Improving Algorithm Communication and Data Cognizance Through Standardizing Documentation

Aaron Kaulfus[1], Kaylin Bugbee[1], Alyssa Harris[2], Rahul Ramachandran[3], Sean Harkins[2], Sean Bailey[4], Aimee Barciauskas[4]

(1) University of Alabama in Huntsville
(2) Development Seed
(3) NASA Marshall Space Flight Center
(4) NASA Goddard Space Flight Center

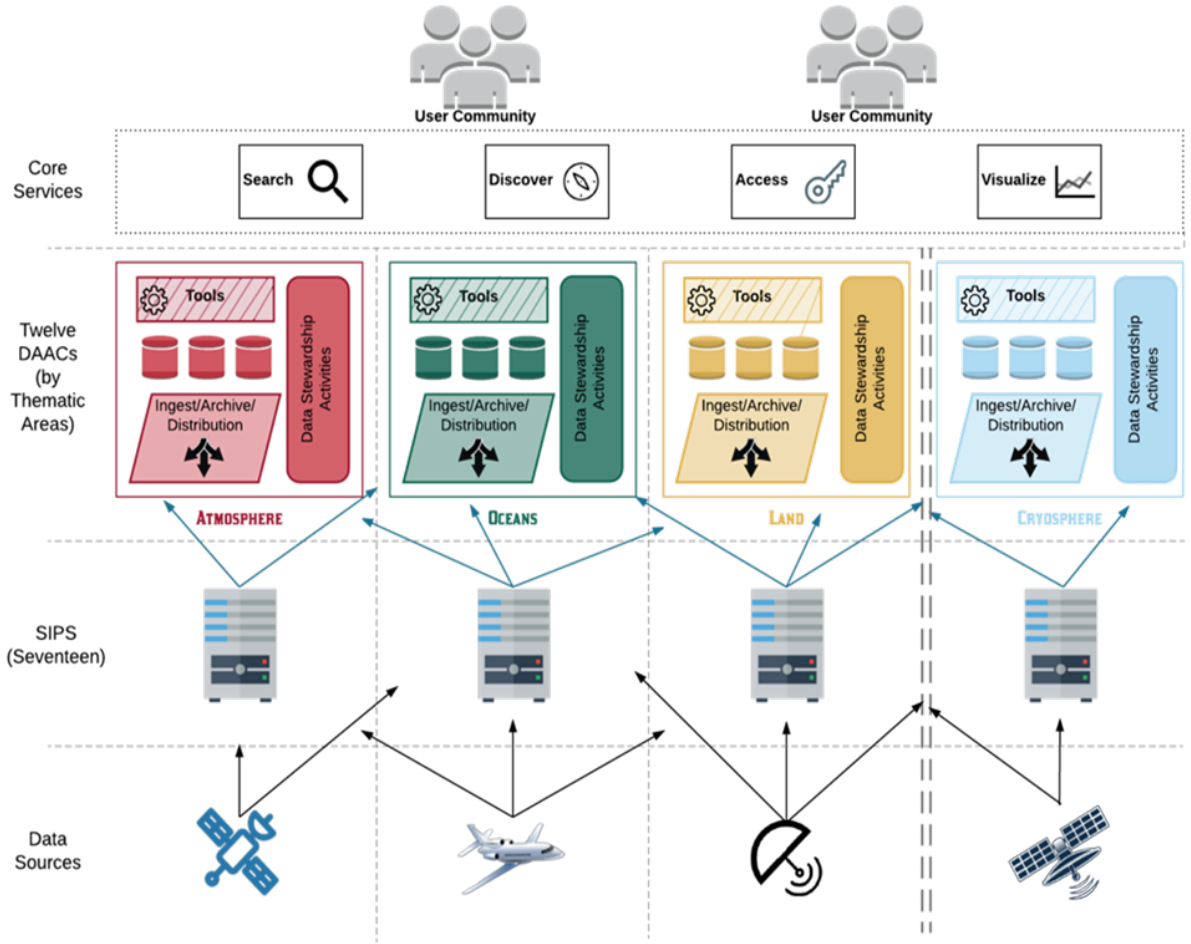
2020 American Meteorological Society Annual Meeting. January 12-16. Boston, MA

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:

