

Improving Algorithm Communication and Data Cognizance Through Standardizing Documentation

Aaron Kaulfus¹, Kaylin Bugbee¹, Alyssa Harris², Rahul Ramachandran³, Sean Harkins², Sean Bailey⁴, Aimee Barciauskas⁴

(1) University of Alabama in Huntsville

(2) Development Seed

(3) NASA Marshall Space Flight Center

(4) NASA Goddard Space Flight Center



development SEED

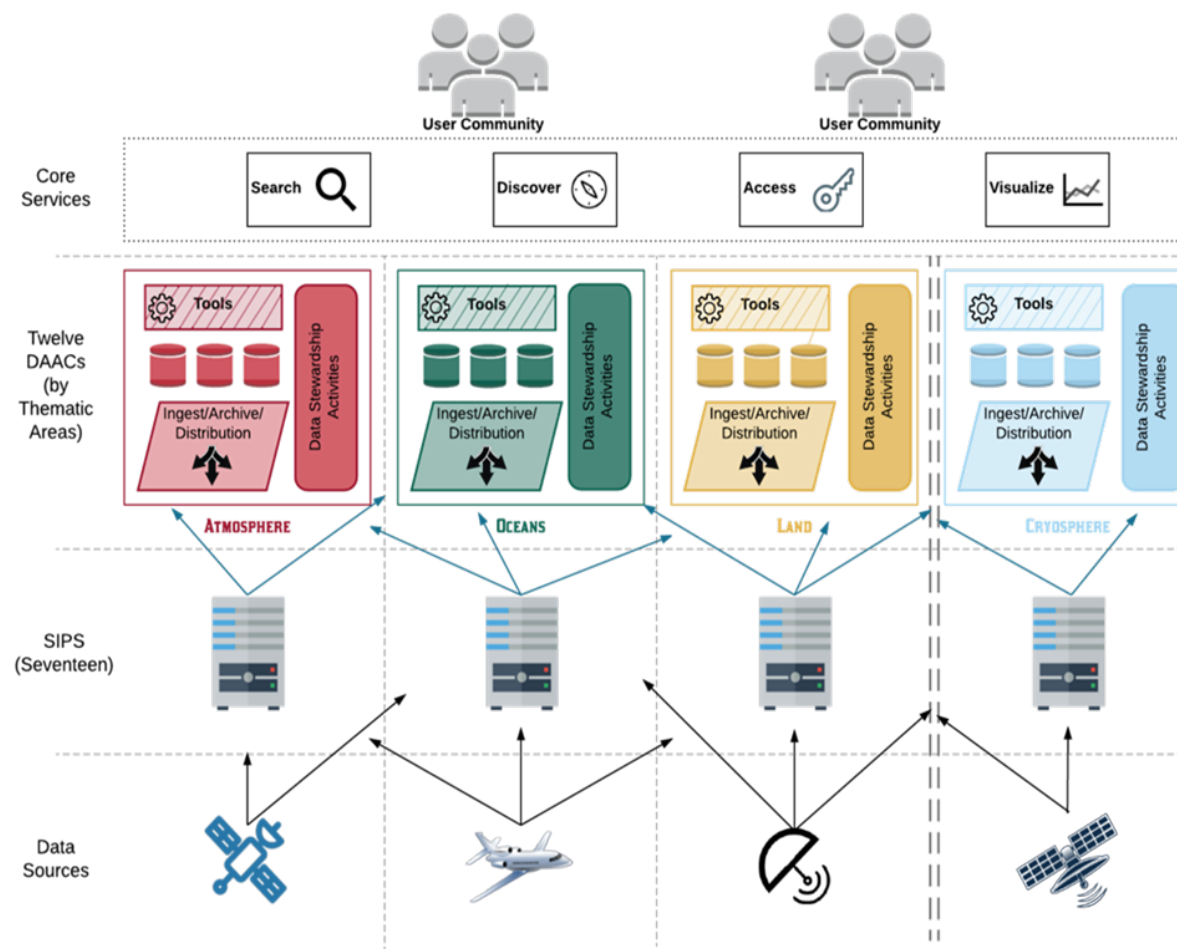


Current Data Infrastructure

Earth Observing System Data and Information System (EOSDIS) provides end-to-end data management capabilities

