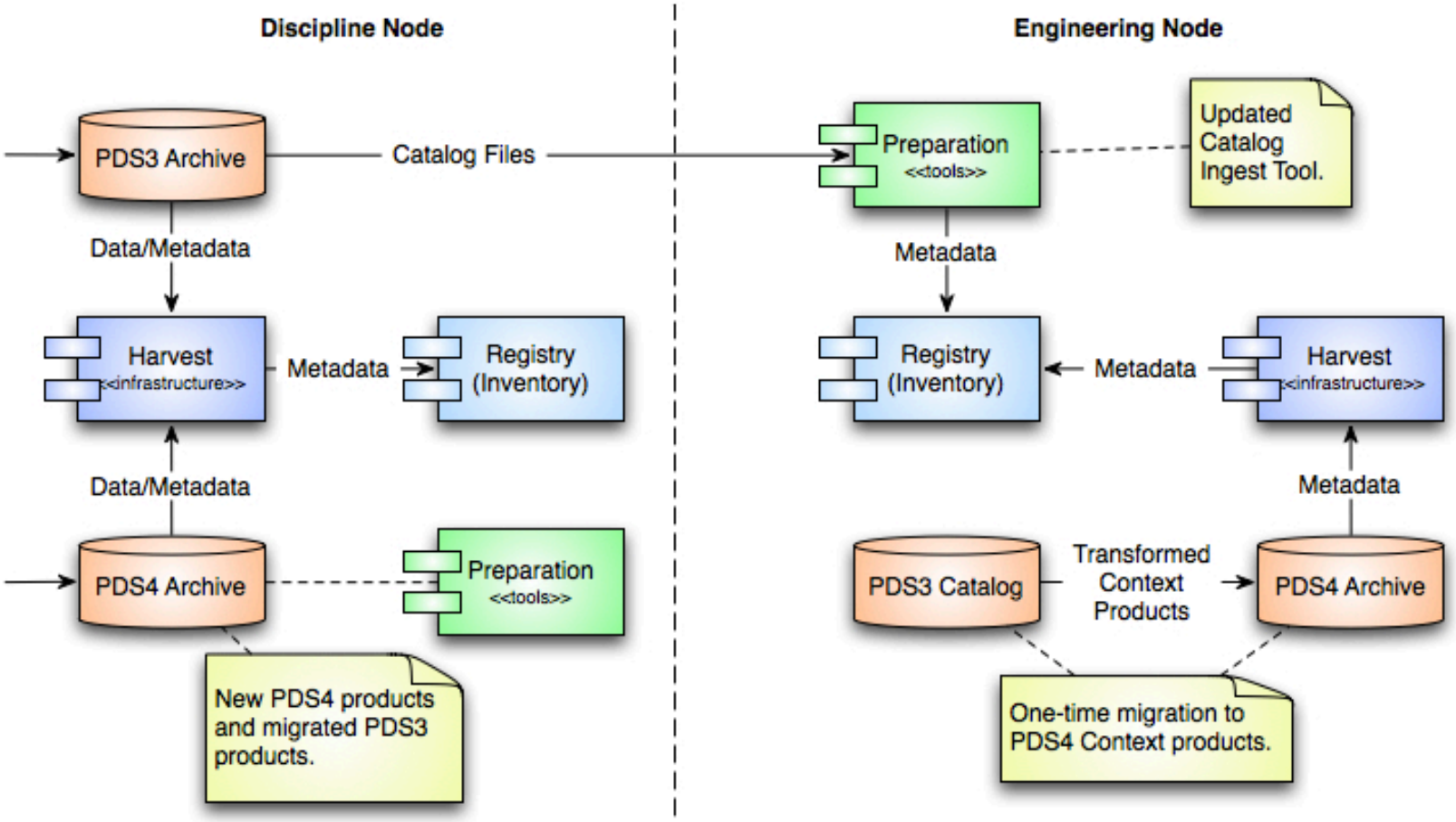
- Member
  - Represents the logical/physical tree of the archive.
- Context
  - Informational to facilitate search.
  - Realized at collection or data product level.
- Direct
  - Represents ancillary information.



Harvested Metadata
Other Areas

- The Subject area contains keyword metadata and facilitates text-based search.
- The Data area can be harvested on demand.
  - Depends on the search requirements for the local registry.
- Harvest is configured to extract specific elements from the product label and place them into slots in the registry.

