Strategy for writing a good data management plan for a ROSES proposal

S. P. Joy, R. J. Walker, T. A. King, J. N. Mafi

Motivation

- The Planetary Plasma Interactions (PPI) Node of the PDS is frequently contacted with requests for letters of support as one of the required elements for many proposals coming out of ROSES (Research Opportunities in Space and Earth Sciences)
    - Most of these requests come very close to the proposal submission deadline
- When these requests come in, PPI usually asks the requestor for an overview of the proposal and a summary of the amount and type of data that would be archived if the proposal were funded
    - If the resultant data products would naturally belong at the PPI Node of the PDS, we agree to write a letter of support and one is promptly provided.
    - If the data would more naturally reside at the Atmospheres or Small Bodies Nodes of the PDS, we refer the person to the correct point of contact for the appropriate node.
- If there is sufficient time before the proposal submission due date, we generally ask to see a copy of the proposal data management plan (DMP) in order to get a more complete understanding of the proposed archive and timeline so that we can write a more complete letter of support
- Our experience has been that the DMPs often lack detail and demonstrate very little understanding of the PDS archiving process, standards, and timeline

Overview

- Start Early
- Select the right archive (not always PDS)
- Contact the selected archive well in advance of the proposal deadline
- Learn the nomenclature and standard practices of the selected archive
- Include a realistic work plan and budget
- Archiving data with the PDS

Start Early!
Start Early!!
Start Early!!!

Starting Early

- One of the primary benefits of starting to think about the archive products and process early is that it allows you time to research the archiving options and requirements
  - Various organizations have different requirements for data formatting, metadata, etc.
  - If you are new to these standards, you leave yourself time to familiar yourself with them

- There can be secondary benefits in thinking through the entire proposal to the final data output early in the writing process
  - Sometimes working backwards from the output may help you refine the discussion of the processing required to get there and the potential pitfalls to be avoided along the way

Select the right archive

- ROSES planetary science proposals are all required to archive any resultant data products with the PDS or “equivalent” archive
  - If PDS is not selected, proposers must demonstrate that the selected archive is equivalent
- In general, the PDS archives planetary data (planets [except Earth], moons, comets, asteroids, and dust) from spacecraft (orbiters, landers, flybys) and Earth-based telescopes by scientific discipline.
  - PDS does not curate return samples. These are generally handled at Johnson. PDS will archive data derived from samples (Spectra, composition, etc.)
- The archiving in PDS is organized by sub-discipline (plasma interactions, imaging, geoscience, small bodies, rings, atmospheres)
- Heliospheric data (Sun, Earth, solar wind) are normally archived with the NSSDCA
- Software, models, and simulation codes are archived in the NASA Github site
  - PDS does not archive executable software, only example algorithms as documentation
  - Simulation output are data that can be archived with the PDS or NSSDCA
- Laboratory analysis of samples have multiple potential equivalent archives.
  - As a starting point, users should consider where they would expect to be able to find similar data (i.e. HIgh-resolution TRANsmission molecular absorption database [HITRAN])
  - Some PDS Nodes will accept laboratory data. However if you don’t find similar data at that node, it is unlikely that others will look there for your results.

Understand the nomenclature & standard practices of the archive

- Users familiar with the PDS from years past (PDS3) remember terms like volumes, data sets, catalogs, etc. The current PDS standard (PDS4) uses new terminology like *bundles*, *collections/products*, and *context files* (https://pds.jpl.nasa.gov).
- Heliophysics archives should be described using the SPASE metadata standard (http://spase-group.org). This standard uses terms like *entities*, *services*, and *data (numerical, display, etc.)/granules*.
- People planning to archive code (Models, Simulations, etc.) should use the NASA Git-hub site. Proposals should demonstrate knowledge of open source coding practices and configuration management (development using feature *branches*, release *roll-ups,* etc.), documentation through a combination of the *git commit* comments and github's *native wiki,* and the use of the built-in *issue tracking tool* for user feedback and bug reporting.
- People planning on creating laboratory sample analysis results that are not archived with the PDS should demonstrate an understanding of the required data formats and metadata standards of their selected archive.