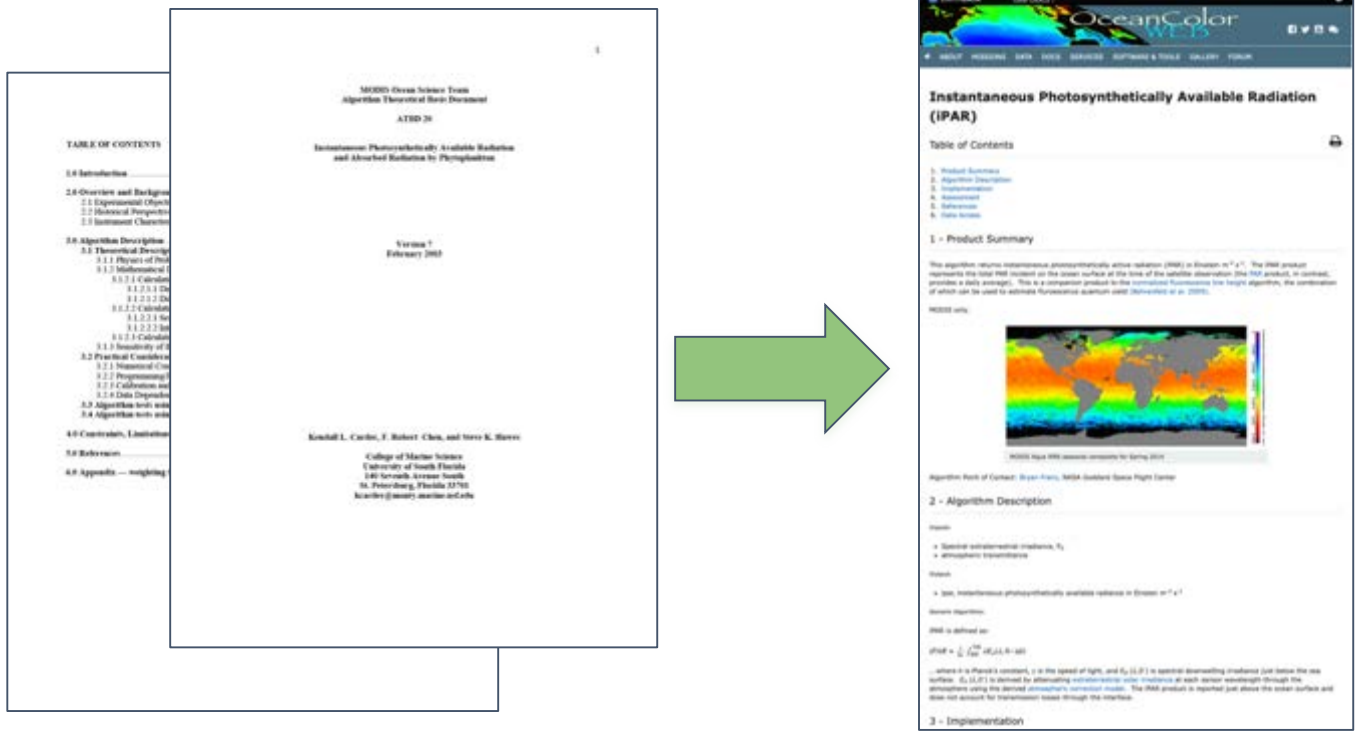
- Difficult to update or maintain
  - Data and associated algorithms may change rapidly
  - Documents must be readily updated for advancements in data processing or when corrections are identified
  - Update process coupled to organizational challenges
- Limited ability for users to efficiently and effectively 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