- *Collect* data by Earth observing remote sensing instruments
- *Process* at Science Investigator-led Processing Systems (SIPS)
- *Archive* at Distributed Active Archive Centers (DAACs)
- *Distribute* to user community through key services

Detailed data documentation and curation is a service that aids in distribution of data by enhancing search and discovery while promoting transparency and scientific reproducibility



Current NASA EOSDIS Architecture

Algorithm Theoretical Basis Documents

Algorithm Theoretical Basis Documents, or ATBDs, provide data users the physical theory, mathematical procedures and assumptions made for developing algorithms which convert radiances received by remote sensing instruments into geophysical quantities

ATBDs are *required* for every NASA Earth Observing System (EOS) instrument product



The Documentation Problem

Organizational challenges:

- No standard template or content requirements
 - An ATBD may address multiple products or a product may be addressed by multiple ATBDs
 - Creates confusion and uncertainty for scientists writing ATBDs compounded by large volume of products and associated authoring science teams
- No central repository for search and discovery
 - Documents are delivered to archival centers for preservation and distribution
 - Important for data distribution velocity and science reproducibility that discovery is consistent and efficient

Technical challenges:

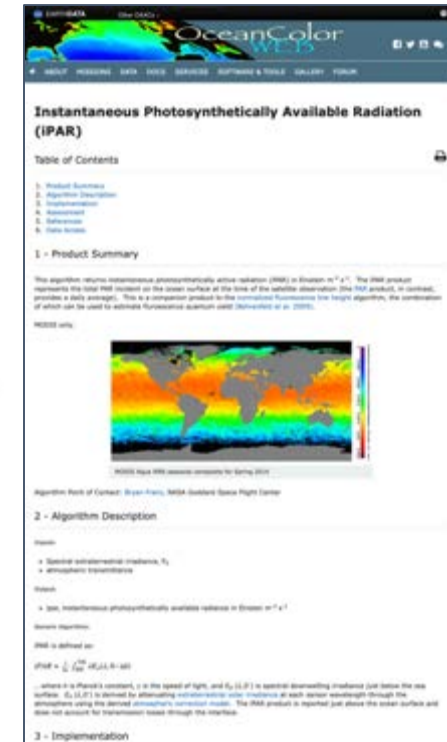
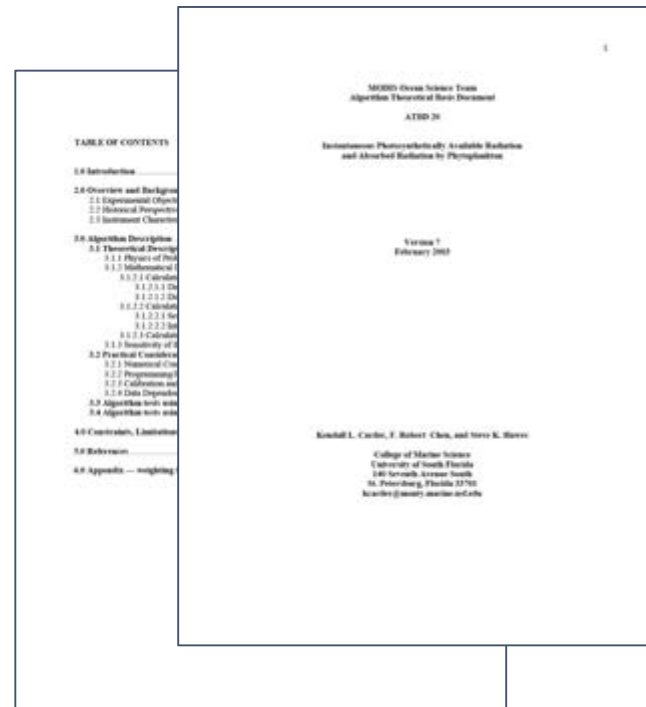
- Difficult to update or maintain
 - Data and associated algorithms may change rapidly
 - Documents must be readily updated for advancements in data processing or when corrections are identified
 - Update process coupled to organizational challenges
- Limited ability for users to efficiently and effectively parse content
 - Often available as PDFs; not search engine optimized

What is APT?

