PDS Technical Session
April 6-7, 2005

Wednesday, April 6, 2005

Attending by phone:
Joel Wilf, EN
Erin Means, EN
Sean Kelly EN
Tyler Brown, EN
Patty Garcia, Imaging, Flagstaff

Attending in person:
Anne Raugh, Small Bodies, UM
David Tarseo, Small Bodies, Tucson
Rose Early, Small Bodies, Tucson
Chris Isbell, Imaging, Flagstaff
Myche McAuley, Imaging, JPL
Lyle Huber, Atmospheres
Steve Hughes, EN
Alice Stanboli, Imaging JPL
Ron Joyner, EN
Mark Showalter, Rings
Ray Walker, PPI
Ed Guinness, Geosciences
Susie Slavney, Geosciences
Steve Joy, PPI
Ludmilla Kolokolova, Small Bodies, UM
Todd King, PPI
Puneet Khetarpal, Small Bodies, UM

Introduction and procedural issues

Questions:
First of all, what does this group do under the new way of doing things?

The Technical Session's job is to make recommendations to the Management Council, which then advises Dan Crichton as EN lead.
Do we need formal rules for doing this? Or can we come to consensus informally? Majority rule?

We'll never reach consensus on every issue. We agree that decisions will be reached by majority rule, one node, one vote, plus one vote from the Radio Science representative. The EN *does* get a vote. Votes and dissenting opinions will be recorded. Decisions will be presented to MC.

Dan Crichton is unanimously elected to be the chair of the Tech Session. In his absence, Anne will run the meeting just to keep us on track. She will also generate a report of the meeting's conclusions to deliver to Management Council. Susie Slavney is taking notes.

Steve H: Who represents PDS as a whole from a technical perspective?
Susie: We need to designate a spokesperson to present an issue to MC. Answer: It should be the chair (Dan).

Ed proposes we consider revising the agenda in light of yesterday's events.

Todd: There are three main topics to discuss:

1. SRD status and plan
2. Top priority tools for the nodes, to present to MC
3. PDS System Architecture

Geo's standards presentation is removed from the agenda, to be presented at a later meeting.

Overview of EN Tasks – Roles and Responsibilities of EN determined by MC – Ed Guinness

(Presentation of slides from MC meeting showing EN Roles and Responsibilities [R&R]. These have been revised and agreed upon by the MC, and have been sent to Bill Knopf.)

Procedure: Geo distributed a strawman set of R&R to the nodes for comments before the meeting. Comments were compiled into a revised set of R&R, which were then discussed, revised again and agreed upon during the MC meeting. They were sent to Bill Knopf to take to NASA HQ, and if approved by HQ will become the EN Charter.

Note: The final list of R&R has *not* been prioritized.

There were no major questions or discussion about the presentation.

Overview of EN Tasks – Overview by Steve Hughes

(Slide presentation from Steve)

Part 1: Development Life Cycle
This is an old presentation (Sept. 2004) so some of it may be outdated.

"The PDS Online System" is the current implementation of the PDS Architecture. It includes components at EN and the DNs, including DN repositories, data bases and catalogs, and links to resources outside PDS. Integration of components is key. There are now >2000 links to node resources.

Mark S: we have new resources that have not been included in the integrated system. The questions are raised: How do new resources get into the system? Can the architecture query nodes for their resources, rather than have the nodes submit them? Need a tool to update (discussion postponed; this and other issues raised are recorded in the Issues and Action Items document).

Anne: How can we input our software products as well as our data products?

Todd: The system should be active rather than based on builds.

Anne disagrees: there are good reasons for having a system with builds.
(Discussion postponed.)

History of D01, D02, D03: >90% of PDS holdings are accessible as of PDS-D02. Anne: What about the other 10%? Steve: The missing data are things like Magellan SAR-EDRs, which are stored offline.