Include a realistic work plan and budget

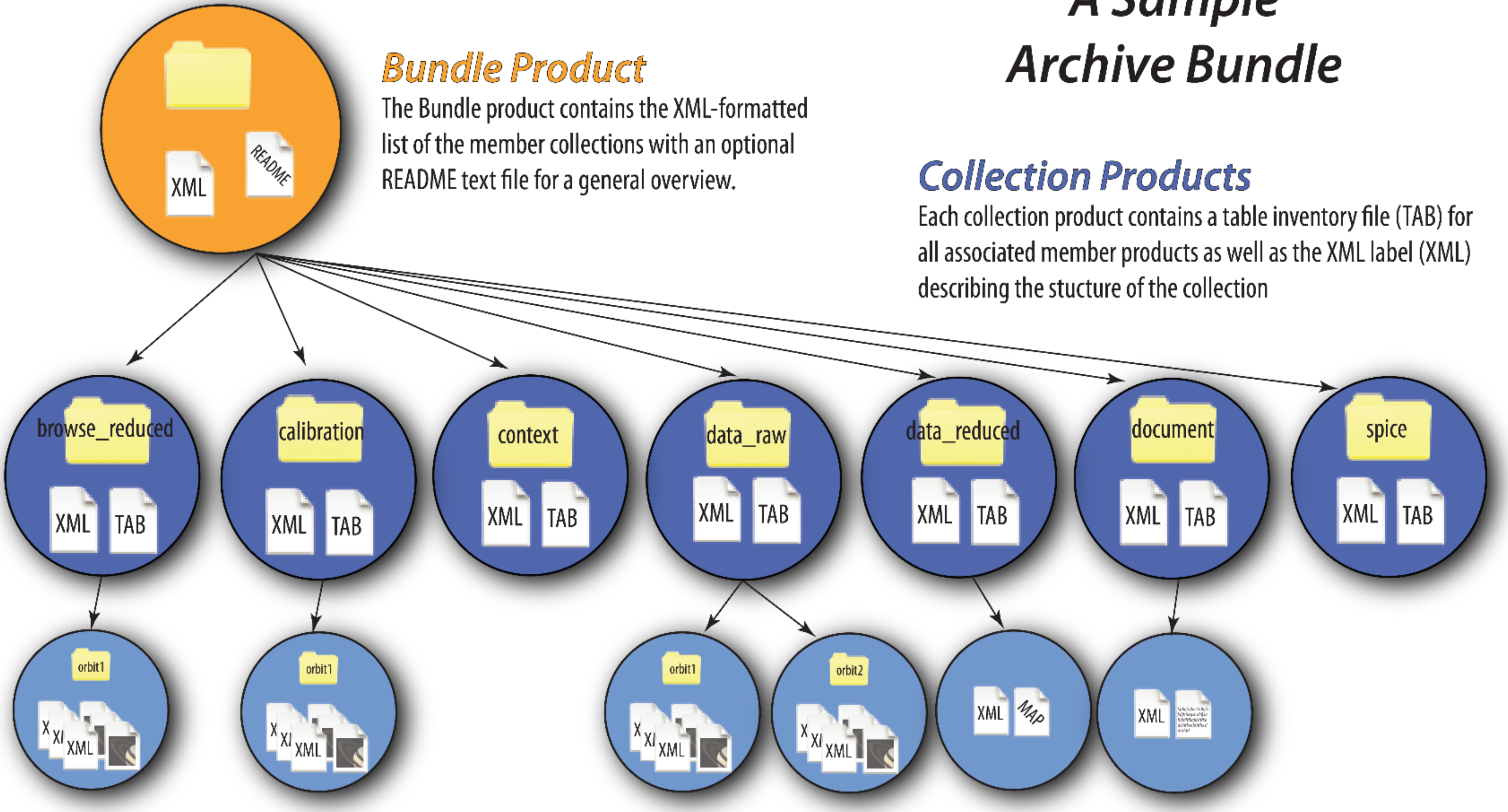
- There are three common problems with most DMPs:
  1. Wrong archive timeline
     - Archive delivery timed to be at the end of the performance period
       - Products should be documented and archived when they are produced
       - There should be time after archive delivery to fix issues that are reported
     - Not enough time to produce and validate the archive
       - Correct time allocation will depend on many factors including archive standards and the teams familiarity with them, data volume, and data complexity
       - In general, at least several months should be allocated to producing and validating the archive
  2. Not enough money allocated to the task
     - Enough said
  3. Wrong personnel assigned to the tasks
     - Archiving tasks are often completely allocated to low cost student workers
     - Timeline and budget should show some senior scientist effort in archive planning/design, documentation, and data validation

