Day 3, 01: Objectives and Introduction to Services

Objectives and Introduction to Services, Dan

Series as way to ties PDS together

- not just as nodes, but internationally as well (ESA PSA)
- What can we do to bring ourselves together as a federation?

Day 3, 02: Data ingestion and registration

Map

Data ingestion and registration-Sean

Ingestion: harvesting metadata and registering it with PDS products

- this talk will focus on PDS4
- For PDS3 products, we don’t have consistent product IDs

Harvest Tool and it’s configuration

- crawler-based tool, usually run on-demand
- low-impact: we don’t want to impact how nodes currently work, behind the scenes
  - phased integration of the software into various node environments, only a few
- harvest: facilitate access and tracking of products

Harvested Metadata

- title is display name: please try to use unique titles so the display is differentiated
- Question: is this title use explained somewhere?
  - No, I don’t think it’s been captured anywhere.
  - This should maybe be in the DPH? Citation description
  - Go back to official definitions in the IM for descriptions of fields
  - How metadata is being used is not well documented
  - Various levels of integration at different nodes: users can customize titles
- This is where they are pulling metadata for most search terms
- Harvest has probably been re-designed 3 times because of changes to the IM since dev started
- Comment: Query models produce constraints on observational data
  - Query model will be discussed later
- references: take internal reference and change it into single slot value in the registry
  - there are many references, they are flattened out for registry service

- Used to have heavy reliance on association object in registry service, now use references more
- most of the harvest ignores node-specific areas, but can be configured to extract other elements
    - tailored search config for Atmospheres Node

Label Element to Registry Slot Mapping

- Used spreadsheet to track mapping in the past, want to move towards capturing these in the IM
    - generate directly from the IM will be must easier, new versions will be updated quickly and easily
- label —> registry, registry —> search service
- have harvested attributed and registered them in the service

Harvest Configuration

- comes pre-configured with a number of product types, can easily be overwritten
- not included in global policy
    - running locally will likely need configuration

Sample Example Bundle

- list collection products up front
- Question: if I have a new collection, I will have to edit every time?
    - Yes. We can get around that, will discuss later
- Root directory where bundle is
- Access URL feature: used in the registry at EN a lot
- Checksums: if you provided one in the label, it will verify the checksum
- Question: Access url, does each product in registry have an associated base URL?
    - I'll get to that in a slide or two, not for each product, for for file
- Candidate: names the specific products
- Question: This is for all of your bundles, or bundle specific?
    - They are bundle specific
- Question: Is there a way to test harvesting to see what did or did not go right?
    - Yes, I use local registry or multiple installations. Content in registry won't be displayed until approved
    - What about with search?
        - once you go through harvest and registry, you should be mostly okay. Missing Validation piece that we haven't made available yet
- Several people have version numbers in file names for accumulating collections
    - don't want to change config file every time there is an update
    - Will out in JIRA ticket for this, not next build but soon

Registry Service

- Based on CCDSD Registry and repository reference model
    - need something that performs better, will change in the future

- using config file from IM, slight conversion from registry service

Ingestion Process

- harvest registers package with registry
- Extracts metadata from product label, validates checksum if provided product registered as specified type
- Question: What if checksum fails?
    - registry for that product fails, but it will continue on. You can bypass that, though.
- Product File Repository associates each files
- Question: If you want to register one product that was rejected?
    - will display warnings about all of the duplicates
- Question: What happens if you have moved files on disc? How will registry pick this up?
    - Tool that hasn't been written yet. Let's write up a ticket for it.
- Question: If there is a problem with the checksums, is it a successful load, is there just an error message?
    - will just show which files failed. Looking into developing helper tool for checking for errors before running through harvest
- Question: If checksums are not supplied, can there still be a valid registration?
    - Yes, the IM does not require it to be in the label. Model requires checksum manifest in the package, can be done in validate tool. Don't want to run checksum again.
    - Should we set it so that harvest runs checksum every time?
- Ron: In NSSDC work, it was unclear whether some bundles had checksums run
    - Validate doesn't read the manifest currently, but it could. Need to verify.
    - Checksums will likely be run long before validate, but sometimes checksum issues still occur later on.