PDS3 Support



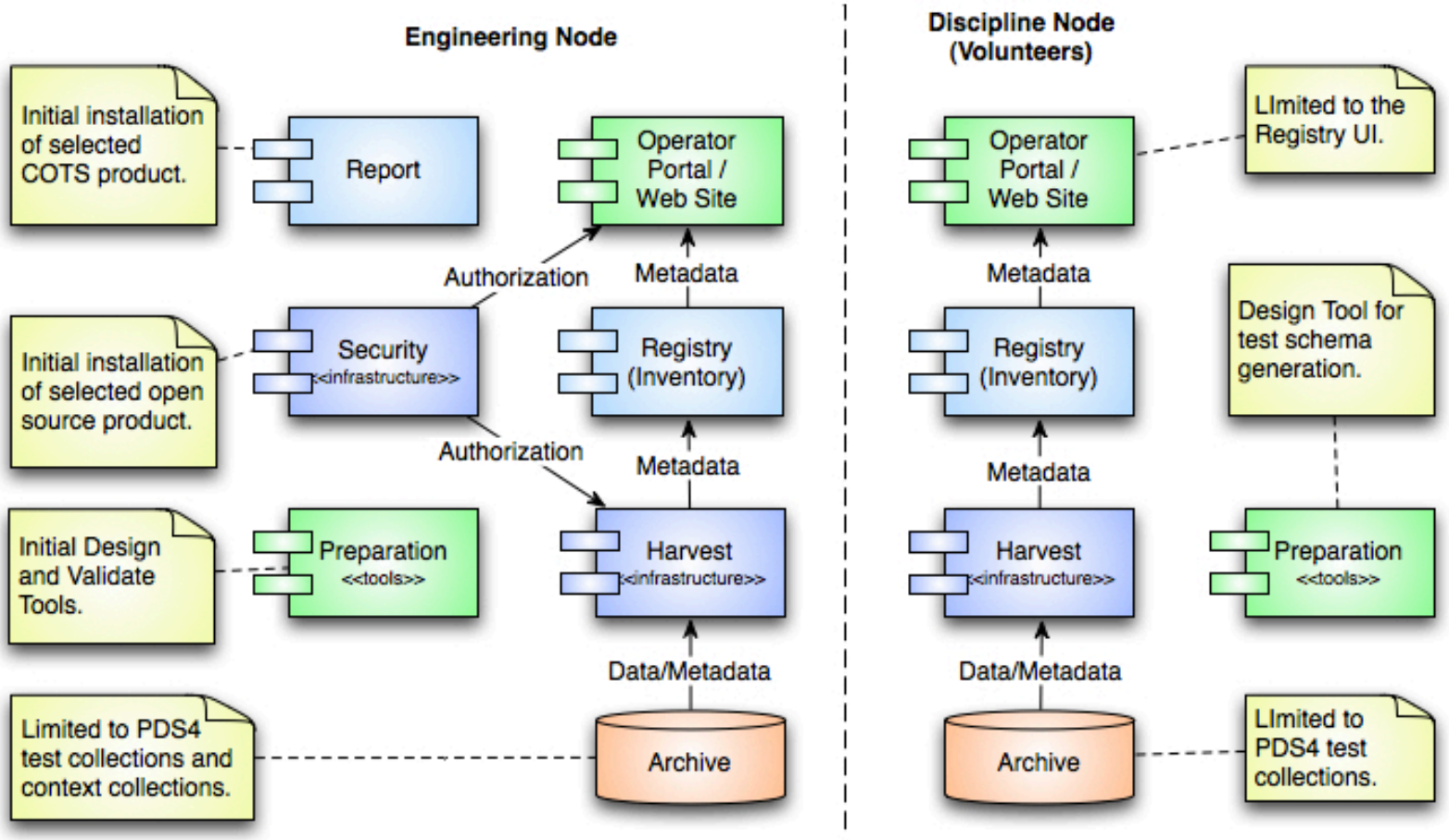
PDS3 Support
Additional Details

- The Harvest Tool supports both PDS4 and PDS3 registration.
  - PDS3 support consists of converting PDS3 labels into PDS4 proxy labels.
  - Registered for tracking and reporting purposes.
  - Will be replaced when the corresponding PDS3 data set is migrated to PDS4.
- The current Catalog Ingest Tool is updated to convert the catalog files to context products and register them with the registry.

Topics

- Overview and Design Process
- Ingestion Related Components and Flow
- Registry Service
- Harvested Metadata
- Deployment and Plans
- Wrap Up

Build 1 Deployment



Build 1 Deployment
Additional Details

- Intended as a prototype build for the core components.
- Deployment of the software at the Nodes was voluntary.
  - The Atmospheres Node is the only taker so far.
  - It was a good experience working out system requirements.
- The exception to voluntary Node deployment was the use of the off-the-shelf Design Tools.
- No integration was required with this build.

Preparation for Build 2

- Add support for querying slot content and for registry aggregation for the Registry Service.
- Add support for registering file objects, test PDS3 harvesting.
- Add support for bundle validation to the Validate Tool.
- Complete development of the Catalog Ingest Tool.
- Design and develop the Tracking Application.

Topics

- Overview and Design Process
- Ingestion Related Components and Flow
- Registry Service
- Harvested Metadata
- Deployment and Plans
- Wrap Up

Wrap Up

- The ingestion process and flow closely mimic existing processes at the Nodes.
- This allows us to minimize impact on current Node operations.
- The Harvest and Registry components provide the core functionality for the rest of the system to be built upon.
- Facilitates tracking, reporting and ultimately product-level search.

Questions/Comments