The Algorithm Publication Tool (APT) is a cloud-based tool to

- Standardize and author ATBDs
- Streamline the authoring process
- Enhance search and discovery

APT has the goal of moving from a static to dynamic model of documentation with intelligent connections to software, data and other supporting resources to improve transparency and promote scientific reproducibility



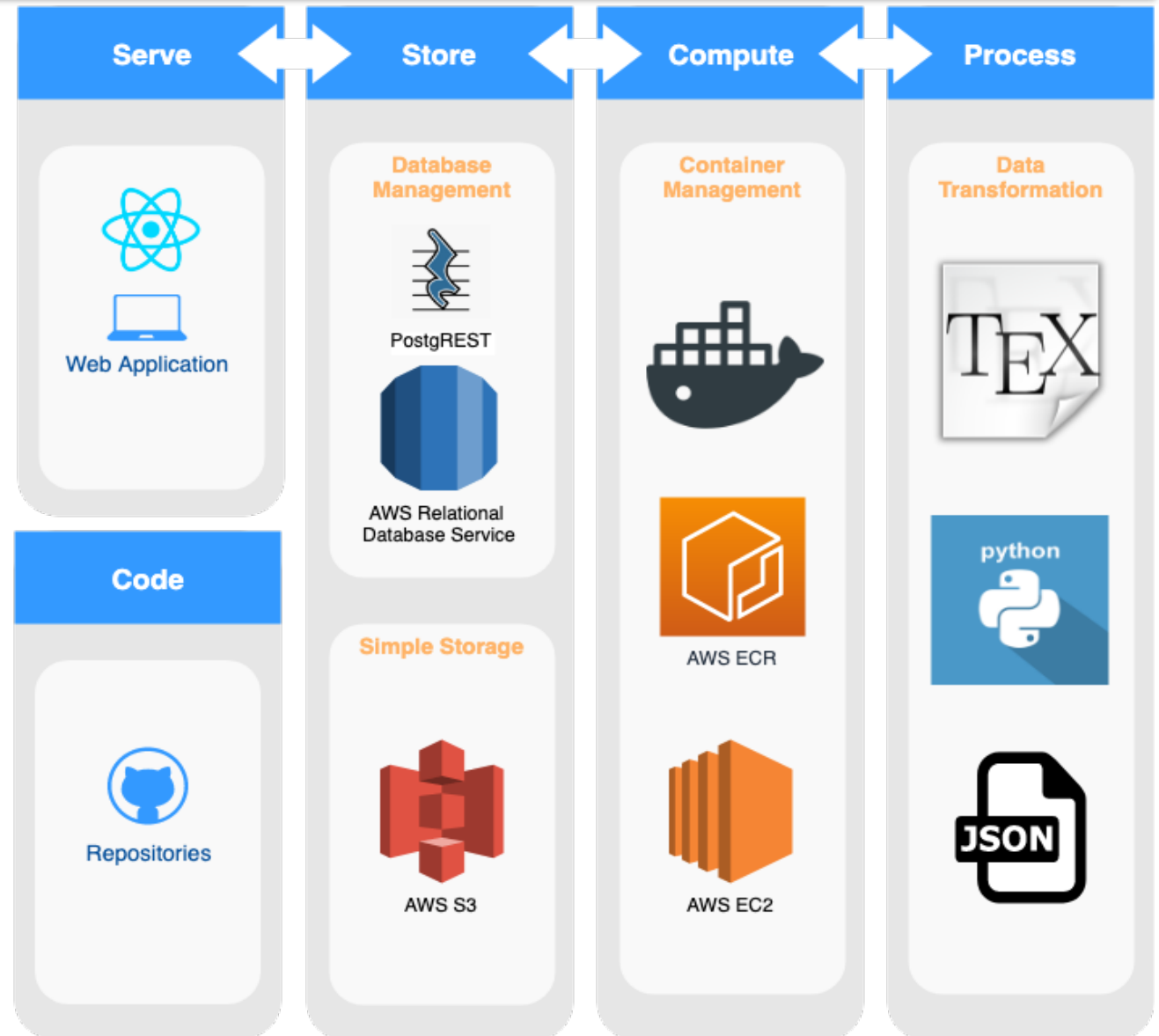
High Level Architecture

Web application front-end which serves, and is served by, a content model schema implemented as a database