Mark: Why do all queries have to go through JPL? Susie: Nothing precludes your node from having its own public resource. Mark: Our issue is that our MAC O/S 10 is not supported for product servers. Sean: Actually we have a great product server for O/S 10. Mark: I mean query server. Why is there only one query server at JPL? (Discussion postponed.)

Todd: Go back to the missing 10% -- why are they missing? Steve: EN has only "archived" data sets. Susie: Ongoing mission data sets are not marked STATUS=ARCHIVED until the end of the mission, yet they are accessible through EN. 100% of Geo archives are *accessible*, even those offline, because you can request them. Issue: What does "archived" mean? Steve thinks once a data set is archived, it does not change. (Big discussion about whether the 10% is really missing. Added to issues list.)

Back to the presentation. Gathering of requirements for 5 year plan, including the "smackdown"… further editing created the SRD "which is currently in signature cycle to Project Manager, Project Scientist, Project Systems Engineer, and NASA Sponsor". Steve says there was a management decision (as of Sept. 8, 2004) to put this document into signature cycle. Lyle: nobody seems to remember this happening. (Discussion postponed; resolved later in the meeting that the SRD will be reviewed by a Tech Session team and then sent to MC for approval and signature.)

MRO review: identified IV&T and other tasks from requirements.

Development process: formal process adopted by EN. One major release per year. "Incremental iterative approach."

There will always be 2 systems: operational and developmental. Some changes must occur independently of yearly release (bug fixes, ingesting metadata, etc.). There's always thorough regression testing after each change. Todd: Why one release per year?

Steve: Arbitrary choice, but corresponds to funding cycle. Plan included design review of each release, but never got to do that for D02 (?).

Lesson learned from D02: Add PDS-wide acceptance testing to catch errors due to data missing or wrongly ingested.

PDS Online System is now under Configuration Management (CM) using CCC/Harvest system supported by JPL. CM controls development life cycle. Development is done on a test system. Erin: The development system gets copied onto the operational system when a build occurs. Anne: Best practice is to do a clean build on the operational system so as not to perpetuate errors from the dev system to ops. (**Action:** Anne will discuss with Joel offline.)

Open Discussion of SRD: What does the Tech Session recommend to MC that we do with the SRD?

Include the IV&T Addendum in this SRD discussion? Steve: No. All: Yes! The Tiger Team completed its task of developing requirements. Those requirements still stand. Two issues caused the IV&T work to be stopped: 1. Making requirements after a prototype has already been designed is not the right way to do this. 2. The impetus for IV&T work came from HQ or from MRO review or from Bill K., NOT from MC, so MC has taken the reins and decided that the IV&T work is not a priority. Hence the ROI was canceled and IV&T work stopped. (*Much* discussion.)

Ray W: MC agrees that the process that was followed for the IV&T work was wrong. It's now up to the MC to move expeditiously to determine what to do next with the IV&T requirements. It was not fair to the EN to have its work tweaked by several different bosses.

SRD Addendum: We keep it. It's now part of the SRD. [General agreement]

Todd: What should happen to the SRD now? Ray W: We want to see it as it is today. This group may then take action on its own to update it or not. Todd: I need to look at it again to know whether or not it's ready for signature.

Ron: We need another tiger team to go through the SRD and perhaps make changes.

Ed: How do we do any development before we have requirements?

Mark: Propose we table the SRD discussion until we have direction from MC, and to give Dan a chance to be involved.

Steve: Managers need to decide whether the SRD represents where we want to go (not every little detail).
Anne: Techs should give their managers a list of high priority things to do, #1 of which is to consider the SRD.

Ed: What about Level 1 and 2 reqts? Steve: They were done by MC in December 2004. Ray W: They are almost signed off, truly minor changes were proposed, can be considered done. Steve: Dick's sub-bullets were effectively the beginning of the level 3 reqts, which should map closely to the SRD. What is needed now is a mapping starting with Level 3 and working down to SRD. One of the tasks on the EN list is to do this reconciliation. (The IV&T tiger team has already done part of it.) Ron: We could finish the Level 3 work now. **Ed: Propose that this group explain the state of the requirements to MC and ask if we can finish this work and deliver to them a complete set of Level 3 reqts reconciled with SRD for review and approval, in order for development to begin.**
Vote: The proposal is accepted by voice vote.

