PDS Tech Session

13 February 2018 (Day 1 of 3)

Westin Pasadena (Plaza Room)
191 North Los Robles, Pasadena, CA
(https://pds-engineering.jpl.nasa.gov/content/2018_tech_session)

Attendees:

Rafael Alanis, Maria Banks (by phone), Carole Boyles, Mike Cayanan, Dan Crichton, Cristina de Cesare, Mike Drum, Mitch Gordon, Kevin Grimes, Ed Guinness, Sean Hardman, Lyle Huber, Steve Hughes, Chris Isbell (by phone), Joni Johnson, Todd King, Connor Kingston, Bill Knopf, Emily Law, Tania Lim (by phone), Maria Liukis, Joe Mafi, Tom Morgan, Lynn Neakrase, Anne Raugh, Boris Semenov, Dan Scholes, Dick Simpson, Susie Slavney (by phone), Tom Stein, Jesse Stone, and Kathryn Sweebe.

Welcome (Crichton):

Dan asked that attendees introduce themselves. He wants to make sure that action items are captured during the meeting.

Best Practices for Bundle Creation (Huber):

Anticipating IM v2.0, Lyle reviewed what has been done so far, what works, and what doesn't. Atmospheres thinks that bringing in v2.0 by March 2020 would be good but perhaps not realistic. So far the following numbers of bundles have been registered in the central Registry: ATM 16, CIS 3, GEO 2, NAIF 1, PPI 14, RMS 0, and SBN 8. This does not capture the volume of data, just the number of bundles recognized by EN. Guinness questioned the GEO number, and Huber said his own(ATM) count is different. Hardman said that locally registered bundles have not been automatically registered; DNs register bundles locally while they are still in development, then register them with EN later. Raugh asked whether there are formal procedures for registering bundles (and collections); Hardman said a procedure is available, but Richard Chen has it, and Sean is not sure what is in it. The procedure for registering bundles should be made more publicly available.

Huber then moved to content of bundles. PDS4 is more than new labels; LIDs interconnect products, and Context products have new significance. Hardman thinks the currently available Context collections have not been put together well (in many cases PDS3 CATALOG files were copied to <description> values in PDS4 Context products); they are not useful in many cases. Hardman added that the PDS3 migrated context products were meant only to be placeholders. However, the EN-curated bundle of current Context products will be sent to NSSDC soon. Guinness noted that the Standards Reference says the Context products should contain no information not available somewhere else in the bundle; so he views the Context products as not useful.

Simpson asked what should be done to improve Context products; Sean said updates should be sent to Richard Chen. There is a need to clean up almost every Context product in the system — for example, coordinates for optical telescopes that are given to 6 decimal places with no information about when and how the values were obtained. Raugh noted that SBN did not want to have the PDS3 CATALOG files dumped into PDS4; SBN would prefer to update Context products as the associated data sets are migrated forward, not be forced to update them outside a migration context. There was some agreement that lead nodes for missions should have primary responsibility for updating 'their' Context products. Neakrase said that he is preparing to ingest PDART data sets; he can't wait for lead nodes to get around to updating those Context products.

Lyle doesn't believe the Schema collections are very useful to users; Raugh said that mission dictionaries should be included, but 'dictionaries' is not a Schema collection issue.

There was discussion about when bundle versions change; if collections are referenced by LID only, there is no need to increment the bundle version when a collection is updated. A follow-up question is when bundles go to NSSDCA; there have been conflicting directions from NSSDCA, which need to be clarified.

Search: Searches can be local or global; they can scan bundles, collections, and/or products (though not all search levels have been fully implemented). There is a question of when to search using attributes with enumerated values; some lists of enumerated values are long. Huber argued for user-defined values rather than long lists; King countered that this leads to synonyms in the resulting, ever expanding list, which will never be equated. Raugh said that choices that cannot be organized into a taxonomy should be free form. Hughes said there is an SCR in JIRA that will reopen discussion on "instrument type"; Lim said PSA is considering another. Few in the room admitted to using the PDS4 Search function in its present form.