Latex backend supports rich content required for scientific writing

Generates PDF and HTML documents

All components implemented using AWS cloud resources



Algorithm Metadata Model

Basis of the tool are rich models for storing ATBD content

- Traditional metadata - information about the document
- ATBD content as metadata - standardize and simplify content needed to describe how data is generated

ATBD information model is based on review of Earth observations community algorithm documentation

Element Name	Element Description	Type	Constraints	Required	Cardinality	Element Used to Build/Create Document?
Algorithm Description						
AlgorithmDescription/AlgorithmInputIntroduction	Provides a brief contextual, introduction for the InputVariables table	String	For prototype: 1024 maximum characters	Yes	1	Yes
AlgorithmDescription/AlgorithmInputVariableName ***[Variable/Name in UMM-Var] See comment column.	The name(s) of the variables that are inputs into the algorithm as they are named in the data. A variable is a named set of data that contains the recorded values of a measurement. A variable can also be the output of a model.	For prototype: String Longer term: Build from UMM-Var	For prototype: 1024 maximum characters	Yes	1..n	Yes

ATBD Publication

APT initial development focused on the development of the content model and publication web interface

- Ease document publication burden

Simplify embedding and generation of rich content

- Equation building with in-line validation

The screenshot displays a web browser window with the URL `nasa-apt-application.s3-website-us-east-1.amazonaws.com/atbdsedit/30/drafts/1/algorithm_description`. The page title is "[Demo Only] MOD20: Instantaneous Photosynthetically Availa...". The document content includes several paragraphs of text and two mathematical equations. The first equation is $E_d(\lambda, 0^+) = E_{dd}(\lambda, 0^-) + E_{ds}(\lambda, 0^+)$. The second equation is $E_d(\lambda, 0^+) = E_{dd}(\lambda, 0^-) + E_{ds}(\lambda, 0^+)$. The interface includes a "Save" button, a "Scientific theory assumptions" section, and a rich text editor with various formatting options.

ATBD Publication

APT initial development focused on the development of the content model and publication web interface

- Ease document publication burden

Simplify embedding and generation of rich content

- Equation building with in-line validation
- UI for table construction

The screenshot shows a web browser window with the URL `nasa-apt-application.s3-website-us-east-1.amazonaws.com/atbdsedit/30/drafts/1/algorithm_description`. The page title is "[Demo Only] MOD20: Instantaneous Photosynthetically Availa...". The main content area is titled "Scientific theory assumptions" and contains the following text: "The Gregg and Carder [1990] model is an extension and simplification of the Bird et al. [1986] model. The first step in the algorithm is to compute the downwelling irradiance just above the sea surface at 1 nm resolution." Below the text are two equations:
$$E_d(\lambda, 0^+) = E_{dd}(\lambda, 0^-) + E_{da}(\lambda, 0^+)$$
 and
$$E_d(\lambda, 0^+) = E_{dd}(\lambda, 0^-) + E_{da}(\lambda, 0^+)$$
. A table construction interface is visible at the bottom, with a table containing two rows: one with "Symbol" and "Long Name" headers, and another with "\$\lambda\$" and "Wavelength" values. A "Save" button is located at the bottom left of the editor.

This close-up view shows the table construction interface. It features a table with two columns: "Symbol" and "Long Name". The first row contains the values "\$\lambda\$" and "Wavelength". Below the table, there is a red-bordered input field for adding a new row. A "Save" button is located at the bottom left of the interface.

