PDS Virtual Tool Summit
23 February 2015 (8 AM – 2 PM PST)

The PDS4 Virtual Tool Summit was a teleconference organized by the PDS Engineering Node (EN) to discuss PDS4 tools and plans for their development.

Participants:

Mike A'Hearn (SBN)	Ron Joyner (EN)	Carol Neese (SBN/PSI)
Tilden Barnes (SBN)	Todd King (PPI)	Eric Palmer (SBN/PSI)
Dan Crichton (EN)	Ludmilla Kolokolova (SBN)	Anne Raugh (SBN)
Lyle Huber (ATM)	Emily Law (EN)	Mark Rose (UCD)
Patty Garcia (IMG/USFS)	Mike Martin (Mgmt)	Boris Semenov (NAIF)
Mitch Gordon (RINGS)	Stef McLaughlin (Mgmt)	Ed Shaya (SBN)
Sean Hardman (EN)	Tom Morgan (Mgmt)	Dick Simpson (RS)
Trent Hare (IMG/USGS)	Lev Nagdimunov (SBN)	Susie Slavney (GEO)
Steve Hughes (EN)	Lynn Neakrase (ATM)	Irma Trejo (ATM)

Introduction and Objectives (Crichton):

Purpose of the meeting was to discuss tool development across PDS — so that key requirements are known, gaps are identified, and an integrated plan for the future can be developed and implemented. Crichton anticipates future meetings at approximately six month intervals to ensure coordination.

EN has inventoried and classified tools and presented current status to MC. Tool classifications include core tools that are needed for preparing and submitting data under PDS4 standards. There are also discipline node (DN) tools being developed locally. Tools being shared by international partners are of increasing interest, although IPDA members are also eager to use PDS tools. IPDA will be assessing existing tools and identifying gaps. PDS may be able to use tools developed by international partners, but those will not be discussed today. Hardman will be creating a data base of tools supported directly or indirectly by PDS, including rules for shared development. EN will push open source licensing, which will facilitate international partnering.

Core tools are those that are tightly tied to design, generation, validation, and transformation of archival data, which are central functions of PDS4. There is both PDS3 legacy software and also new PDS4 tools. Core tools must be traceable to PDS Requirements. Discipline and mission-specific tools follow in importance. Names and descriptions of additional tools being developed within PDS and within missions that should be added to the data base should be sent to EN. There may also be user-provided tools that enhance usability of certain types of data products.

When PDS advertises a tool, the organization is obligated to provide some support; what support is to be expected or required needs to be discussed.

Role of IM in Tool Development (Hughes):

Principles of "Domain of Discourse" define how humans interact with machines and how machines interact with other machines. The PDS4 Information Model (IM), which is based on input from DNs, contains the PDS4 principles. Developers should be able to create new software based largely on the IM and products derived from it, such as the Data Dictionary.

EN Core Tool Status (Hardman):

The EN philosophy has been to build a set of common libraries based on open source utilities; the libraries then provide a foundation upon which more powerful tools can be built. A shared code repository is needed to facilitate this.

Generate Tool: *generate* can accept a PDS3 label and translate it to PDS4, or it can be used to generate a new PDS4 label. In command-line mode it can be included in pipelines;. *generate* also has a Java-based interface. Hardman wants to streamline the Java library; the tool has become a 'library' in itself. He is working to improve the *velocity* tool and support additional input sources.

Validate Tool: *validate* checks product labels against the IM and the schemas and Schematron files derived from it. This tool works only for PDS4 (there are separate validation tools for PDS3). The user can specify which version of the IM (and other controlling files) is to be used; alternatively, *validate* can determine the IM version(s) from the label itself. The tool checks for case, checksums (when provided), referential integrity, and other issues. Its output is text and an XML file; the latter can be processed into other formats, such as for reports. EN is working with UCD to add validation of data product content (tables, fields, binary, etc). A lot of the functionality will be moved into the Core Library over coming months, so that validations can be carried out in a number of different contexts which are outside *validate* itself. Raugh said she cannot use *validate* in her production environment; she and Hardman will set up a time during the next week to discuss this problem.