Migration: One question has been what to do with original (PDS3) labels; there was agreement that the PDS3 label provides history even if there is no value added in reading and using the data. ATM has included old labels in File_Area_Observational_Supplemental. If the digital object does not have to be changed for PDS4 compliance, then the data, PDS3 label, and PDS4 label can be stored together. GEO and CIS have been saving PDS3 labels as Header objects; Slavney would prefer that attached labels be separated into distinct files. Guinness wondered whether Stream_Text would be better than Header. Huber suggested that DDWG consider whether there should be a single solution for consistency across PDS.

Fixes and Point Builds: As context products are updated and enumerated value lists refined, what is the best way to incorporate these into the Information Model? Waiting 6 months for a new release is often inconvenient. Perhaps, under v2.0, point builds to accommodate such changes will be more common and full builds less common.

Best Practices for Local Data Dictionaries (Raugh):

Discipline dictionaries were expected to be developed quickly, then be subject to configuration control with occasional updates. Mission dictionaries were not expected to adhere rigorously to common constraints. Surprises included that LDDs have come more often from ROSES archivists than missions; discipline dictionaries have lagged; the relationships between LDDs and the common dictionary are not easy to explain; and LDDTool, originally expected to be a stop-gap internal tool, has become a cornerstone of LDD development. The relationship between LDD and IM versions has become a problem. CCB-156 was filed to prompt investigation of inconsistencies. CCB-203, CCB-204, and CCB-205 have identified specific issues. These should be resolved before PDS goes to IM v2.0; also driving toward improved consistency are the Roadmap, ROSES, migration, and missions.

Versioning and Provenance: Developers want to develop LDDs independent of IM development, track the IM development history, ensure that all dictionaries referenced in the same label are part of the same IM version, and locate dictionaries belonging to a specific IM version. Tracking changes is both best and standard practice. Modification_History is available to do exactly this; it needs to be used.

Build and Release: Discipline dictionaries should be released simultaneously with new versions of the core IM, but they should not be required to have cosmetic updates. Huber noted that mission dictionaries are developed on schedules which are largely independent of PDS4 builds, There should be a way to notify PDS and the public when new LDDs are being released. Everything after submission of the new Ingest_LDD file should be automatic. There should be a robust testing program for discipline dictionaries; details need to be discussed and implemented.

Development and Configuration Control: All users need a problem reporting system for both the core and discipline dictionaries. Discipline dictionaries should be included in impact assessments for core change proposals. Changes should be trackable. Validation for discipline dictionaries should be as rigorous as for common. The concept of a small team of experts developing a discipline dictionary is the exception (*e.g.*, the geometry dictionary); most discipline dictionaries have single stewards and single developers. There is no standard for what constitutes configuration control for LDDs. To develop an LDD requires proficiency in xPath, which is seriously lacking in PDS.

Gordon asked that attendees study Raugh's slides, which are posted on-line, in preparation for tomorrow's discussion on LDD 'solutions'.

Information Model v2.0 (Hughes):

PDS became operational in 1990; the archive now includes data from over 40 years of solar system exploration. PDS4 v1.0 was released in 2013. The early design focused on the 85% of digital formats that could be described simply; there has been little study of the remaining 15%.

There are 4 fundamental data structures (FDSs): array, table, encoded byte stream, and parsable byte stream.

PDS v2.0 should represent a relatively stable version of the IM; this suggests that the design should be sufficient for the remaining 15% of data. Hughes proposed that the other digital objects be composites of the four fundamental data structures. There are three cases: data that can be normalized with respect to the 4 FDSs (*e.g.*, ISIS3), structures that can be normalized using "sound practical sense" (*e.g.*, 3-color images with interleaved lines), and structures that can only be normalized through "exceptional efforts". Required will be identification of structures in PDS3 holdings that cannot be described using the FDSs and estimation of their frequency of occurrence; this effort should be coordinated with the PDS3 to PDS4 migration work. King wondered whether we already know the answer; JPEG2000 images are not PDS4-compliant, and Huber thought some old radio science data may also not be compliant (though Simpson was not convinced).

Discussion then turned to archiving MAVEN telemetry data, possibly as Product_Native. Simpson felt that PDS should not spend its resources on telemetry archiving, since a lot of effort would be required to ensure that enough information was included in the archive to allow use of the data.