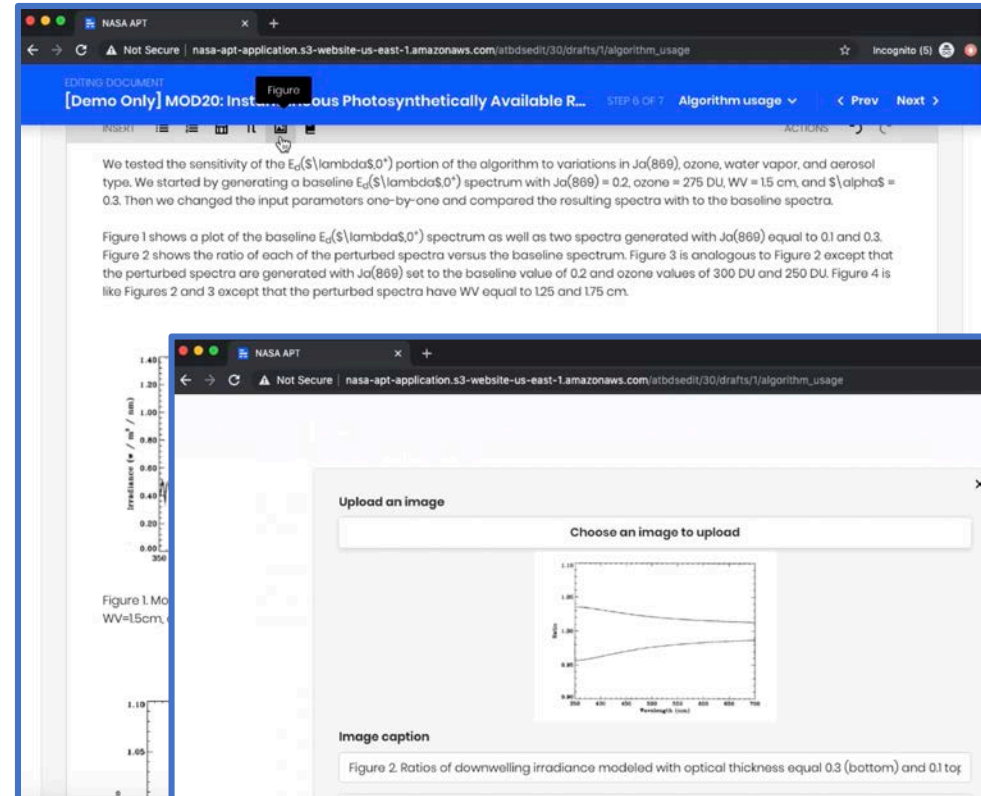
ATBD Publication

APT initial development focused on the development of the content model and publication web interface

- Ease document publication burden

Simplify embedding and generation of rich content

- Equation building with in-line validation
- UI for table construction *and* inserting figures



ATBD Publication

APT initial development focused on the development of the content model and publication web interface