Transform Tool: *transform* calls *generate* to translate PDS3 labels to PDS4. FITS transformations will be added in the next release. Transformation of other formats (*e.g.*, ARRAY, CDF, etc) will be considered. Certain types of VICAR conversions are possible now; but they are prototypes of label manipulations rather than full product translations. PDS4 Tables can be converted to CSV. Some image label transformations are possible; but Hardman needs more examples. Future enhancements include the ability to process more than one image object per product. Martin asked about pre-PDS3 products in PDS; are those files (such as Viking and Voyager compressed files) on the list for transformation?

Hardman would like to add them, but he thinks other transformations have higher priority. Since they can be very large, Gordon asked whether there is a size limit on converting binary tables to CSV; Rose said the conversion is on a record-by-record basis, so there is no practical limit.

Ames Core Tool Status (Rose):

UCD has developed several tools for PDS3, which are in maintenance mode. The *LACE* label editor is for PDS4 only; Mark is tracking it for maintenance, since there has only been sporadic use. Additional features are planned for *LACE* such as better support of XML, but these are on hold pending completion of work on validation code.

Volume Validator needs to sync with *Validate.*

Image Browser works on PDS3 Mars data sets; it could be upgraded to PDS4 but it relies on ODE, which does not yet work with PDS4 data. Currently there is no funding to go farther.

Table Explorer was upgraded over the summer by an intern; but it doesn't support all PDS4 tables, and it never supported all PDS3 tables. Martin submitted some bug reports; Rose said they have been partially addressed, and the others are in progress.

Planned PDS4 development includes the PDS4 Core Library, which is upgraded as the Information Model evolves. Raugh asked whether any of the web-based tools are deployable. Rose said 'no', but deployment is possible in the future; Raugh volunteered to be a tester when one is needed. Martin asked whether there is any hope of getting a simple users guide for the Core Library; Rose said 'no' because development and debugging are taking precedence. Martin then asked whether anyone at EN is available to create a users guide. Hardman said he can put it on his list of things to do, and an upcoming handover of the Core Library to EN would be a good time to create one.

Hare noted that the Google Earth API is being retired; Rose said UCD has not decided how to react, since future availability of the existing software is not clear.

A locally installable validation tool should be available by the end of FY15.

Core Tools and Software — Gaps and New Capabilities (Crichton):

EN has made a list of possible tool and software issues within its own current plan for tool development; the list will guide prioritization of work.

Information Model: Hughes is expecting to expand terminological entries (making them more specific) in the Information Model; there have been no other requests. Martin asked whether Hughes is documenting the steps he goes through in maintaining and updating the model; Steve is working on that, but the simple answer is that he "pushes a button" to

update the model. The trick is to know where all of the files going into an update are located.

Local Data Dictionaries: (1) Gordon asked whether he can get an IM Spec and UML diagrams from LDD Tool (as is possible with the common Data Dictionary); Hughes will add those (and JSON files) to his list. (2) Mitch is working on the Geometry LDD; he would like to have an LDD with common geometry building blocks, then separate LDDs for orbiters, flybys missions, landers, etc. He hasn't been able to do this. (3) Gordon also would like a way to reference available LDDs (such as development versions) that have not been released. Mitch and Hughes will discuss possibilities off line. (4) Slavney asked whether *IngestLDD* is the only way to get an LDD into the system. Gordon and Hughes agreed that *IngestLDD* is the practical answer; there is no formal requirement. Slavney said InSight wants to add to an existing LDD manually; but there are questions about whether this will work. Raugh noted there are a lot of ways to write schema files, which may or may not be compatible with PDS4; Hughes added that *IngestLDD* limits unacceptable free lancing.

Tool Interfaces: Martin asked whether a GUI could be wrapped around some of the tools, such as *generate*. Crichton replied that GUIs represent another layer of maintenance. Hardman said he is not planning to add GUIs, since most are intended to be used in pipelines.