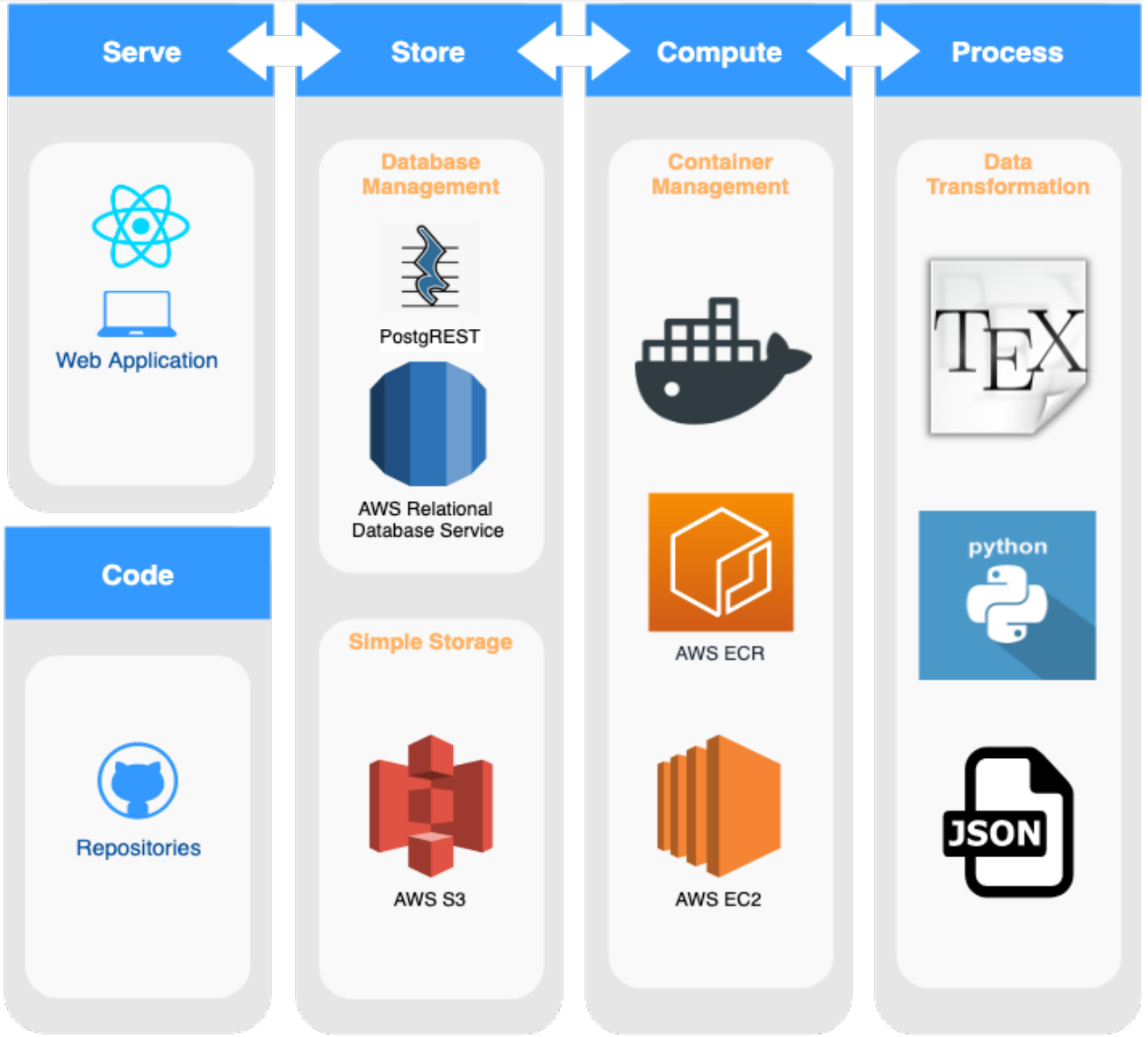
High Level Architecture

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

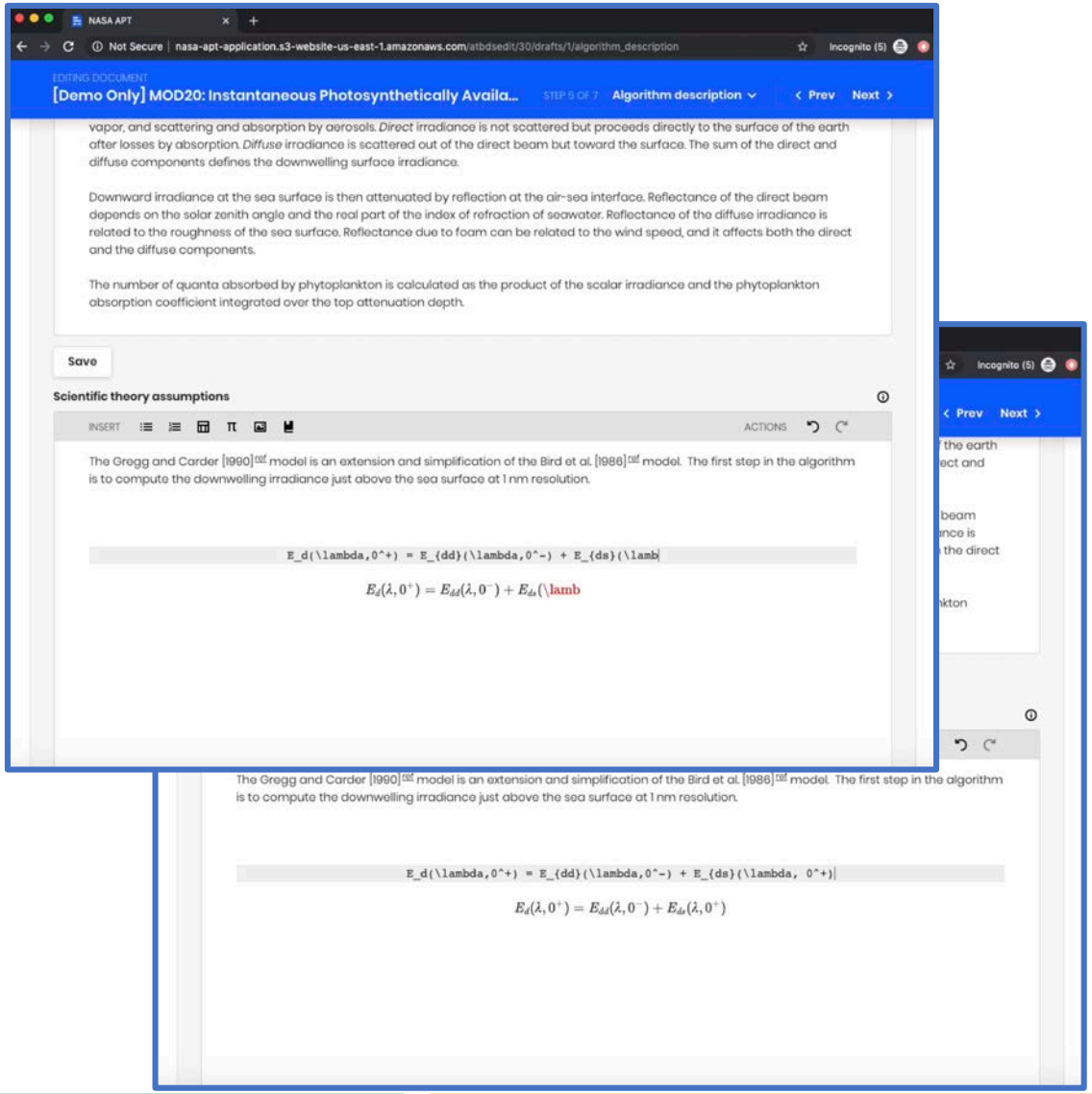
