01: Welcome and Tech Session Overview-Dan

Welcome and Tech Session Overview-Dan

- Welcome and logistics: breaks, Caltech Guest wifi, please send presentations to Sean & Emily
- Purpose of the Tech Session
    - meant to be a forum for the nodes to discuss standards, tools, and services
    - should be discussion-oriented
    - questions are welcomed
- Day 1: Standards Support and Plans
        - Sharing Lessons learned
            - improve implementations across PDS
            - Identify gaps to address in FY17
            - understand how current and planned missions will use PDS4
            - Understand scope and need for support
- Day 2: Tool Support and plans
- Day 3: Services support and plans

02: Overview of the PDS4 Data Standards-Steve

Map

Overview of the PDS4 Data Standards-Steve

Overview: Standards —> IM —> Documentation
PDS IM: Address data Variety

- PDS4 Im plays key role in defining the PDS info reqs
    - defines entities and relationships
- PDS4 system is enabled by an informations model-driven approach where the IM is the cornerstone of the system
    - write software to information model
- Sources and productions of standard themselves: What feeds the IM?
    - PDS-specific: information requirements, Domain knowledge
    - Reference models of Info systems: OAIS
    - PDS reqs and policies set the foundation for the information requirements

        - PDS-Specific Domain knowledge: expert information about the “things” in the domain that should be described and associated with the data to make and keep it useful
            - fundamental structures, contact, integrity, reference, documents
- OAIS Open archive information systems reference model
    - digital object: an object which is real data
    - physical object: an object which is physical or tangible i.e. Saturn
    - conceptual object: an object which is intangible i.e. Cassini mission
    - It’s the description side that we have been working on most in the past few years
- Data Dictionary Reference Model: Governance entities, registration authority, steward, namespace
- Registry reference model: ebXML (Electronic business XML) standardizes the secure exchange of data
    - defines: registry database schema, generic registry object, core attributes

IM Database

- all of the source information for the PDS IM database is managed using Protege
- Database content is a merge of the domain model and the data dictionary
- specialized tool written to export the database to formatted files used by the data providers

IM Specification

- HTML description of of how info is being stored in the database
- Examples of XML Schema and Schematron files
- Question: Schema and Schmatron rules: are the written by hand or generated automatically?
    - Answer: automatically, IMTool uses IM model database to generate most schemas, some exceptions, everything comes from information model
- JSON: entire IM database is dumped into JSON files, defines how the information fits into the model, can be adapted to other tools
    - any tool developer can go to this info to learn about the IM
- Registration configuration parameters: tells harvester what is coming through the pipes to be ingested into the system
- Query Model: defined constrains on mission science data collections, set of required attributes
    - Schematron file will validate that the proper attribute exist in the file

Standards and product development lifecycle: what resources do we have that address and guide all of the products? Is it the data dictionary?

- PDS4 Edifice example of how PDS4 is built: the base is the formation of almost all archives
    - the the classifications of data are applicable across most digital archives, 10 categories

Post presentation discussion: Will we be able to use the same version of labels forever?
No. backwards compatibility doesn't always happen, but core is pretty much the same. Some nodes have needs that others do not, but we are a confederation and some sacrifices or pains for overall better good.

- how do we make sure we are considering technological impact? Who determines non-backwards compatible changes? Part of technical assessment. This is reviewed, engineering node will head reviews, but will also need feedback from community CCB
    - can we elevate changes in non-backwards compliance to wider audience?

03: PDS4 Product Development & Process Discussion-Ron

Map

PDS4 Product Development & Process Discussion-Ron

Lessons learned, identify areas of improvement, use archive lifecycle as a baseline
Seven steps of archive process
1) Orientation: establish contact with PDS, IPAG, MPAG
2) Archive Planning: what to archive, when, how, request unique identifiers, establish common schema, DMAP
3) Design: design bubble, collections and data projects, organize data and documents into collections, design production process (design XML labels for all products in the bundle)
LDD tool (LDD Tool)
4) Bundle Generation & Assembly: generate documentation products specific to the mission, generate LDD, generate collections in bundle
4) Validation: quality check: validate the metadata and XML labels, peer-review, verify data-product pipeline

- visualization tool? currently at SBN

5) Ingestion/registration
6) Search and distribution: product search, product distribution to data users, bundle delivery to deep archive
Future wants and needs?

- PDS4 Requirements and policies: provenance? chain of authenticity
    - Earth science is looking at this already W3C, we have tools to monitor changes-do we need a requirement?
- Standards and PDS4 Documentation: identity gaps ambiguities, and misconceptions

How do we test local data dictionaries? Need better documentation about how to do this, or a tool