Validation: There will be a few EN *validate* improvements in the next PDS4 build; there is a longer list planned for the following build. Crichton asked whether 'visual inspection' is a high priority at the DNs; several people responded 'yes'. NASAVIEW is one 'visual inspection' tool; a set of requirements will be needed for a PDS4 successor; simply upgrading NASAVIEW may not be the best solution.

Transform: King suggested that transformations of Array to other formats (such as ASCII Table or Delimited Table) be given high priority. If a generic Array transformation is available, then transforming image products (possibly the biggest need) becomes a special case.

Schedule: King noted that maintaining lists of requirements for common tools is appropriate in many cases; but mission needs cannot always be met on common schedules. Sometimes DNs have to develop ad hoc tools or make ad hoc improvements to existing tools because waiting for the long-term common tool takes too long.

Deployment: A discussion followed on network versus deployable software. Raugh said her community likes locally installed software using industry standard installation procedures; then she doesn't have to guide installations at many sites, which often require custom links and libraries. Hare said he has been working with environment issues for some time; there is not a good solution for everyone yet. Crichton suggested a survey of DNs on future directions in deployable software; Martin agreed. Hare said adopting a 'virtual machine' model could allow PDS to release a single version so long as all users had

adopted the same 'virtual machine'. Some reasonable virtual machine candidates are appearing.

PDS4 Portal and Search (Hardman):

The Search Service is based on the Apache Solar protocol. The Registry Service is based on the Information Model; parameters are harvested from incoming products and made available to Search, but the process needs more automation. There are currently three sources for search: PSA, PDS3, and PDS4; ISRO has provided some information and may be integrated by the end of this year. The Registry is rebuilt each time it is updated; currently this takes about an hour for each update. Sean would like to implement an incremental mode, which could save time as the number of products increases.

Search is deployable so long as there is a local Tomcat server. The Query model is still under development. Multiple query formats can be supported; in principle, any parameter collected by the *harvest* tool can be used for Search, but this general capability is still a goal. Product-level search is being developed. Ultimately, tools will be searchable as well as data.

Martin asked whether future plans include moving metadata from DN registries to the Registry at EN. Hardman said bundle and collection information will be copied to EN to support 'context' searches, but product metadata will not be copied; it is sufficient that the registries are linked. King asked whether there is on-line help for each of the services; Hardman said on-line help is available for Registry but not for Search, but Search help will be added.

DN Portals and Search (Law):

King will discuss PPI search during his node presentation. Gordon made a few remarks about plans at RINGS. There were questions about how the interface between queries and local search interfaces should be modeled; Hardman noted that this is covered in the PDS4 architecture document.

Portals and Search — Gaps and New Capabilities (Crichton):

Crichton asked whether users should be able to search across PDS4 at the product level. King replied that the search should be a 'cascade' so that the system does not become overloaded with the total number of hits and the user gets a response in a reasonable amount of time. LIDVID search and retrieval should be possible for any product. There appeared to be consensus on returning the latest 'certified' version when a LID search is requested, possibly with status information on past and 'in development' versions. Search should be driven by the IM.

Discussion followed on what parameters should be supported in a top level search; time, target, mission, etc. were proposed. Raugh said 'mission' and 'time' aren't useful in many SBN searches; she also said that a generic 'no' is sometimes helpful when there are not going to be any hits. But Martin pointed out that the Image Atlas will return a hit when a search parameter has no value.

King mentioned that PPI has adopted some internal guidelines on 'mark down' in <description> values; this can provide helpful output formatting. Raugh questioned whether 'mark down' is good practice in archival material; she also wondered whether there would be consequences for the PDS-to-ADS interface, but then decided that only simple text is usually sent to ADS. King will provide an example of what PPI is doing.