ATBD Publication

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



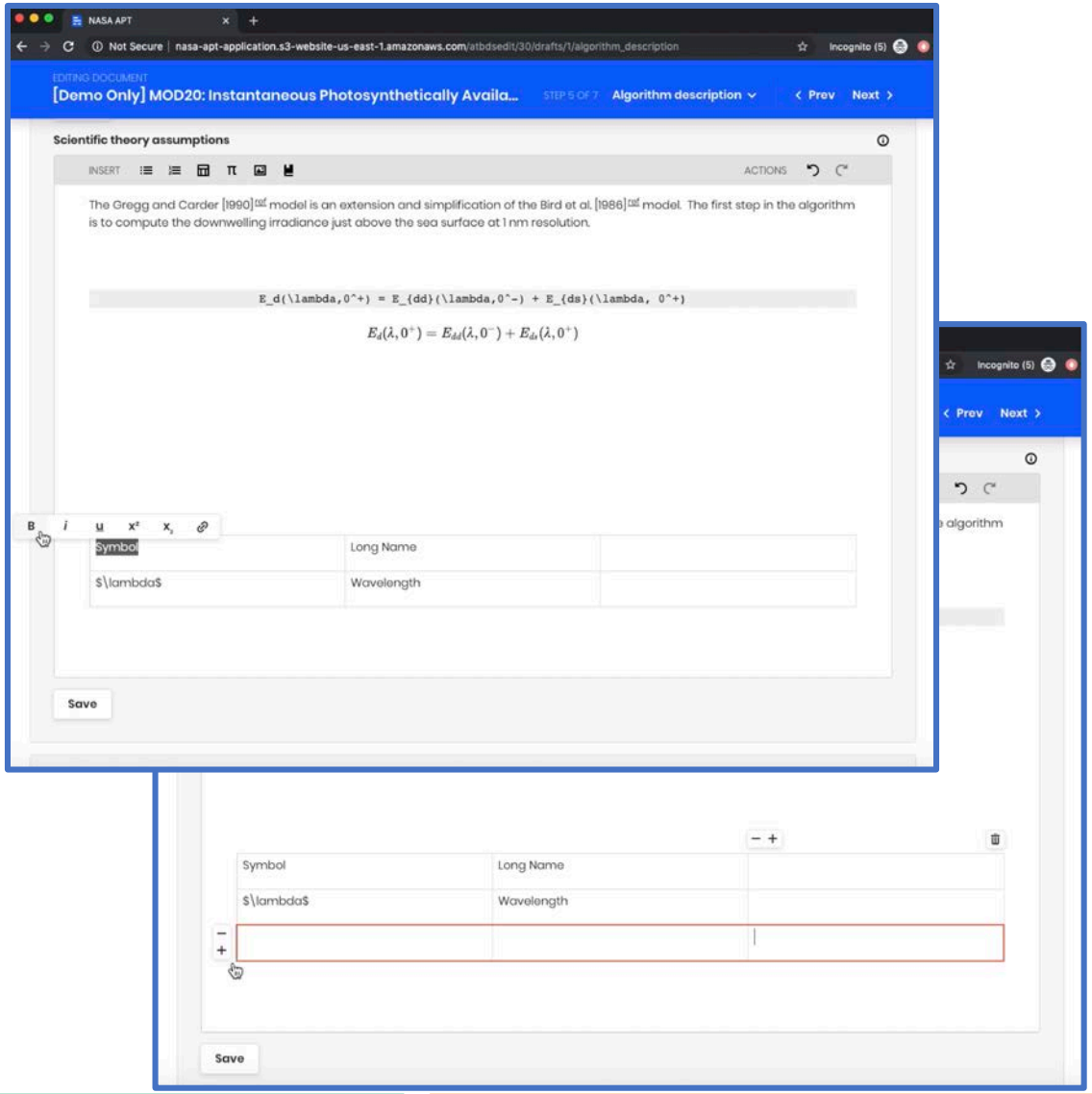
ATBD Publication

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