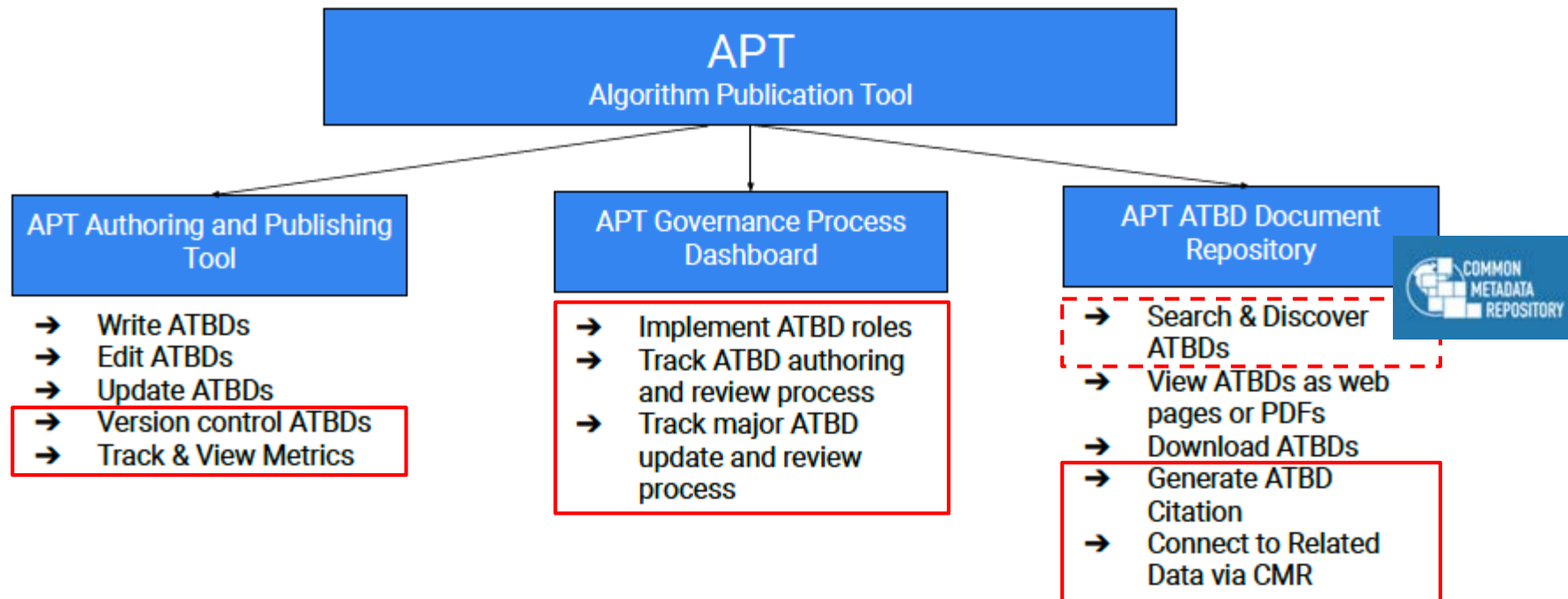
- Ease document publication burden
- Simplify embedding and generation of rich content
- Equation building with in-line validation
- UI for table construction and inserting figures
- Bibtex upload for reference management
- Relational database implementation promotes content reuse and consistency

The screenshot displays the NASA APT web interface. A file manager window is open, showing a list of files including 'atbd_mod20' and several image files. A BibTeX file named 'S0034425718304139.bib' is selected. The BibTeX content is visible in the editor, showing metadata for an article titled 'The Harmonized Landsat and Sentinel-2 surface reflectance data set'. Below the editor, an 'Import references' dialog box is shown, with the selected file name 'S0034425718304139.bib' and a 'Proceed' button. In the background, the main interface shows a form for entering reference details, including fields for Title, Authors, Series, Edition, Volume, Issue, Report Number, Publication Place, Year, Publisher, Pages, ISBN, and DOI.

Long-term Vision

Integrate into NASA EOSDIS data curation and preservation system

- Adoption of APT for future EO mission documentation requirements
- Promote an interconnected and open data ecosystem
 - Interconnected with *input* and *output data* through CMR metadata
 - Leverage future metadata schemas including that for *variables* and *software*