Archiving with the PDS

- Most ROSEs planetary proposals will end up producing data products that will end up with the PDS
- All archives submitted to the PDS must comply with the new PDS4 standard
  - Some missions that began archiving with the PDS prior to the release of PDS4 are grandfathered into the PDS3 standard.
- All of the previous discussion applies equally to PDS archiving but there are a few PDS specific items or issues that should be included in your DMP
- The mantra of “start early” is exceptionally important for PDS archives
  - Contact the PDS node that you think is the likely best fit for your data. They may direct you to another node based on initial discussions of data types, sources, or targets.
- If you begin working with a PDS node early enough in the process, many nodes will help you with some thoughts on the initial design of the archive and help you guesstimate the amount of time it will likely take to generate, validate, review, and update your archive products
- **If selected, please inform the Node ASAP so that they can begin to integrate support of your project into their already full schedules**

PDS4 Concepts

A Sample Archive Bundle



Bundle Product

The Bundle product contains the XML-formatted list of the member collections with an optional README text file for a general overview.

Collection Products

Each collection product contains a table inventory file (TAB) for all associated member products as well as the XML label (XML) describing the stucture of the collection

Basic Products

Basic products consist of individual or groups of data files (images, headers, documents, tables, etc.) with their associated XML labels that can be placed into logical groupings (collections)

PDS4 Labels: *A single XML label uniquely identifies the product and its component pieces, describes their structure and relative locations, lists related metadata, and provides linkages (references) to related products.*

PDS4 Documents

- Documents for ROSES Proposal Writers

  At the time of proposal writing, the most important resources available to you are the PDS4 Concepts document (https://pds.nasa.gov/pds4/doc/concepts/Concepts_1.8.0_170406_clean.pdf)

  the Individual Proposer's Archive Guide or IPAG (https://pds.jpl.nasa.gov/documents/Individual-Proposers-Archive-Guide-v11.pdf)

  and NASA's FAQ page on Data Management Plans

  and ROSES Data Management Plan Template

- Documents to help generate PDS4 compatible archives

  After selection, you will need to review the PDS4 Standards reference (https://pds.nasa.gov/pds4/doc/sr/current/StdRef_1.4.0_150922.pdf) and may find the Wiki that has been set up by the Small Bodies Node to be useful (http://sbndev.astro.umd.edu/wiki/SBN_PDS4_Wiki)

PDS4 Tools

- PDS provides tools to facilitate the creation and validation of PDS4 archives and other tools to read and display data described by this standard
- All of the generic PDS4 tools (*Local Data Dictionary Tool*, *Generate Tool*, *Validate Tool*) and libraries to support user tool development can be accessed from the main PDS site at: https://pds.jpl.nasa.gov/pds4/software/index.shtml
- Various PDS Nodes also provide tools for PDS4 label generation, validation, or data visualization
  - PPI – tools to read CDF formatted data to extract metadata and tools to generate PDS3 or PDS4 labels using Velocity Template Language (VTL) (https://pds-ppi.igpp.ucla.edu/software/index.jsp)
  - SBN – tool to read or view PDS4 labeled data, file and label formatting / verifying routines (https://pds-smallbodies.astro.umd.edu/tools/software.shtml)

Archive Overview

- The DMP should begin with an archive overview
- For a ROSES proposal PDS4 archive, this will typically consist of:
  - 1 *Bundle* that links all of the archive components together
  - 1 or more *Data Collections*
    - Data files within a collection should be uniform in structure and purpose
  - 1 *Document Collection* to hold the SIS and any other documents (data user’s handbook, data processing algorithms and descriptions, etc.) that might be included with the archive
  - 0 or more *Browse Image Collections* to store quick-look data plots or images
  - 0 or more *Calibration Collections* to store calibration products
- A DMP that discusses the archive contents and organization using the PDS4 nomenclature demonstrates an understanding of the PDS archive process