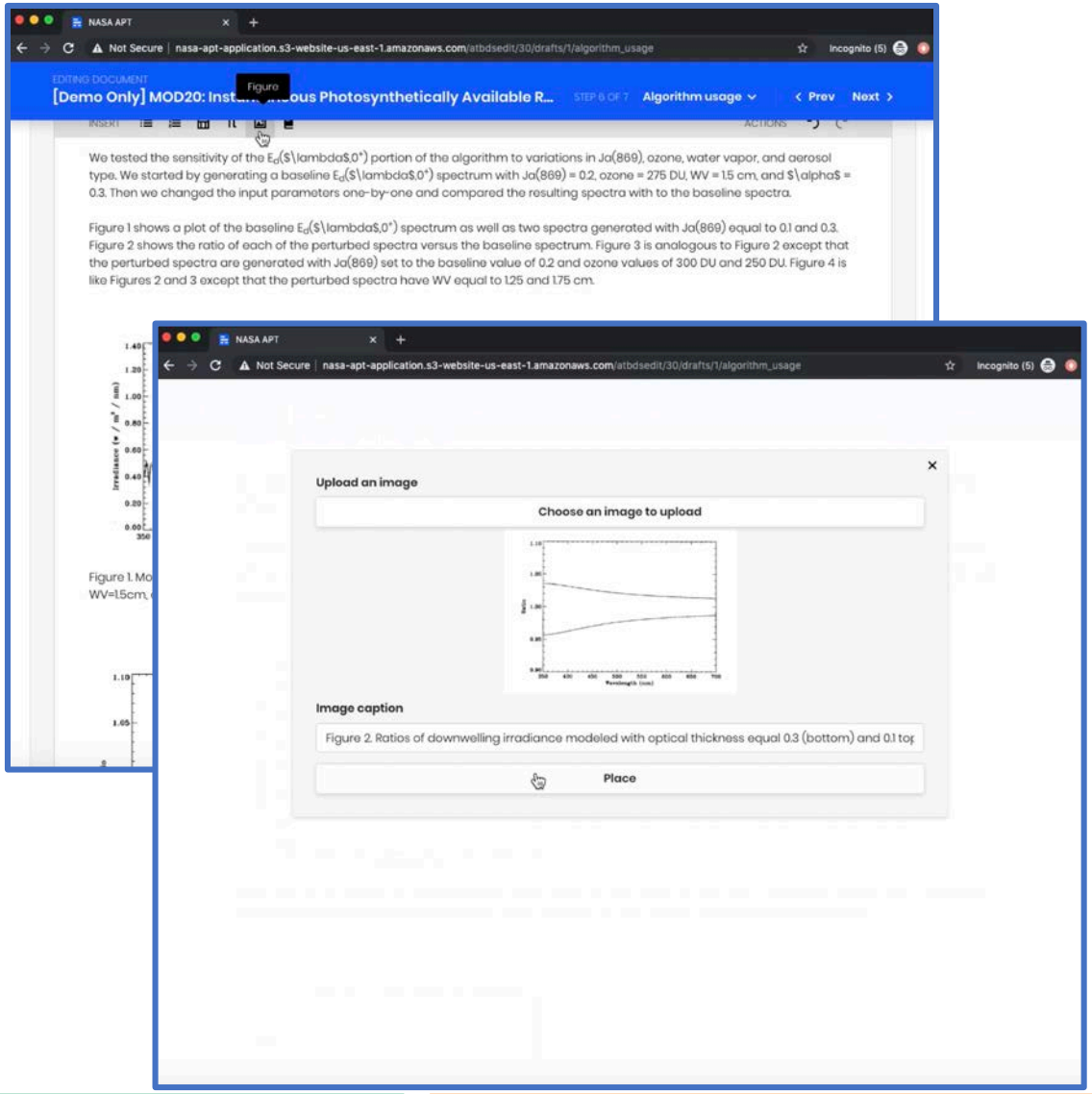
ATBD Publication

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



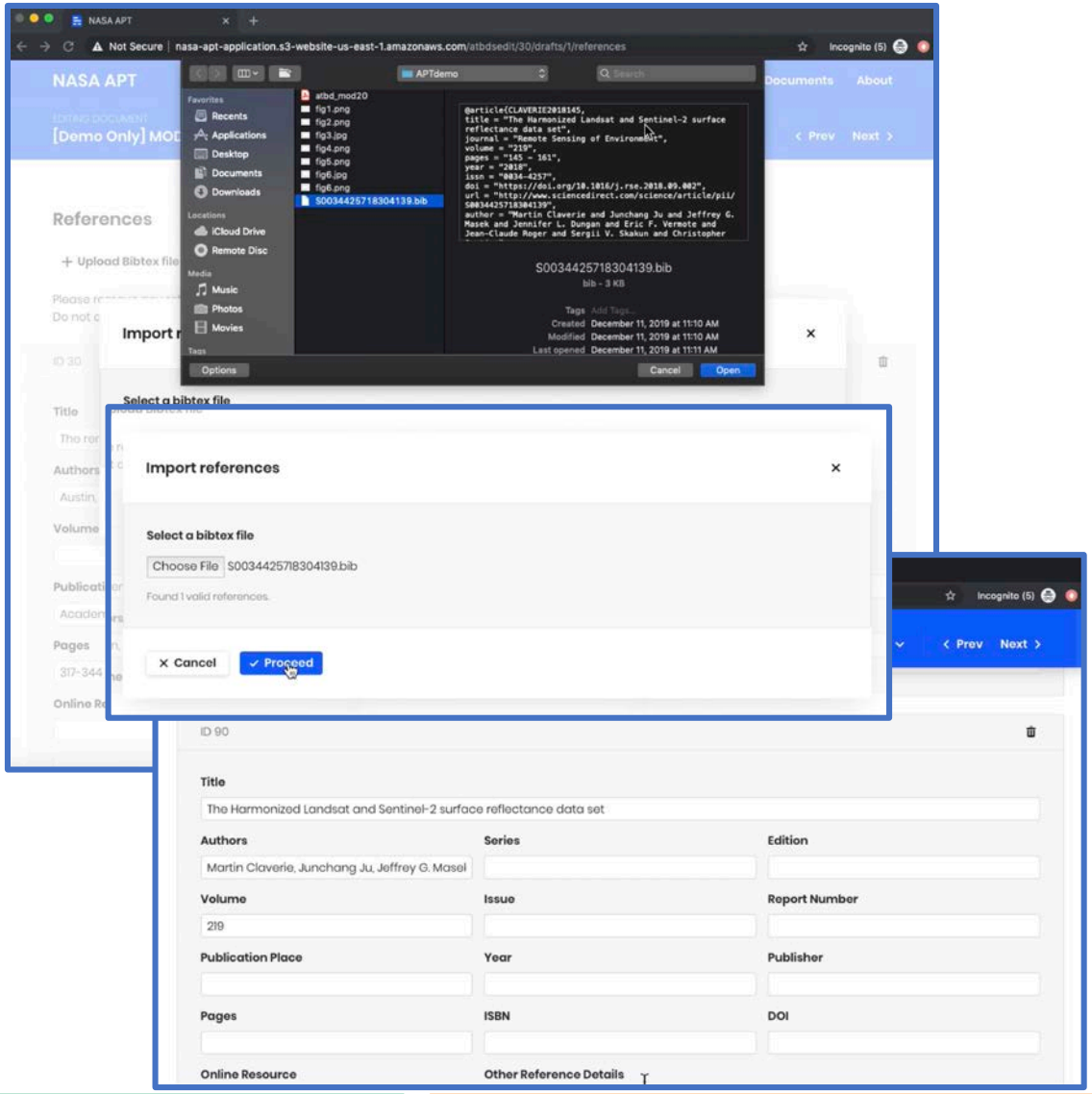