PPI Tools (King):

igppdocgen is used for creating labels at PPI. It is supported by a library of locally written functions, such as code to do time conversions. PPI uses this tool for label generations since PPI has requirements that are different from the ones guiding development of the EN *generate* tool. King will send the PPI list of requirements to Hardman.

mimic was needed because PPI had trouble reproducing and transferring large groups of files; it gains efficiency by assuming that the volume to be transferred is large. It can be used in incremental mode, transferring only the material added since the last request.

ditdos is the back end of the PPI user interface; it now operates on top of Apache Solar. It can do search, selected format conversions, and display data. Although Todd's impression is that users don't want a lot of web interface complications, he thinks they also don't want to install software. So the default at PPI is to interact with users through a browser where they can locate and display data quickly and easily.

SBN/PSI Tools (Palmer and Neese):

Small Bodies Image Browser (*SBIB*): SBN needed a tool to locate and display selected Deep Impact images; it has been adapted for Vesta and now displays cube data. SBIB provides an image data base which can be downloaded in a few seconds; the 'images' are then downloaded as the user needs them. Users can quickly leaf through the data products, stopping as needed to study them more closely. The tool will be upgraded for Eros and Ceres data. Palmer has resisted requests for analytical add-ons; users should download the same products through regular SBN channels and apply IDL (or equivalent) analysis tools.

OLAF: *OLAF* was designed to make it easy for members of the SBN community to ingest data into PDS; data providers do not have to understand much about PDS. *OLAF* has been in operation since 2003; it is being upgraded for PDS4. Users enter metadata through an interface; *OLAF* then creates labels. *OLAF* is limited to data sets smaller than about 5 GB. *OLAF* is especially good for ground-based observations, but it can have mission applications.

It works with images, tables, and several special cases of tables. A migration mode will allow translation of PDS3 data to PDS4.

NEO Survey Browser: NEO Surveys have been conducted for detection of small bodies. The data consist of hundreds of terabytes of images of the sky; the software takes the name of a known object (or a set of coordinates) and returns a list of images including that target. Only some survey data have been incorporated so far. The tool does not depend on whether the data have been archived under PDS3 or PDS4.

Small Bodies Data Ferret: This tool allows users to search the SBN holdings based on target and data set name. It does not cover large data sets from spacecraft missions, which can be searched using other tools. *Ferret* will be upgraded to handle PDS4 data.

PDS4 to IDL Conversion (Shaya):

Shaya wanted to read XML and put statements into IDL structures; he then extended *READ_XML* code to become *READ_PDS*, which extracts data and metadata into IDL at a level at which products can be processed. He demonstrated how to use the software, which can be downloaded from the TOOLS portion of the SBN web site. It is presently in 'beta' mode while PDS4 evolves. Several dozen people have downloaded *READ_XML*, because it reads generic XML; not many have downloaded *READ_PDS*.

Other Topics

Mike Martin wondered whether there were tools that might be applicable to APIs and access to the deep archive. A discussion followed on the utility of PDS Challenge for tool development and how those results can be best integrated into mainline PDS development. Morgan suggested that the PDS people who have Challenge experience discuss the subject and provide recommendations at some future time.

Next Steps (Crichton):

Dan plans to have today's presentations uploaded to the EN web site; he will also make a report to MC. Dan has been taking notes; he and Emily will integrate theirs with Simpson's notes and distribute the results. Morgan thought this was an excellent start.

Things to Work:

1. Document IM process (being worked)
2. Output HTML, UML, and JSON files
3. Allow reference of any LDDs available (not just the released ones)
4. Provide repository for code sharing and open source

5. Define requirements for Visual inspection which is high priority
6. Include ARRAY transformation which is high priority
7. Form a working group to develop PDS multiplatform deployment model (e.g., VM, Docker)
8. Provide search service online help
9. Access every product using LID
10. Form a group discussion of tool development using PDS Challenge

Original (Simpson): 2015-02-23
Revision (Emily): 2015-02-25
Additional revisions (Simpson): 2015-02-25
Minor Edit (Emily): 2015-02-26