Tool Overview (Hardman):

EN focuses on developing core tools; priorities are set by the Tool Working Group (TWG), and there is a new release every 6 months.

Generate Tool: Generate creates PDS4 labels from PDS3 metadata; there are similarities with PPI's Docgen tool, and there is a long-term plan to merge the two. A couple requests for enhancements are in the queue.

Validate Tool: Validate supports the Build 8a release; new features will be added as priorities are set by TWG. Build 8b will include content validation of arrays, flags to disable content validation, and improved support for XML Catalog files. In Build 9a there should be validation of naming rules; in Build 9b referential integrity for local identifiers will be checked, statistics for fields and array elements will be checked if given in the label, special constant use will be validated, and external parsing standards are confirmed. Raugh questioned validation of file naming rules because there are no "rules", just recommendations.

Transformation Tool: Transform supports 29 transformations; new capabilities are added based on TWG priorities. The current version can convert PDS3 labeled tables to PDS4 labeled tables and translate between binary and character PDS4 labeled tables. Transform depends on several underlying libraries, which complicates open source efforts. Build 9 planned enhancements include Array_2D_Map to GeoTIFF and Array_2D_Image to PDS3 Image.

Inspect Tool: Inspect allows visualization of PDS3 and PDS4 products; it parallels SBN's PDS4 Viewer and should eventually replace NASAView. Requirements are being reviewed by TWG; a development release should be available to TWG by the end of February.

Tools Service (Hardman):

EN is looking at wrapping command-line based tools with a service-based interface utilizing software developed by JPL's MGSS organization. The service offers a REST-based interface for passing parameters to the command-line application; results are returned in JSON format. Single-file or batch processing are possible. The software would be at EN; if the file is local, a valid URL would have to be provided so that the executable could find it. Scholes asked about time-out limitations; Sean said EN has not looked into those, but single file operations are 'comfortable'. Development work will continue, including packaging the software for deployment to users.

Validation In Depth (Cayanan):

Validation is possible by using the core model, user-specified criteria, the label, and/or an XML Catalog file. Validation can compare data with labels, references, and content validation. Table content validation includes that values match data type, field format, max/min limits, bit limits for Field_Bits, etc. Scenarios include single product validation, validation of directories of products, validation of bundles and/or collections, and schema referencing (supplied with the tool, referenced in the label, and/or referenced in an XML Catalog file). Discussion followed on which schema referencing scenario is appropriate for acceptance testing. Sean said we have never decided what the answer is; the answer may depend on how long each option takes. Development will continue through Build 10a and 10b with priorities set by TWG.

Mike ran an example validation using a bundle in the DPH examples area; it included a table, a few images, and a few documents and took less than 15 seconds. Raugh asked about bundles that include products generated under different versions of the IM; caching of IMs improves speed, but the variable versions may be an obstacle to efficient validation. In the case of SBN, the products are relatively simple; one recent version of the IM may be sufficient. Huber suggested that a warning be issued when a recent IM clashes with content that was acceptable under an earlier version.

Guinness asked whether files in directories that are not part of the PDS4 bundle/collection will be flagged as errors; this is important because some migration scenarios place PDS4 labels into PDS3 data sets, and the old PDS3 labels would remain in place whether incorporated into the PDS product or not. Simpson and Raugh expressed somewhat different views on whether PDS4 archives have required structures and file/directory names. Hughes said he liked a suggestion made by Slavney that unexplained files be identified with informational messages (not errors or wanrings). Huber asked whether PDS has an obligation to send NSSDCA copies of PDS4 archives if the data have already been archived under PDS3. Simpson noted that NSSDCA us our backup; we should back up our PDS4 holdings even if the data files have already been backed up once before under PDS3. Crichton suggested an action item for future discussion. Raugh said SBN intends to make its PDS4 migrations *better* than original PDS3 holdings, so they have few reservations about backing up both.

Tools Open Discussion (Hardman):

Raugh said it took her an hour to find an old version of Validate; is there a way that versioned software could be located more quickly? The fact that software version numbers are not related to Build numbers adds to confusion. Since the software is behind a firewall, users (who will not know the password) will not be able to access this material. Hardmann cannot solve the firewall problem; he suggested looking at Change Logs to determine which version of software would be appropriate.