Lyle is opposed, because we have already completed and reviewed this SRD. Mapping it to L1 and L2 reqts seems redundant because they are so general.
Joel: SRD is supposed to be just reqts for system and tools, not for whole organization. So not one document. Need feedback sooner rather than later.

Todd: Using the IV&T work as a model: we created an addendum for the MC to review. Same approach? Ask MC to review SRD for weaknesses and then task us to make an addendum to fix them? Ed: If we do the review first we will find those weaknesses anyway. Mark: Don't present managers with an unfinished document. Integrate addendum into the main document. Steve: This will take too long – 6-7 weeks!

Lyle: Set a deadline.

Steve: Can be done by one person, a technical editor who can do the analysis and deliver a report. Ron: alternate approach. We know (from IV&T work) that there is work to be done in these areas. Need a tiger team to look at it.

Anne: So will the review be done by committee or by technical editor? General agreement that it can't be done by a tech editor.

Ron will lead the group, which is composed of Lyle, Todd, Susie, Myche or Patty. Ron to make a schedule. Will include initial review period, one telecon to discuss, another period for an editor to make changes, then final review by email. Finish by week of April 25. Deliver report to MC.

EN Tasks List as requested by Geosciences – Steve Hughes

(Presentation of Jan-Dec 2005 Plans spreadsheet)

Work shown in order of system build, lists priority, status, driver, and source.

IV&T work: Can EN stop working on it now, or wait for direction from Bill Knopf? Ray W: Stop now. Assume Bill Knopf has MC's direction to stop the work. If consequences

come back from HQ, it's an MC job to manage them, not an EN job. Even though there's a JPL award fee from CalTech riding on this work.

Joel: Work is underway now (due by 4/30) on updating requirements for catalog templates because currently they do not contain the necessary information for updating the PDS Catalog. (Spreadsheet lines 14-16. Driver: Must provide consistent standards and validation tools for the customer.) Also includes update of Data Dictionary to allow lvtool to correctly validate catalog files (due by 5/30). Steve: Deadlines are fixed because of award criteria; we can't change them.

API/OODT task: developer's package is now out for review.

Issue: Mailing lists are out of date, e.g. Mark is not on the tech list, but he's not sure. **Action to Valerie**: We need to know who is on mailing lists and how to update them.

Todd: It's an issue that EN is charged to do work by JPL award criteria, and their work is not necessarily what the nodes would have them be doing. We have no input. Also, we are impacted by their work; we have to do some work to help them, yet we have no input.

Tech Session has a concern about success criteria imposed on EN without node input, and asks MC to fix this situation. (Anne will word better and include in her report.) The JPL award fee is NOT just within PDS, but part of the JPL-wide "report card".

Ed: How do items get on this list, and how do they get prioritized?
Steve: Input to issue tracker queue comes from the following: the web site gripe button, with a judgment call about whether it can be fixed. HQ stuff automatically is Level 1 priority. MC provides input items. Internally they are prioritized. We now see we need priority recommendations from MC (with advice to MC from Tech Session). Todd: EN needs some discretion for gripe-button issues. MC should not micro-manage. Anne: Propose we discuss by email a triage system for handling such issues. Todd: Should be up to EN. Ray W: MC should shield the EN from big issues handed down from HQ or MRO or whoever. Steve: In the past, everything was high priority and we had to do it. It's a relief to have the direction from MC.

Lunch break, reconvene 1:15.

PDS-Wide Tools Development – Joel Wilf

Joel has an online tool survey, organized by functionality, not by existing tools. (http://pds.jpl.nasa.gov/survey/Tools_survey.html) Also lists legacy tools – which ones need continuing support.