Post-harvest:

- package and associated products have to be approved to be seen in search

Registered Content

- Never made metadata query interface that useful, more focused on metadata extraction
    - registry is just one area of search index, other products will go into index (i.e. product updates, tracking service)
    - All interfaces now query the search service
- Question: How to find object/field that isn't in the registry?
    - Using concept of index table with additional metadata fields. Will ultimately be available, not yet released
    - index table is used a lot at imaging node
- Comment: if you have a product update, do you validate content of table? Read from table, then dynamically replace values in label
    - Future DDWG conversation, don't want to generate new labels on the fly for

product update

Service status/updates

- Both PDS3 and PDS4 registries are online (see slide 19 for links)
- Also registry for IPDA
    - has all of the PSA data points harvested from their site
    - No set harvesting plan, but happens about every month or two
    - Were also working with Indian Space Agency, but not yet
        - work in more international datasets
- registry service can be queried for metadata harvesting
    - stopped indexing all PDS4 context objects; only for those with data
- Mitch: We need to have context objects for items that have data on the way
    - only effects search
- Questions: Tracking releases?
    - tracking service will track: releases, archive status- these are not in the IM
- **Ponder**: Two labels (Product A and Product B): label B wrongly points to product A, how do we catch/resolve this?
    - possible in the file supplemental file
    - Anne wants multiple products pointing to a single file
- Question: How many PDS4 bundles are registered with you?
    - 10 bundles and 49 collections. Spice bundles and LADEE
    - How come there aren't more bundles registered yet? There should be more.
        - New PDS4 data coming. The bundles are not ready to be put into search. They are registered, purposely not pulled. Registered at node, but not considered registered for search?
- **Question**: When do we start to show these products? Do we just pull automatically? Do we need review process?
    - Local node registry vs. Public registry- what products are made public?
    - Why would I register things that I don't made public?
    - Richard Chen and Emily Law will work to put process together
    - **Process to migrate data**
    - Need something to document process of how to get items into search
- **Sean: Review total bundles registered**

Day 3, 03: Search, Access, Transformation and Distribution

Map

Overview

- Search core: indexing
- Search service: Apache solr with PDS configuration
- transport component facilitate access to data products

Search Service

- returns metadata about products, will have links to products inside
- Manipulation of metadata to get it into the index, allows flexibility to get info into idea that isn't in registry
    - also pulls from tools registry

Search Architecture

- Has several interfaces built on top of data sources :
    - PDS API
    - PDAP API
    - Other APIs

Index generation

- Index generation Index makes use of associations in registry to find related products
- There is one common configuration, but intent is to tailor config for discipline nodes
    - hopefully using similar field names
    - Cross-node search is being tested, will work on more in coming year
        - Use cases?
            - There is one. Not a lot of situations where cross-node searching makes sense.
- Question: In update table, each field would have something defined in a dictionary, would the field name identify the namespace? Common namespace across nodes
    - Yes, it should. Common tailored config across nodes would pick up on these fields/dictionaries

Search Core Configuration

- secondary registry references: PDS4 data products can have multiple registries at 2 nodes

Example: Product observational

- second aspect of mapping: what we find in registry slot mapped back to search term used

Question: Mission dictionary ingest, simple or complicated?