Guinness suggested information on program functions be more easily accessible from the Tools web page; typically users have to download an entire package to reach the documentation, which eventually may tell them they have the wrong program. Web site reorganization is in the works with a goal of making it easier for users to navigate to the things they really want.

Several people commented that the PDS4 Dictionary look-up tool should be made more functional; Mafi said it is difficult to find enumerated values, for example. Gordon did a search for "instrument type" and got 124 hits, most of which are PDS3; he got no hits on instrument_type.

PPI is using the Volume Validator for Juno data; King asked what options are available if the tool is deprecated. He likes the way it reports errors. Three people raised hands when asked how many other people use Volume Validator. Its long-term future is problematical. It can only be used over the network using Firefox; Chrome and Safari won't work because of security issues. Sean could make a deployable version. The software is available on-line and can be installed locally.

High Speed Data Transfer at GEO (Scholes):

Large data deliveries are typically by ftp, http, and data brick. Washington University has a 40 Gbps internal network (WURN) and 10 Gbps internet2 connectivity to the outside. Aspera is the high-speed platform; its propriety software uses bandwidth efficiently. Typical downloads of selected GEO data take place at 600-700 Mbps, so 1 TB can be transferred in about 3 hrs (compared with a week to ship a data brick with the same data). "Automated Cart fulfillment" allows 10 GB transfer to CIS/JPL in less than 2 minutes. Scholes wants to test with other DNs, support for the entire GEO archive. Because of the expense of the Aspera license, GEO also needs to look into licensing options which would allow broader use. Hardman said that he did some experiments with GEO; JPL IT Security cut the connection because they though Sean's machine was the target of a denial-of-service attack.

Preparation for LPSC (Stein):

A triple-sized booth has been reserved; Chanover is assembling material that can be used for informal trainings on Tuesday-Thursday. They are hoping to announce availability of the training before the meeting so that more people will know about it. Sign-up sheets are open for people to staff the booth and to take the training.

Training:

Law asked about the status of the PDS training coordinator. Banks said that the plan is to hire a full-time coordinator for 1-2 years to get the effort off the ground. Morgan said that everything is contingent on what PDS finds in its budget for this and following years; even FY18 funding may not be known until the end of March.

Guinness asked whether the recommendations from the previous coordinator (intern Nicole) are available; Banks replied that the recommendations have been organized into a package that can be passed to the new coordinator. When asked for details, Banks said that most of the package appears to be training materials that have already been circulated (see https://pds.nasa.gov/pds4/training/). Morgan added that we are still in the information gathering phase, learning from each new experience; during the summer there will be more chance to distill the lessons learned into more formal guidelines.

Discussion:

Huber asked whether PDS has an obligation to back up at least one copy to hard media. Morgan said he doesn't know; if there is a policy, it is probably very old. Crichton said he believes there is only a requirement for three copies; media is not part of the policy.

Huber then asked about the status of security issues needing repair. Morgan said Pat Michaels has been moved to another job; there is a list of remediations drafted last August that are awaiting funding.

PDS Tech Session

14 February 2018 (Day 2 of 3)

Westin Pasadena (Plaza Room)

191 North Los Robles, Pasadena, CA

(https://pds-engineering.jpl.nasa.gov/content/2018_tech_session)

Attendees:

Rafael Alanis, Maria Banks (by phone), Carole Boyles, Mike Cayanan, Dan Crichton, Cristina de Cesare, Mike Drum, Mitch Gordon, Kevin Grimes, Ed Guinness, Sean Hardman, Lyle Huber, Steve Hughes, Chris Isbell (by phone), Joni Johnson, Todd King, Connor Kingston, Bill Knopf, Emily Law, Tania Lim (by phone), Maria Liukis, Joe Mafi, Tom Morgan, Lynn Neakrase, Anne Raugh, Boris Semenov, Dan Scholes, Susie Slavney (by phone), Tom Stein, Jesse Stone, Kathryn Sweebe and Catherine Suh.

Welcome (Crichton):

Crichton greets everyone and opens the second day of the tech session; he also introduces Catherine as note-taker for the day.