Purpose of the tool survey is to collect priorities from those people who use the tools, also to get feed back for the requirements.

Ed: How does the survey fit in with new way of business in which tool development is directed by MC? Joel: This would be the Tech Session's recommendation for tool priorities to MC. Ed: Have one response from each node, collate responses from all nodes. Need **deadline: end of next week**. **(Action on each node.)**

Ron: also prioritize categories? Joel: Yes, we can add a box beside each category if you want. Mark: not necessary; it should be obvious from responses. Too hard to rank relative priorities among categories. Lyle would like to prioritize categories. Ed: Let's use the comment box on the form to add explanations or whatever the form did not cover. (So: do not add a box to prioritize categories.)

Todd: Some items in the survey don't come from the SRD; is that how we want to do things? Anne: There must be a way to introduce new items during the life of the system as needed. Is this the system? Yes, for now, because we are catching up and reorganizing. Todd: An alternative is for each node to nominate its 3 most important tools. Todd would like to see MC direct EN to begin particular tool developments ASAP. Proposal by Anne: **Each node list its top 3 tools in the comment box.** [General agreement.]

Categories with a "to be discussed" box: If you have suggestions for these, put them in the comment box.

PDS Architecture – Steve Hughes

Comments on OODT collected from nodes (Steve's list)
Mark has not responded to Steve's survey but he has some comments to send.

Responses from Sean Kelly (slides):

Improved product retrieval speed (Joyner) by factor of 10 by increasing block size, but still slower than HTTP or FTP. Experimenting with another kinds of server that streams data rather than using blocks, expected speed increase of 30X, no additional memory usage at the EN.

REST = Representational State Transfer: Architectural model for web services. Would not affect users except by increased speed. Nodes would need to run a web application container (e.g. Tomcat). Server manager no longer needed, removing a security risk and returning complete control to nodes. Lyle: ATM has had trouble with Tomcat using up 95% of the CPU. Are there alternatives? Sean: Yes, both free and commercial. **Action: Sean will look into alternatives to Tomcat to help Lyle.**

Todd: Now that we are at the point of defining the PDS architecture, should we reconsider the use of OODT and perhaps develop something else? Steve: Architecture is separate from implementation.

Anne: We have no OODT application now but we are developing one. Should we continue OODT development, using current interfaces? Steve: Yes, as Sean said, the interfaces will not change when OODT begins using REST.

EN (Sean) has an action to work with the Rings Node on specific issues of an interface to their catalog.

PPI Node Server -- Todd King

(See Todd's slides)

DITDOS in now in 3rd generation of providing products online from PPI; i.e., a product server before there was OODT.

Basis (premises on which DITDOS was developed): Efficiency is important; layers of software are not necessarily the best implementation. Metadata is the enabler; should be portable. The volume structure is for people, not software. Data labels are persistent, but data bases are transient. Inventories should reside with products.

Steve H: Volume structure is for people? I thought it was both for people and computers, not to be debated now, but as part of the architecture discussion. Also, the label is persistent but not the data base? What is the source of the search attributes, the label or the index table? Another topic for later discussion.

Todd: we use the label as the basis for metadata; if there's a correction to be made, we fix the label first, then the data base (or index). We think that's the best approach for PPI, though there's no requirement or convention in PDS to do it this way.

Application server: A servlet for each function; provides quick response, extensibility, and portability. Can be personalized (".pds" file type, "portal" philosophy). Shopping cart model.

Data cart: like the usual web site shopping cart. Contents are delivered as a zip file, organized as a PDS volume.

Steve H: What kind of data products does it understand? Todd: A product = a label plus everything it points to.

Tools and servlets are written in Java.

Issue (from Steve H): There are already 2 canonical parsers in PDS; should EN look at Todd's and see if it's also canonical? Myche: Existing canonical parser(s) need an API. Lyle: ESA has a detailed Java validator for PDS labels. EN should look at that. Anne: it's commercial software. Steve: We should validate the ESA parser to see if it's canonical. (Why??)