ATBD Publication

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

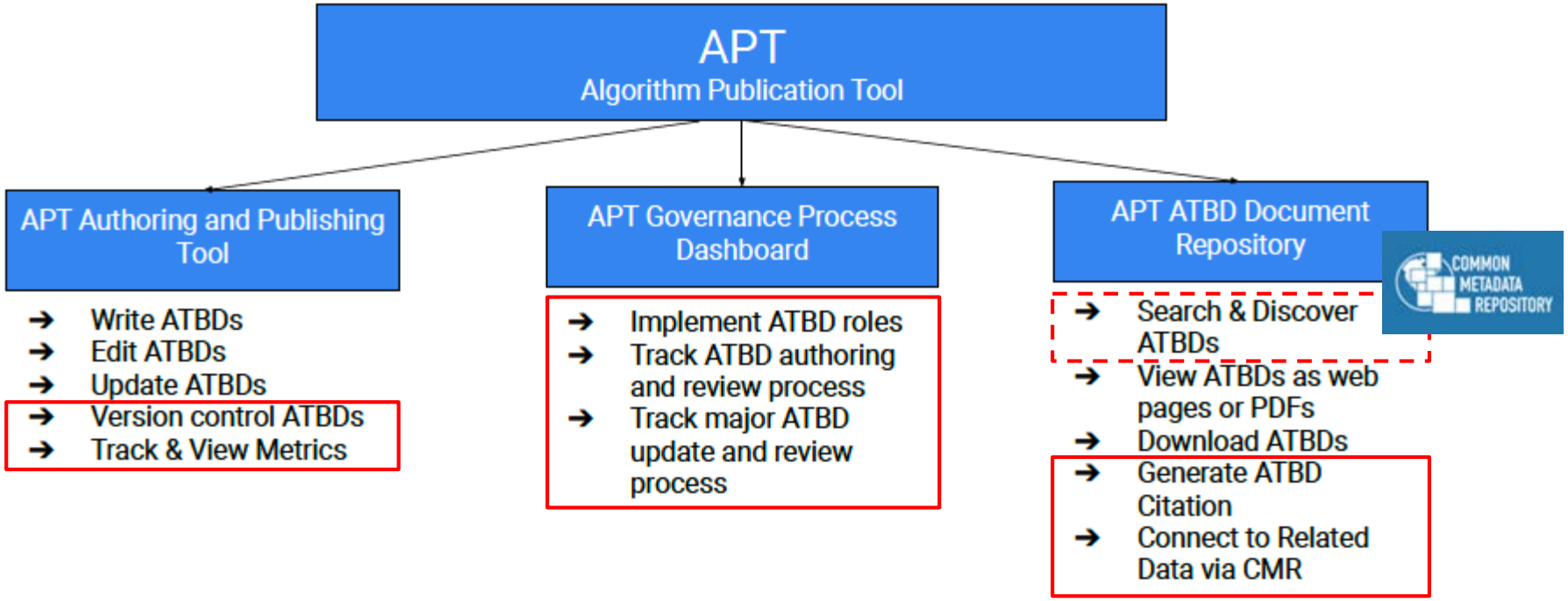
- Equation building with in-line validation
- UI for table construction and inserting figures
- Bibtex upload for reference management
- Relational database implementation promotes content reuse and consistency