Namespace Dictionaries – Part 2 (Gordon):

In response to Raugh's LDD presentation from yesterday.

Versioning & Provenance: There should be a record of changes that are made to the local dictionaries — i.e. what version was used to create the LDD, etc. Mitch suggests that Ingest_LDD *not* be discarded, which is the current practice, and that Modification History would go in Ingest_LDD which would then be captured in the two XML files generated. He would also like to create an extra XML label which would be a permanent part of the PDS archive, available internally and not to the user.

Regarding dictionaries and Information Model versions, they are difficult to navigate. The Schema Page should be reorganized by IM number and not by dictionary; the WITT team will discuss how to help a user easily navigate the namespaces and find the relevant dictionaries in the correct versions. There is discussion of the naming conventions for the dictionaries and its versions, of how to distinguish between version and build of the IMs and LDDs. King suggests following the naming convention that Geometry uses, which is PDS_namespace_version#_build#. Raugh suggests doing away with all the individual files and packaging all the latest dictionaries into one .zip file which users can download.

Raugh elucidates on an issue stemming from dependencies between and among discipline dictionaries and things dependent on those – that if one discipline dictionary changes, there is no way for dependent discipline dictionaries to know unless the IM is run and built again. This could be addressed by generating a new version of each discipline dictionary with each update of the core dictionary and vice versa, but this will be postponed until there is a pressing need for it.

Ed says that there should be documentation of the dependencies and references between discipline dictionaries. Jordan responds that there was a previous attempt but it wasn't very successful, and Steve adds that the LDD_Tool *should* be able to do this but that it hadn't. Steve and Jordan will perhaps take another stab at it, but in the meantime, point builds can be scheduled every 2 or 3 months.

Require all LDDs be built using LDDTool with no additional hand editing. Is this Policy or Best Practice? This should be a policy, standards reference, or requirement for discipline dictionaries. Raugh recommends that we first implement and demonstrate the various upgrades and changes to be made to LDDTool; in this way, the documentation and tools, validation and sanity-checking will all be in place. Then there will be much more confidence and weight to the SCR. There will likely be a half-dozen SCRs that will be submitted through the stages. Mitch Gordon will be responsible for these SCRs and for the timing in which they are done.

Ingest_LDD class needs to be changed to include IM and Modification History (Mitch will write an SCR), so that the LDDTool can write the label for the LDD product. Specific SCRs will be submitted for specific requests (Mitch). A dictionary steward may need to adjust his or her Ingest_LDD class, but this will not apply for a few months yet.

(Note: Mission dictionaries need to be specifically, separately addressed at some point in the future.)

Action item: Revisit release process and make it applicable to point builds of the IM. Assigned to EN and tied to the calendars referenced in the following paragraph.

Build & Release: Mostly addressed in previous theme. Additionally, Anne Raugh will write a 'Best Practices' document and post it on the wiki. (Which wiki?) Also, it would be helpful to have a calendar with all deadlines and regressions to which dictionary stewards can refer. One should be posted on an internal site for all nodes to access, and another should be made available to the public. There is a question to be addressed of how detailed each calendar should be.

Development & Configuration Control: Problem reporting mechanism and tracking changes can be handled by PDS wide GitHub as it already has an interface in which users can report and track issues and changes. Additionally, there is steward for each data dictionary. Another is to have a 'bug report' on the main site which would be passed to the appropriate steward. Note that not every issue reported needs to trigger an SCR, so there needs to be a gatekeeper. This gatekeeper could also be responsible for alerting the appropriate node of said issue or bug. Implementing this problem reporting mechanism linking the PDS GitHub account from the main

PDS site is to be done by Engineering. Reminder: problem reporting and tracking changes are only a part of the question of configuration management. EN must make sure every node has access to the PDS GitHub, and DDWG will set the plan on how LDDs will be deposited into GitHub. Sean and Todd have begun implementing the first and latter half respectively.