The interface (web page) is not public yet; still in beta testing.

Data repository does not necessarily store data in PDS volumes (though it could); volumes are generated when requested.

Lessons learned: Use servlets (resident) instead of cgi (non-resident) to speed up initial data base connections. Use different connection for each type of query to improve performance, because of result caching. Scalability is an issue; the larger the inventory the longer it takes to search.

Considerations: Could OODT be used for this? Yes, but… multiple layers of abstraction are unnecessary. MySQL database is used on same server as DITDOS, but any dbms could be used. Steve H: Could repository be separate from data base? Todd: Yes. Data, query server, and DBMS can be on different servers.

System design ("architecture"): Define messages, not implementations. If profile or product query is done using HTML encoding, implementation can use any method (Perl, CGI, servlet, etc.) Profile response is HTML or XML, reformatted for display. Product delivery is either a URL or a data stream, mime-typed or zipped.

Steve H: Architectural issue: What should the URL look like to request a product? Also, what should a product ID look like? Todd: Our system uses an internally generated product ID that's just a unique number. Steve: Adding the node id would make it unique across the PDS system. Joel's method is data set id, volume id, directory path, product label.

Steve H: Another architectural issue is the location service: how to find out where products are located and in how many places in PDS?

How often to have Tech Session meetings? Ray W: Way more often. Mark: Reta wants to have fewer MC meetings. Do we have to travel? Telecons would work *if people are prepared*.

Thursday, April 7, 2005

Attending in person:

Chris Isbell, Lyle Huber, Ray Walker, Ludmilla Kolokolova, Steve Hughes, Ron Joyner, Ed Guinness, Susie Slavney, David Tarseo, Rose Early, Puneet Khetarpal

Attending by phone:

Patty Garcia, Sean Kelly

Continuation of Architecture Discussion – Steve Hughes

(Slide presentation PDS Core Infrastructure Design and Implementation)

Design approach: Modular components, open accessibility for search and retrieval through APIs, distributed incremental development. Use COTS, open source and free software, build on work done at JPL, NASA, Nodes, etc.

System is built in layers, limiting most development costs to top layer (PDS Custom software). Architecture should focus on the middleware level.

Requirements are assigned to system components. Architecture provides framework for understanding components. Strategy: Map architecture to components to functions.

Software design -- 4 tier architecture:

1. Storage tier (repositories, catalogs). PDS-D began with repositories organized as volumes online. Catalogs were built from index tables. Enabled PDS-D to be up and running quickly for distributed systems at nodes.

2. Service tier. Product server provides API for product retrieval; profile server provides API for product search. Both are part of OODT middleware distributed to nodes.

David: Even though DITDOS doesn't use OODT, do you consider it a product server? Steve: Yes, maybe. Depends on what API he (Todd) uses. If we settle on the right API, and DITDOS uses it, then yes it's a product server. The API currently used by OODT is included in the OODT document recently sent out to the nodes for review.

While we have product servers at all nodes, we have very few profile servers; they have to be customized for each node. Defining the API for a profile server is more difficult.

Currently the API is an HTML-encoded keyword=value string with the usual logical operators. We need to agree on what the query should look like, but more importantly, we need to decide what a profile should look like. Currently it's returned as an XML

document (one per product). Another example: the PDS-PSA interface work is defining the profile for retrieving data from the PSA.

David: what profile servers are now in use in PDS? Steve: Mark Sykes has one (The Sykes Page). The PDS single point of entry web page uses one.
Susie: Can the profile server simply search existing index tables? Steve: If that's what you want to do, then EN will provide the functionality to do that. Ron: there's a performance hit to doing it that way. Ray W: performance is very important to our user community; we don't want people pounding on the door. Steve: It sounds like the Tech Session is recommending to MC that they levy a performance requirement on the system. There is a performance requirement in the SRD, which is going to be revised, so now is the time to do this.