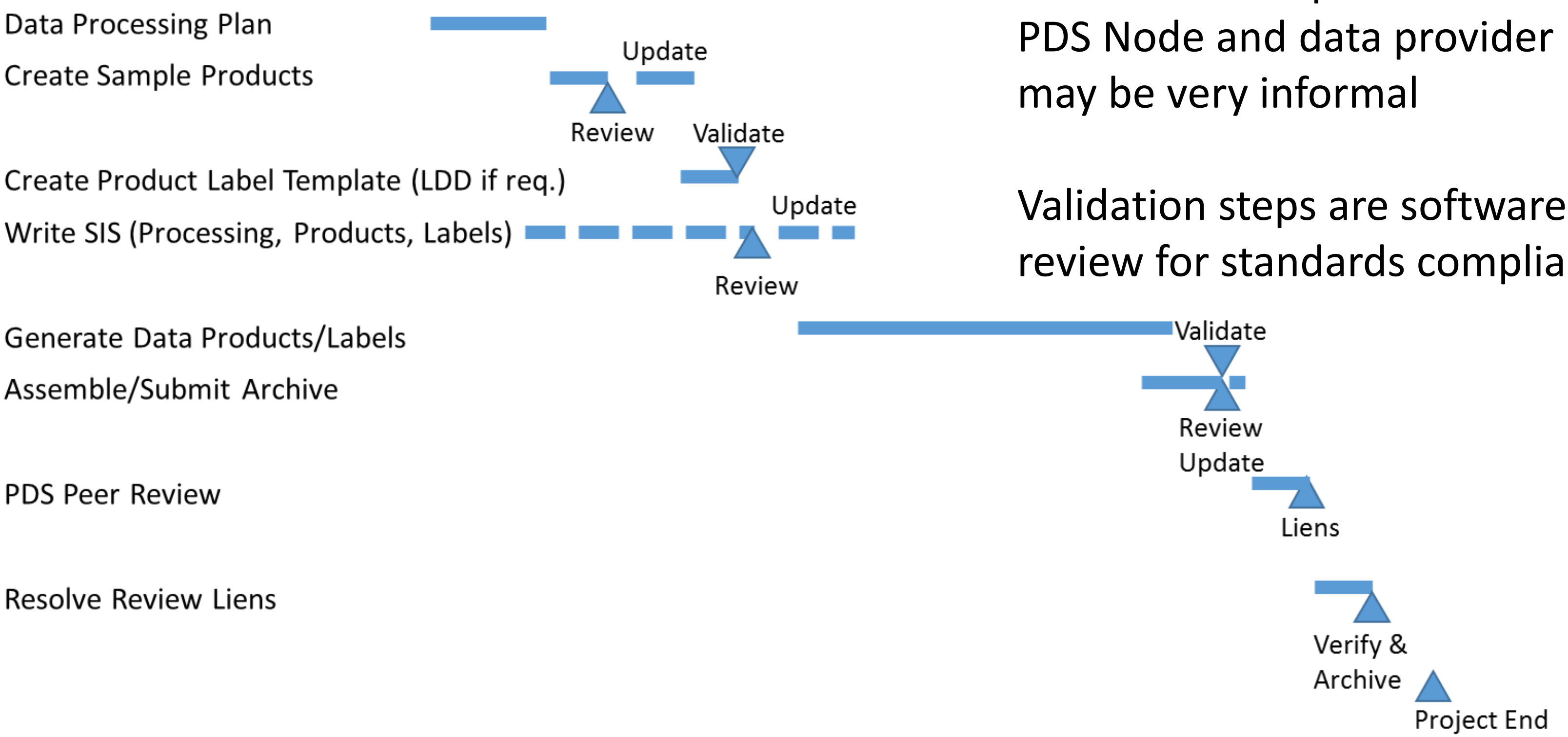
Size and Complexity of the Archive

- All good DMPs should provide an estimate of the size and complexity (number of different file structures and formats) of the expected archive which includes:
  - Total Expected Archive Volume (MB, GB, or TB, ie. 20-40 GB)
    - Estimates are just for scaling purposes and are not expected to be precise
  - Number of data file structures (i.e. Four different ASCII table structures + 2 types of FITS images, full frame and windowed)
    - Each different file structure will require a label template adding time/cost
    - One off file labels can be manually populated, scripts or other tools need to be used to populate label templates when numerous files are expected
  - Approximate number of each type/formats of data files (i.e. the ASCII tables are each single files containing calibration parameters [4 one-off labels], there are between 10 and 20K image files, primarily 2MB full frame with a few hundred 120KB ¼ frame images [two label generation tools/scripts])