Overlaying a PDS4 Bundle onto a PDS3 Volume (Sue Slavney):

Assumptions are that some PDS3 data does not need to be touched or altered, there will not be duplicate data files, one PDS3 volume can equal to one PDS4 bundle, there will be a pointer in the PDS4 data back to the PDS3 label, PDS4 will not import PDS3 .XML files, and question of how to handle data without .XML files. Looking at the bundle directories, Guinness points out that there's no way of knowing to which collection the PDS3 Volume belongs. Lyle suggests a slight revision of the root directory, perhaps to convert readme and errata to documents.

There was a rather lengthy discussion of data and directory structure. The standards say one thing but many members present say another. There is a PDS policy that DNs can store data any way they like so long as their storage interfaces properly with the central Registry. The Standards Reference only discusses files and directories in the context of data transfers — and then only as a recommendation when the parties to a transfer don't want to invent their own system. One way to answer the question of whether the constraints need to be adhered to is to tested the PDS4 bundles. Susie's PDS4 bundles are able to be validated but have not yet been tested with the harvestor, so Susie and Lyle will both find a structure that works with the harvestor and one that does not. It could be helpful to eventually have a Best Practices document for this.

There is consensus that the rule for <parsing_standards_id> be changed to allow its value to include 'PDS3 Data'. But not all PDS3 data are parsable, for example, JPEG2000.

NSSDCA (Group discussion based on powerpoint slides provided by Stef and Pat):

In response to Best Practice for Bundle Creation.

Concerning the question of when bundles should go to NSSDCA, NSSDCA asks what interval provides a reasonable safety net? This is to be determined individually with each node. Furthermore/Separately?, NSSDCA advises changing the bundle product version even though the bundle product itself hasn't been modified. Steve wonders if the Manifest Generator can handle that, and Raugh agrees that there is a need for a smooth way to do this with as little human intervention as possible.

Sean reminds us that NSSDCA is not ready to officially accept PDS4 data but is accepting and working with test data, even seeking a wider variety of data. Raugh asks whether NSSDCA is capable of handling incremental changes. I did not get Sean's response.

Regarding possible repetition of data: PDS4 bundles need to be delivered to NSSDCA even if the data was already delivered in PDS3 volumes. Any flagging of extraneous files that are not part of a PDS4 product is the responsibility of the node. Any departure from these practices requires an update to the MOU.

Service Overview (Hardman):

The services include Registry, Search, Transport and Report; only Search has public visibility. Sean is searching for a way to extract dataset metadata. These core services have been relatively stable for the last few releases; the current focus is on indexing suggested refinements for the search UI and search results.

There are out-year plans for refactoring of these services. Meanwhile, WITT has been prototyping a solution for improved and incremental indexing. There are also plans for upgrading Apache Solr and taking advantage of any new features as a result.

Report Service: Sawmill not going so well, so they are swapping it out to the open-source ELK stack (Kibana, Elastic Search, and Logstash). Currently, both are being run side-by-side and once a consistent operation is verified, they will work with each Node to develop custom reports.

Tracking Service: 3 main areas of Delivery Tracking, Status Tracking, and Subscription. Delivery Tracking is to track product deliveries from instrument teams to the PDS. (Raugh says that she needs this service for delivery to NSSDC.) It ideally replicates the functionality of Cassini Archive Tracking Service (CATS) in that you can see delivery date, allow delivery to be rejected, etc. This is being implemented because a basic spreadsheet is not sufficient and a new subscription service is needed. Emily asks if the subscription will include a better vehicle of announcing data releases and Sean says yes. Status Tracking looks to capture everything from release status, archive status, NSSDCA status and DOI assignments. This manifest will need to become a permanent part of the system bundle, and there is a process agreed upon with NSSDCA for which Engineering will be responsible. Subscription will be an improved implementation of its current service. These improved services will be rolled out in Builds 8b and 9b.

Services will be maintained with each build. Anomalies and features requests from PDS staff with accounts may be submitted at https://pds-jira.jpl.nasa.gov. Sean's preference is to have the tools open-source but sees no compelling reason to make the services open-sourced.

Containerized Deployment (Hardman):

There are multiple deployments at any given time being handled by Engineering, so having a container would be helpful. EN is looking into container technology for deploying PDS4 software, namely services and not tools, and will use Docker in the first prototype. After testing at the EN, they will draft DN testers and hope to have Build 9a deployments containerized.