We may end up with two pathways into the PDS; the one we have now where you search for pretty specific things, and a more general Google-like search.

Another problem is how do we know for sure we *don't* have something – how to give a definitive No.

3. Web tier components (e.g. validation, tracking, data set view, data set browsers which are currently being converted to use profile servers, subscription manager, volume packer, authentication manager).
4. Client tier (e.g. web browsers, bulk download clients, server manager, etc.)

This 4-tier system is what was proposed in the MRO review.

When we have more profile servers, it will become more important to have a location server to figure out what products are where. Right now that information is hard-coded.

Profiles can exist for documents, software, people, etc., as well as data.

Todd's presentation showed that the PPI system overlays very closely onto Steve's architecture design.

Current status: target system includes only core requirements, no extended system requirements. All future work is stopped waiting for direction from MC. Tech Session can recommend that work on selected components should continue.

Global Product Search – Steve Hughes

(Slide presentation: Global Product Search)

This work came out of Mark Sykes' request to make his products available using a common URL syntax.

Incremental approach: We did a quick prototype for proof of concept using existing profile servers at EN. When we have more profile servers and a location server, we can apply this concept to all the nodes. The goal is to provide a single entry point for users to search for products across the PDS archive, searching on any indexed attribute. Can receive subsets of large data products (e.g. HiRISE). It's up to MC to tell us whether this is something we should pursue.

Ray W. : Where did the idea come from that the index table is the primary source of search attributes? *(Long discussion of the history of the index table.)* The issue was brought up yesterday by Todd as to whether it's the label or the index table that is the primary source of metadata.

This does not preclude the need for any DN Custom Browser; it requires a profile server API to any custom browser catalog.

Drivers for Global Product Search: Increase in volume and number of products predicted. ADS interface with PDS: PDSQUERY Interface using Dublin Core IDENTIFIER attribute. PSA/PDS Interoperability. Users are asking for a single point of entry for searching.

PDS-D subsystem provides the mechanisms to implement the Global Product Search.

Ed: If Geo makes its own profile & product servers, is it necessary for queries to be submitted through the Query Server at JPL? Steve: No, you can have a local profile or product server and write an application that queries them directly. Ed: If we need help to do this…? Steve: Sean is available to help. In fact, if Geo has a profile server available, Steve or anybody else can write software to talk to it if they know about it. The reason for going through the query server at JPL is that the average user wouldn't know about the Geo application.

Goals:

1. Confirm that technology experiment proves the capability. Will ultimately require a profile server interface to all product catalogs.
2. Consult nodes and gain consensus on:
    a. API for global search (query mechanism, e.g. PDSQUERY)
    b. Design and implementation of a PDS location server (a location server works by using profiles of profiles!)
    c. Managing a combination of distributed and centralized catalogs
    d. Location of primary site for product retrieval
3. Explore other search technologies
4. Research performance, reliability, scalability, maintainability etc.

PDS Data Resource Management -- Ron Joyner

This demo was developed because so many people requested an online interface to do updates to the PDS Data Set Catalog, for use by the DNs. Yesterday someone asked for a different kind of system ("pull" technology) that would involve a process at EN that queries the node databases in order to do updates. EN needs direction from the tech session (via MC) as to which of these methods to pursue, if any.

(Demonstration of Data Resource Manager prototype on Ron's password-protected web site.)

Each node should be able to access only its own resources.

Steve: The next step would be a similar tool for adding data sets, not just resources (i.e. links) to data sets. Do we want MC to direct us to do this? Ed: No, this is too far down on the priority list.

David: What does the resource point to – a data set? What is a data set? That is, what are you asking to see when you say "I want a data set"? (On the PDS search page, it's what you get when you click on View Information For <data set id> in the Information About the Data Set column. The resulting page is known to EN as The Sykes Page because Mark Sykes asked for it.)

Chris: If I update a resource, it will automatically update the PDS search page? Answer: Yes.