PDS archiving timeline

Each review step involves the PDS Node and data provider review – may be very informal

Validation steps are software review for standards compliance



At the beginning of the process, the team develops a plan that will result in the generation of data products to be archived with the PDS. Samples of these products are produced and submitted to the PDS Node for review. The Node will likely suggest minor updates to the product formats or contents. Once the all of the product structures are agreed upon, PDS4 label templates that describe the structure and contents of the archive are created and validated. A SIS that describes the archive components, data processing, etc. is the written, reviewed, and updated. Once complete, the data and labels are generated and the archive is assembled and reviewed, first internally and then by domain experts. Any deficiencies in the archive are corrected, the corrections verified, and the data are archived.

Team Members Archive Roles and Responsibilities

- The roles and responsibilities of each team member who is contributing the final archive product should be explicitly called out in the DMP.
- Typically, systems are designed by senior scientists, implemented by staff programmers or post-docs, and executed by junior staff
    - Senior scientists (PI or Co-I) should be involved in writing some components of the documentation (discussion of data processing techniques or algorithms), data processing pipeline validation, and final archive review. This effort should be called out in the narrative and included in the budget. It doesn’t have to be a lot of time/money, just engagement in the end-to-end process.
    - Including senior scientist in the archiving effort demonstrates a commitment to the archive process and quality
- Every person who works on the archive must appear in the budget and its narrative

2017 updates to ROSES pertaining to DMPs

- All proposals to data analysis programs **EXCEPT PDART** that generate data or software must use the two-page DMP template. These pages do not count towards the 15 page limit on the Scientific/Technical/Management section.
- PDART proposals, being an archive-centric program, are expected to include the DMP in the main body of the proposal
- DMP Template (should look familiar)
  1. **Overview of the data**
  2. **Data types, volume, formats**
  3. **Schedule for data archiving**
  4. **Intended repositories for archived data and public access**
  5. **Plan for enabling long-term preservation**
  6. **Software archiving plan**
  7. **Roles and responsibilities of team members**