EN is researching containers rather than virtual machines (VMs) as the latter uses significantly more system resources. Raugh asks if there are any security advantages to justify using containers and there are several responses in the affirmative; Carole points out there's such a list available somewhere and Sean will make it available.

Search In-Depth (Hardman):

Stein brings up the issue of incorrect lat-long values and/or unknown variables (i.e. size of image given only the center lat-long coordinates) and how this affects search results. How should this be addressed?

Sometime before Build 9a, a working group needs to be assembled to review mapping strategy and search field naming rules.

If anyone would like any more (detailed) examples of mapping, contact Sean Hardman.

Service Q&A (Hardman):

Updates registry by adding new things; does the search configuration have to reconfigure from scratch or it's good to go? Currently, adding things regenerates the index and/but the search is good to go.

Is there a way to supercede a label? Irrelevant, because at the moment, search will return only the latest version of a LID.

How easy or difficult would it be to rollback a registry? Is it possible to apply a patch to a single product? Relatively easy and possible. There are a few ways to do it, and Hardman can do it. Configuration control needs to be enforced though.

King regarding LDD and GitHub: we'll need a LDD generate service in which input is the specification file and output is the LDD or schematron. It's a call-to-LDD in a wrapper. Hardman thinks it's a good idea and can be done.

Roadmap Findings (Morgan & Maria):

Roughly $1.2m is the top offer and there are four augmentation proposals whose financial allotments seem about right. Cassini and Messenger are mostly covered; still waiting on Dawn and Rosetta. For the next 2+ years, estimated funds are $1.3m which are yet to be approved. The list of the 'remaining high priority legacy data sets identified for translation to PDS4 through 2021' will be revisited.

The timing of steps listed on slide 3 are TBD by each instrument team. Slide 4's spreadsheet of node efforts in helping active missions make their conversion plans needs to be updated.

DOI Process (King):

A DOI is intended to be a long-lasting, location independent, and globally unique identifier for an immutable PDS product. A DOI is applicable to PDS products that have been or will be ingested into the PDS4 registry and applies to both PDS3 and PDS4 Collections, Bundles, and Documents. Registering for DOIs for PDS3 DataSetsets are not yet available but on the to-do list.

The format of a DOI is "doi:[prefix]/[suffix]"; prefix is a number that identifies the registrant agent and suffix is a number that uniquely identifies the data item and is assigned by the data agent. The registration process is relatively quick and easy.

Several use cases: create and assign a 'new' DOI, update metadata in an existing DOI, and reserve a DOI for use in future PDS products so that it can be put into the metadata but still be kept private until it is ready to be public.

When a DOI is resolved, it'll lead to a landing page (LP) which gives information about the data and clear links to the product cited/referenced by the DOI. Every active registered DOI requires a fully resolvable URL that points to a landing page. This is the ideal, but not all DOIs and LPs are configured this way. The expectation is that there should be a clear and immediate path to the data product.

Include DOI in the label? SCR-206 has been submitted but not yet approved. Take note that the IM would need to be tweaked for this. King says it'd be nice to have it as a tag element, not as a keyword or description. This would allow for the metadata to be pulled directly in the same manner that other metadata such as the LIDVID is harvested. Currently, how is the DOI found? Hardman will integrate it into the search and, whether the DOI is in the label or not, there will be a way to track it. Once the tracking service is online, each landing page will also say its DOI (it does not at the moment).

Need DDWG to review and approve the metadata that are associated with DOI to send to OSTI.

To what extent should the assignment of DOIs be limited i.e. Collections and Documents only? An action for DDWG to determine. Lyle thinks Bundles would be more useful and Raugh likewise products. The question is more so that does there need to be a policy i.e. does every Collection need a DOI? Where do we derive the task to go after a DOI? It would be early on in the process of citing data. Nodes need to lead or set the example for users.