04: Lessons learned and plans for PDS4

Map

A) PSA- Santa Martinez: PDS4 Implementation in the PSA, ESA

- Documentation: PSA PDS4 Archiving Guide, data providers to PSA ICD, and PSA PDS4 Schemas
- Data organization:
  - One PSA Mission bundle: PSA Schematron. LDDs, and contact products
    - standard set of collections for mission bundle: document, context, misc, spice_kernels (under discussion), xml_schema
    - Instrument bundle: Data (raw, processed, calibrated, derived), calibration, document, browse, misc, context, spice, xml schema
- All raw/calibrated data divided into mission phase —> sub instrument, range of days, range of orbits, observation campaigns
- Comparison of LADEE and MAVEN to how they are archiving data (see link on slide [11-12?] for more info)
- Using PDS4 Schemas, PSA PDS4 Schema
  - Project, or mission-specific level schemas, also used for label validation
  - XSL2PDS tool developed to convert PDS4 data product description in Excel to XML label template: Because of length of mission, ESA wants to keep the templates for reference over time
- Planning to use LDD tool because it will be part of the matter IM
- Useful to have more documentation of Data Dictionaries,
  - all mission dictionaries in central repository with explanation of purpose, use for international community
  - How to distinguish between mission dictionaries? How do they know which dictionary to use?
- Only using the validate tool currently
  - Core library package as Java APL interface
- Issue with bundle/collection versions: difficult to follow
  - using accumulating bundles, unto 5 deliveries a day
- Bundle generation takes place at PSA, differs from PDS
  - trying to follow standards of of PDS, but also finding new solutions for their unique structure and organization
- No recommendations on how to upgrade to new version of PDS4
- Interested in lessons learned for workflows
- Issues with Target Context Products: Missing information in existing targets (description)
- Traceability: provenance of how label was generated? want reproducability
- See presentation, not PDF for most recent version- issues, examples and recommendations

B) LADEE Lessons Learned- Lyle Huber:

- LADEE archive history
    - 3 instruments: all were relatively simple, single bundle with raw, data and derived collections

LADEE Archiving Process

- Made early contact with mission and instrument teams: this was the early test of PDS4
- understand products expected
- build XML templates
- review sample
- Peer review
- Sean for archive

Tools Needed

- Oxygen used for template design
- Validation tool used during review
- One instrument used LDD, LDDTool was not operational at time

Question: archived in 1100, how hard would be to move up to 1600, 1700? Most core is easy, one-offs may take as long as all the rest, student-generated Python code

- would you even want to do this for every update? PDS4 datasets are so few, it's best to have them comply to the most current data sets, but might not be useful for older data
- Obsolescence of older versions: does common software still support newer version?

Updating Sean's each tool in the beginning: how to accommodate different search terms? Early success to help inform going forward: **Replacing stock examples with real examples from missions**

C) MAVEN Lessons Learned- Joe Mafi:

MAVEN Archive

- ATMOS: 3 instruments, 6 data bundles, 30 data collections
- PPI: 7 instruments, 12 data bundles, 85 data collections
- Possible radio science data archive
- Spice archive associated with 3 archives as well
- Question: are there things that PDS4 didn't let you do that you wanted to?
    - Answer: trouble finding agreement about what is in core dictionary, local data dictionary

PPI Archive procedure

- Data delivered from MAVEN, decompressed and verified, reorganized, label generation, collection & bundle generation, standards validation
- Procedure Details screen: lists which tools used for each phase of process
    - igpp.docgen from UCLA, more info in presentation tomorrow

    - validate tool generates a lot of output, they use a parse tool to sift through the results
- Question: Why did you generate the labels for PPI?
    - Answer: Early on, there was't a standards reference, many new users who were unfamiliar with PDS, we were more familiar. This is not something we plan on doing into the future

Tool Needs

- Collection & bundle generation: would like to eliminate manual process
    - Collection label generation tool
- Standards validation: doesn't verify logical identifiers, data structure description validation to match structure of data file
    - Question: Is PDS4 working on this?
        - Answer: Sean will present on Validation tool in coming days, should be available in months
- Validation tools do not check product and referenced LIDS against the registry

D) InSight Lessons Learned-Ed Guinnes:

Insight Status: launch slipped to 2018

- Geosciences node is lead, 4 instruments (HP3, SEIS, RISE, IDA)

Label Design: templates and examples developed using Oxygen

- not particularly user-friendly for PDS4 schema, room for improvement

Label generation:

- HP3 will generate labels with their own software
- SEIS: asked for tool
- RISE: likely will make their own software
- IDA: might have to make the labels for them

Validation:

- Need to make sure LIDS new correct

Ingest

- have little experince, using very old tools

Question: how are search services being used? might want to ingrate efforts

- use PDS labels to drive search

Bundle/Collection Organization
Geosciences node is working on a mission local dictionary with input from some other nodes- imaging

Lessons so far:

- mission dictionary with other nodes is challenging

Concerns/Gaps:

- these have largely been brought up to the tools working group
- LID: look up existing context LIDS (might help alleviate Richard's workload)
- Context products: what happens after creation? Updates? viewing, submitting, downloading context products
    - how to get info shared after sent to EN? Need documentation posted about this

Question: new instruments present, terms are not present in IM, problems?

- can fit into high-level generic terms,more of an issue of search usefulness

Question: because the mission is stopped and will re-start, what about a loss of continuity?