Users of the DMP template for PDS archives

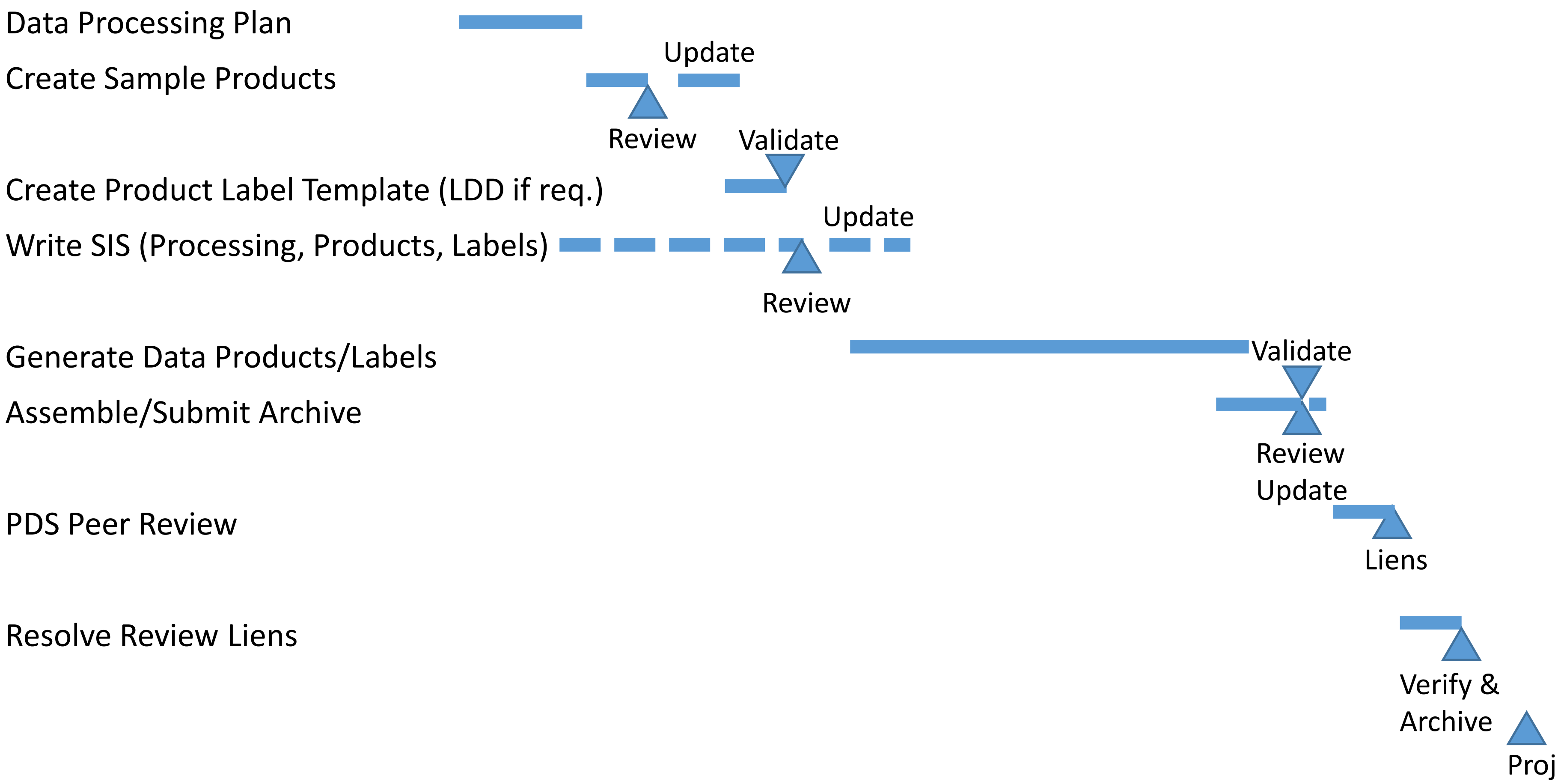
- In section 4 (intended repositories and public access) you can include the statement that all data in the PDS are online and publicly accessible at no charge. Users are not required to register or pay for access.
- In section 5 (long-term preservation) you can state that data that are by the PDS are expected to be preserved for at least 50 years. In order to ensure this longevity, the PDS maintains at least 3 copies of the data that are distributed geographically. In addition, PDS only accepts data in formats that can be fully described by its required metadata (no proprietary or transitory formats)
- In section 6 (Software archiving) you should state that PDS does not archive executable programs but does accept algorithms and source code as forms of documentation. Any software submitted to PDS should be described as elements of your *document collection.* If your output includes executable software, then you should archive these elements at the NASA github site. PDS4 allows these items to be externally referenced in your meta-data.

Summary

- Start Early
    - Good DMPs require more effort than many expect
- Select the right archive
    - not always PDS
- Contact the selected archive well in advance of the proposal deadline
    - They may be able to help with nomenclature and initial archive design which might fold back into effort and budget
- Learn the nomenclature and standard practices of the selected archive
    - PDS4 terminology is very different from PDS3, etc.
- Include a realistic work plan and budget
    - Leave schedule margin for unexpected delays, hurdles, and PDS coordination (Nodes are busy)
- If selected and archiving with the PDS, please let the Node that you will be working with know ASAP so that they can begin working your effort into their schedule

Backup

Schematic PDS archiving timeline



Data Processing Plan

Create Sample Products

Update

Review

Validate

Create Product Label Template (LDD if req.)

Update

Write SIS (Processing, Products, Labels)

Review

Generate Data Products/Labels

Validate

Assemble/Submit Archive

Review

Update

PDS Peer Review

Liens

Resolve Review Liens

Verify &
Archive

Proj