- Will have to do 3 things: re-harvest, update config file for index, search service configuration
- Comment: Query model into mission LDD —> JSON file with all the updates
    - data provide working with mission dictionary should have everything they need

Search service config

- Mostly Apache Solar pre-configured, we add specific fields they want to capture in their search (i.e., mission area)

Search Goals

- Improved integration across nodes and agencies
    - agencies have been integrating well
    - Nodes don't integrate as well
    - Try to use Atlas to map parameters
- Question: Are all of the nodes using REST-based interface, get into APIs?
    - No, but for the ones that aren't, we have services available. Didn't want to write search interface for the nodes, let them do it but use out services to serve search up
    - Will probably get upgrade this year. Possibly significant.Will be general, but can easily tailor facets for node-specific searches
- From search, Rita wants to direct users to PDS mission pages where they can drill down to mission specific info
- Question: Documentation for REST interface?
    - Yes, next presentation
- Jump page to data itself, or to page that describes data the same way that was listed from search?
- What would cross-node look like?
    - if each node had an index, we could just roll it up?
    - present at product level

Service Status

- PDS protocol and PDAP protocol for REST-based APIs
- Documentation for how to use at links?
    - No, in the next presentation

Service usage: search service has multiple registries, all rolled up into one
Access and Distribution

- Transport Service (OFSN) upgrades version of old PDS-D product server: allows you to view old return types, transforms on the fly

- Registry transport service: Will see list of identifiers, runs against search service
- Question: does it return products referenced by other products?
  - Not yet. How do we want to do that? Only bring back files that are referenced in product label
- Question: Size limitations?
  - Better have big temp area, can specify. You can also put your own limitation on it.
  - sub-folders of the zip? Used to be able to retain folder structure, have not done anything like that with this
- Question: what does product mean? Label, product file?
  - Both
- Question: What if you give it a collection
  - Doesn't grab all the items of a collection yet. We could do this.
- Question: Product that references associated products, will it grab it?
  - No, but you can specify associations of certain type
- How to link to bundles from other nodes?
  - Link that takes you to online listing of that bundle.
- Question: Referenced products: can handle with Opus.
  - is PDS3, less for PDS4
  - Checksum and delivery manifest?
    - Just checksum now.

Service status/usage:

- transport registry service is running, see links on slide 24
- is available on the backend, but not the with nice interface

Day 3, 04: APIs and Interfaces (REST protocols, PDS-specific etc)

Map

Registry API: registry protocol implemented as REST-Based interface instead of HHTP

- has own interface, you can get to in the registry, has documentation
  - pds.nasa.gov/services/registry-pds4/docs/

- Will return XML and JSON files
- can write python scripts to query package

Search API

- Data discovery
- data access
- service linking
  - want to work on this in coming year
  - pds.nasa.gov/services/search
- there is document about PDS and PDAP search protocol available through EN site

Next steps

- need to do work on PDAP protocol, low priority

Question: Tracking service: suppose you have data set, you post it, then they recalibrate. How does this work with tracking and registries, to keep both?

- New data set should have new version, will be registered once, but with different versions in the registry
- How do you supersede?
  - Will be tracked in tracking service. How do we roll it back in the registry? Not sure, developing.

Question: Registering pre-delivery, version 0 stead of 1?

- could be version 1, should represent version they're going to deliver to us. Will be tracked but not registered.

Comment: shouldn't matter where users enter into PDS, they should be able to access materials regardless. How does this work with tracking and registering?

- requests for data from instruments on same mission that have been distributed across nodes

Develop REST API plugin to Opus

Question: two different products that include the same file: data products?

- That is forbidden for data products, you're labeling multiple digital products with different labels. Might be okay with documents, though.
- Steve said yes to Anne. No check in the software for this.

Day 3, 05: Wrap-up Discussion

Map

- Process to put migrated data online: Local node registry vs. Public registry- what products are made public

- Registries with PDS4 data
  - NAIF
  - Geo
  - PPI
  - IMG
  - RINGS
  - ATMOS
- Search Service
- REST Adoption
- If you are bringing up services, let’s talk about common REST approach
  - if you have your own REST interface, we can do the mapping
- Bringing back to Mgmt Council meeting: important to have face-to-face meeting/discussions, regularly scheduled
- Sharing tools and development will be brought up with TWG
  - Github interface
  - Email amongst nodes? Can set up a mail list
  - Purpose of tool registry is to make more visible, but excludes projects in development