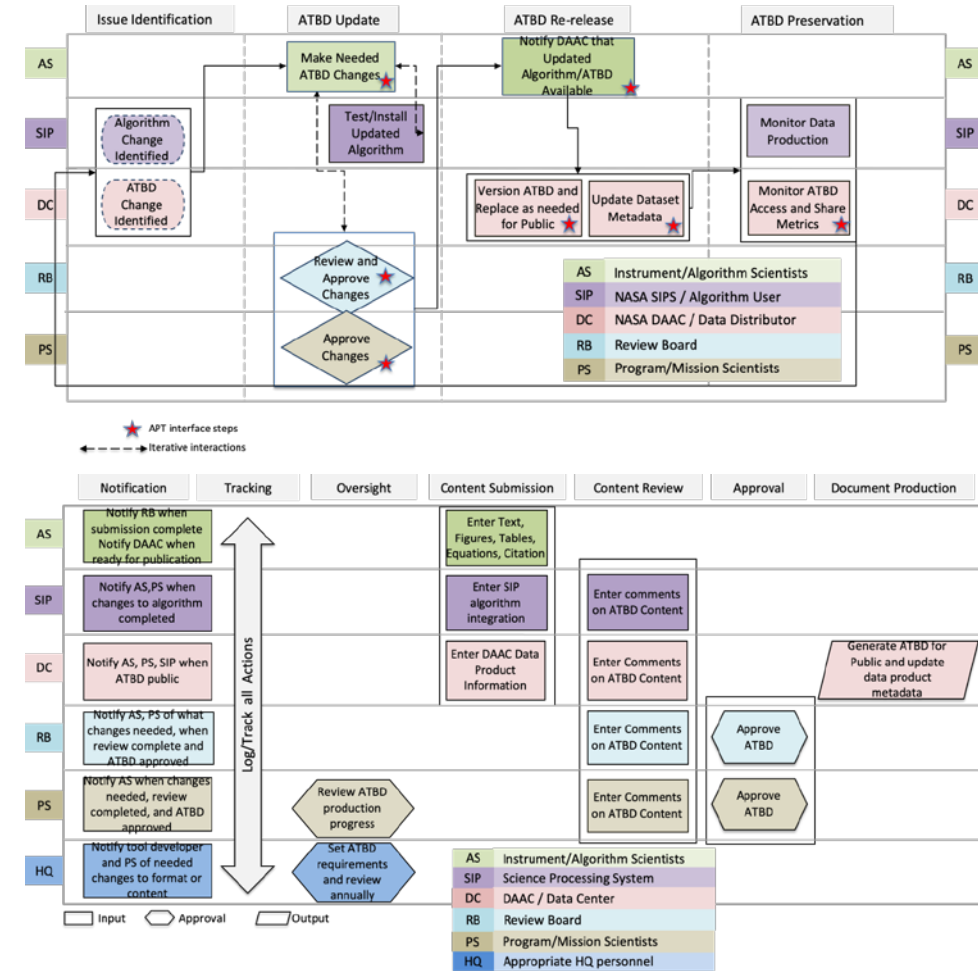


Long-term Vision: Governance

APT will reinforce NASA EOSDIS and data provider responsibilities throughout the data lifecycle providing independent workflows for:

- 1) Initial creation and publication of an ATBD (formulation phase)
- 2) Updates to a ATBD (operations phase)

APT User Interface will include dashboard for tracking/managing all interactions associated with workflows



Sample governance workflow and required interactions

Long-term Vision: Centralized Repository

Single search and distribution service for all NASA Earth Observation ATBDs through integration of existing NASA ATBDs into the APT document repository

Implement interface and capabilities for search and discovery of documents by

- Identifying metadata, such as citation information and keywords *and*
- Document content, such as equations or scientific concepts



The screenshot shows the NASA APT interface. The top navigation bar is blue and contains 'NASA APT' on the left, 'Documents' and 'About' on the right. Below the navigation bar, there is a search bar with 'Demo' entered and a '+ Create' button. The main content area shows 'Showing 1 result for Demo'. Below this, there is a table with columns for STATUS, TITLE, and AUTHORS. The table contains one row with the following data:

STATUS	TITLE	AUTHORS
Draft	[Demo Only] MOD20: Instantaneous Photosynthetically Available Radiation and Absorbed Radiation by Phytoplankton	Aaron Kaulfus

At the end of the row, there are links for 'PDF', 'HTML', and 'Preview'.

Discussion

Versioning for dynamic documents is challenging

- Must decide what changes constitutes a new version of an ATBD
- How are persistent identifiers (document DOI) impacted?
- All versions must be archived and made available for transparency

For repository completeness, existing archived ATBDs should be incorporated

- Significant effort to rewrite all of these documents
- Can all ATBDs and their versions be accounted for?

Usability and simplicity is required for success

- The tool should minimize resistance to adoption of a single typesetting service
- The tool must accommodate all scientific writing needs
- A standard governance plan should not overly/unnecessarily burdensome

Value to science users is maximized when integrated with existing EOSDIS meta(data) resources

- Need a flexible, future-proof tool to fit an evolving documentation vision and supporting suite of tools



Questions?

ak0033@uah.edu