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*

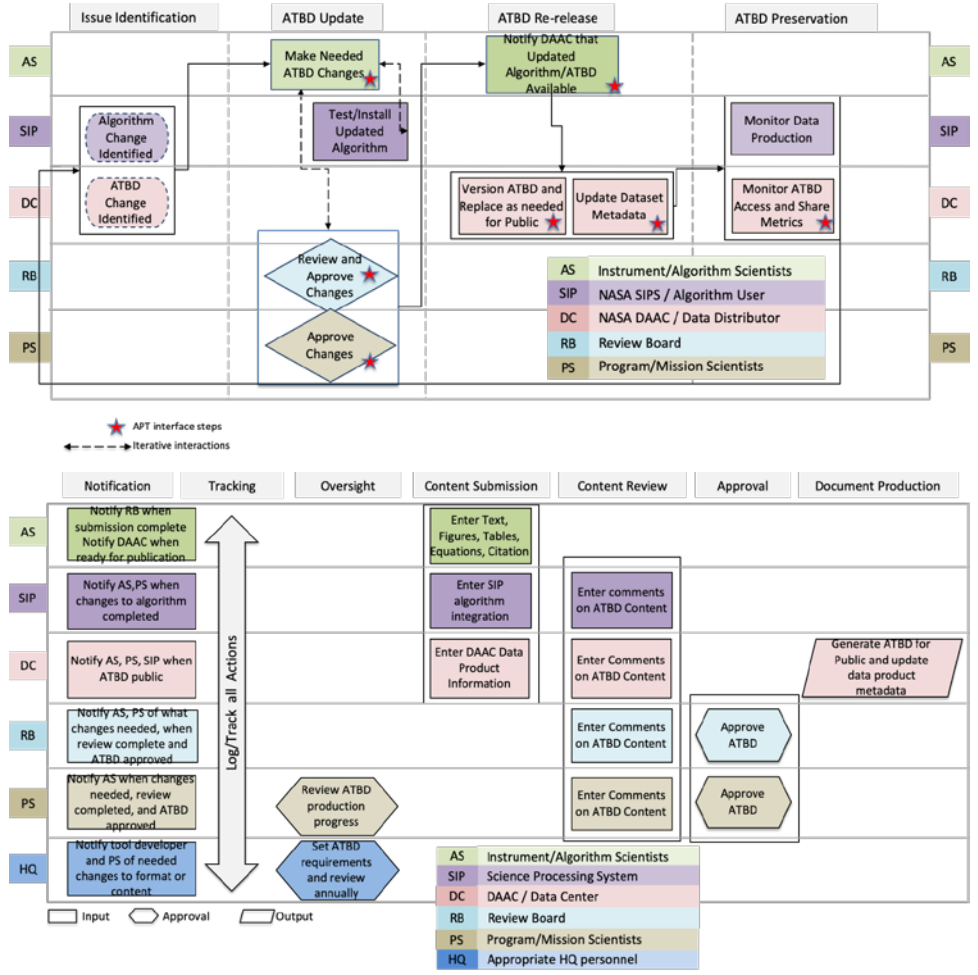


Long-term Vision: Governance

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



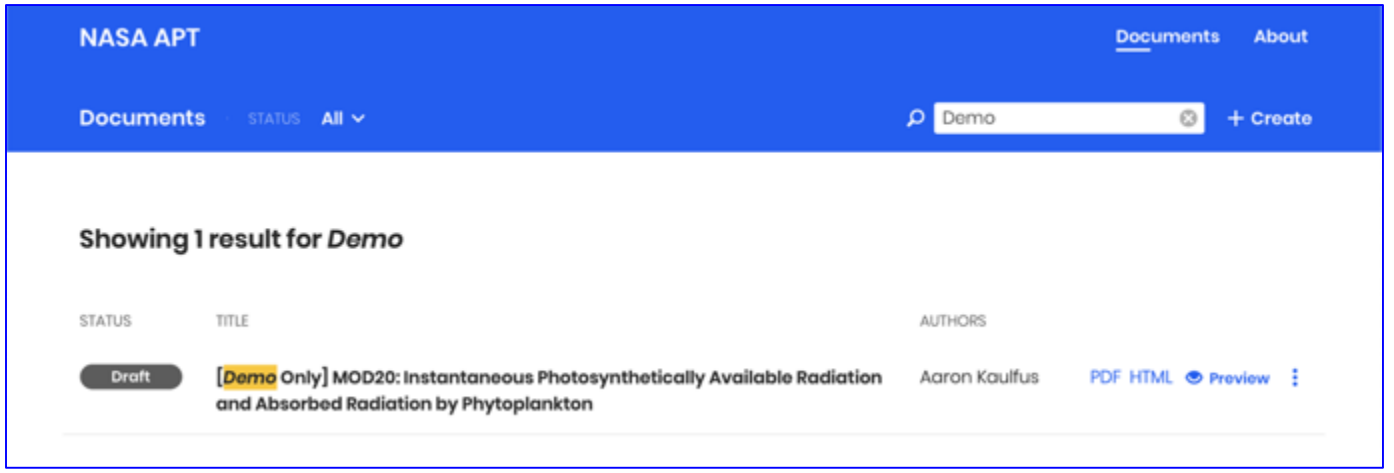
Sample governance workflow and required interactions

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



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