Wrap-up: Raugh strongly advocates (1) reviewing the dictionaries in some way and (2) writing a process for nodes and reviewers to be aware of what it is to review missions?... She will put these more eloquently on presentation

PDS Tech Session

15 February 2018 (Day 3 of 3)

Westin Pasadena (Plaza Room)
191 North Los Robles, Pasadena, CA
(https://pds-engineering.jpl.nasa.gov/content/2018_tech_session)

Attendees:

Rafael Alanis, Maria Banks (by phone), Carole Boyles, Mike Cayanan, Dan Crichton, Cristina de Cesare, Mike Drum, Mitch Gordon, Ed Guinness, Sean Hardman, Lyle Huber, Steve Hughes, Joni Johnson, Todd King, Connor Kingston, Bill Knopf, Emily Law, Tania Lim (by phone), Maria Liukis, Joe Mafi, Tom Morgan, Lynn Neakrase, Anne Raugh, Boris Semenov, Dan Scholes, Dick Simpson, Susie Slavney (by phone), Tom Stein, Jesse Stone, and Kathryn Sweebe. Catherine Suh

Data Dictionary (Raugh):

Rigorous Validation
Have to write schematon rules to validate. List of general rules, templates, guidelines, best practices will be needed to help write schematron. Provide examples for missions, but tricky for disciplines.
Regression testing – important to have schematron rules to validate. Stewards should be the ones to submit the rules for testing with LDDs, provide test labels. May need a steward support group can come up with a standard ways to develop the schematron rules, and develop a list of things that schematron rules must have. Validation can only be enforced and verified by human eyes.

In conjunction with the next MC F2F, organize one day to talk about schematron rules, dictionary validation issues.

Discipline vs project (mission) dictionaries
What requirements on discipline dictionaries might be only best practices/recommendations? PDS does not have man power to enforce project dictionaries other than what's constrained by LDDTool which is required to use. PDS can provide training of best practices and recommendations. So, again, should there be a policy to require projects to use LDDTool?

Level of support and integration to be provided for active and archived project dictionaries.

Need to identify the metadata and file set need to be preserved for archived project dictionary. Need to think about for all dictionaries.

Need the same level rigor for validating both type of dictionaries.

Discipline Dictionaries
Need review process – what does it mean, and what criteria to use – to help quality control, to help develop a uniform set of dictionaries.
In order to help develop the procedures, come up with best practices and something informal to help the stewards would be a good step forward. What can be done .
Will be helpful to have delivery requirements for discipline dictionaries including user documentation, integration into search interface, a set of regression tests.
A small steward group to work these will be good.

Project Dictionaries
How to present project dictionary to the peer reviewers to ensure it does get reviewed?
May be same support for both project and discipline dictionaries should be in place.

Do we need a formal hand-off procedure from project stewardship to archival stewardship of the dictionary by PDS? In mission bundle, or documentation bundle? EN should be the administer and keeper of project dictionaries.

Develop a procedure to take over the project dictionaries and register them.

DOI Wiki (Raugh):

Anne walked through DataCite's Oct 2017 v 4.1 schema/metadata shown on:
http://sbndev.astro.umd.edu/wiki/DataCite_Schema
Notes are highlighted specifically for PDS consideration and applicability.

SBN is working with DataCite directly for their citation, vs, OSTI.
OSTI has their own schema and requirements.
DDWG will take a look of the mapping to have a PDS-wide agreement.
Results of mapping can eventually be incorporated into IM.
OSTI is willing to work with users to develop a more detailed set of metadata, going the right direction.
If we don't coordinate assign and relate DOI, we would have data management issues from the process , IM and system standpoints.

Data Migration Discussion:
NAIF's registry is fully up to date with LIDVID that's searchable, as a stop gap using proxy labels. NAIF's data is identical for PDS3 and PDS4. NAIF's data are ever cumulating. For most part, NAIF's labels are identical from mission to mission. NAIF is asking if proxy label constitutes as data migration. PDS MC will need to decide.

Develop a simple set of best practices for constructing context products and pass it onto DDWG.

Context Products Best Practices (Gordon):
Need a simple set of best practices. For PDS3, we have description from catalog files, but are not appropriate, sufficient for PDS4 context. There’s also inconsistency in including LID vs LIDVID, vary wildly from mission to mission.
Mission references instrument hosts which reference instruments.

Action items (All):
Went through captured action items for all 3 days. Last action is for all attendees to review the notes and action items and send comments/updates to Law.