- will want to use most updated model, might have to change. No loss of staff yet

New proposers: this is happening now, increasing support

- this data is coming in before the missions: need to use this as a driver to build infrastructure

E) Osiris-Rex Lessons Learned- Michael Wendell

One bundle per instrument, collections organized by processing level
Currently doing review planning
Discipline dictionaries not ready to support O-Rex

- wanted better documentation for designing labels
- version changes during development process

PDS4 Viewer is used heavily
Lessons Learned:

- tools, documentation, and support are main areas of need
- end-to end archive development documentation being put together as they go
- Changes to standards as development is occurring-is this necessary?
    - the newer dictionaries might be necessary (Geometry dictionary —> IM 1.6)

05: PDART Lessons Learned-Moses Milazzo

Map

PDART Lessons Learned-Moses Milazzo

USGS
DAPS were about 2/year, PDARTS are 10 and growing

- 6 inactive, possibly because of frustrations from PIs about tools and documentation
- Question: are some of these just inactive because they're waiting until last minute?
    - Worried about push for last-minute help, some people just run out of money, some people come back long after the fact to follow-up
    - burden will be on us, we don't have resources to give them so they can help themselves
- Question: how do we discuss quality of DMAP as qualifier?
    - writing a poor DMAP currently doesn't disqualify, merely advisory
    - PDART is the odd ball out: requires high quality DMAP, downgraded if there isn't high quality archiving responsibility
- PIs are not typically data scientists, unsure of how to approach PDS, need better instructions
    - Question: Does the IPAG help?
        - Shared with PIs, not sure of usefulness, fine for proposal, but almost nothing about detailed instructions, no examples, "not a training guide at all"-Lisa
        - "PDS is broken" how to reconcile using PDS3 data and then archiving in PDS4

ROSES 2016 requires DMP: wide variety of requests for support

- major requests to archive DTMs, DEMs, GeoTiff images, GIS maps, GIS projects
    - a number of these requests are in formats that are non-compliant
    - Question: any hope of getting GIS product that is complaint?
        - I don't know how this could happen, probably not, needs to be considered going forward, these formats are quite popular
            - USGS is archiving these kinds of materials all the time: maybe that can be an example of how to make them compliant in PDS
            - many formats have been intentionally excluded to ensure long-term preservation

Plans to support non-compliant products by reformatting and using miscellaneous directories

- Products are derived from PIs products, but not actually the PIs data product
- Question: What is the product type you are planning to use for non-compliant products?
    - depends on the product, in misc? unknown. MAVE is using product ancillary
    - at what point does this become a dumping ground?
    - need another product to continue support community; collection contemporary
        - underscores importance of transform tools: need to convert from archival format to contemporary format
            - have to be very careful with conversions, user might not know what works best in the long-term
    - ACTION: figure out how to solve GIS issues, PDART proposal?

Needs/Wants

- label design software, start w/existing labels
- public library of labels to share appropriate examples with PIs
- Need to use language that is accessible for scientists, not just data scientists
- Training for users: step-by-step handbook, in-person training
    - Planetary Data Workshops: want to have tool to train DAP investigators

Tomorrow: spend some time talking about GIS issue (3:15-4:30p)

- we should talk about multiple formats that are currently non-compliant and how to deal with them
- Goal: have a list of types of data that we need to have a plan for

06: LDDTool, Ingest_LDD, and Dictionary Stacks-Steve

Map

LDDTool, Ingest_LDD, and Dictionary Stacks-Steve

LDD Tool

- performs a temporary load of a dictionary into the IM database
    - parses and merges each file into IM database
    - Validates and report errors
    - reports XML, schema, schematron, JSON and other files
- Same as IMTTool: you should see same structure in LDD runs as IM
- Question: When the LDD tool, do you also get the database?
    - Yes. Data directory in zip
- Capability of producing HTML file?
    - it is an option, outputting IM specification with dictionary included
        - is of entire model, common and LDD
- There is one model, everything is validated against the one model
    - is a temporary ingest
- Configuration-oriented approach to options of LDD tool
- Question: for -s why not let me define the file name root to use?
    - config file coming to help with this, it is an open issue
- Temporary ingest into IM database, against everything in IM

Ingest LDD: purpose is to define the dictionary, classes and attributes that you want to ingest

- 1) define class: class name
- 2) inheritance: info about what it is a type of
- 3) composition: definition, association
- Model created in Protege database: displays class definition, and is nested to show inheritance, composition in detail window

Dictionary Stacks: purpose is to serve as information resource

- a list of dependent and consistent dictionaries
- What does consistency mean?
    - green light in oxygen for local dictionary, testing against different versions of the IM, its okay? Yes.
- Question: can you have multiple version of the ingest in a stack?
    - Answer: stack consists of consistent dictionaries, do not have multiple versions
- Question: what are the next steps?
    - Is Bottom-up or top-down approach better? Maybe it's both?
        - two sets of stacks
    - Versioning: if you don't make a change, don't version it
        - version of local dictionary tied to IM model which it's based

    - develop stack registry out to DWG in the next few weeks for review: recommended stacks

07: Over view/recap of 9/21

Map

- Make sure we have a list of the data formats that we think we need to address
    - What are the gaps?
- Working on prototype for stacks
- A lot of tool discussion: tool gaps? Installation/use
- Documentation: data lifecycle that tracks tools and documentation needs