# STANDARDIZED ALGORITHM DOCUMENTATION FOR IMPROVED SCIENTIFIC DATA UNDERSTANDING: THE ALGORITHM PUBLICATION TOOL PROTOTYPE

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

[1]University of Alabama in Huntsville
[2] DevelopmentSeed
[3] NASA Goddard Space Flight Center
[4] NASA Marshall Space Flight Center

## ABSTRACT

Algorithm Theoretical Basis Documents (ATBDs) are documents which accompany Earth observation data products generated from algorithms. While ATBDs are essential to scientific reproducibility, these key documents are not standardized and are often difficult to find. In this paper, we present the prototype Algorithm Publication Tool (APT), a cloud-based ATBD authoring and editing tool for NASA's Earth science data systems. A standardized ATBD information model is also described as well as lessons learned from developing the prototype tool.

***Index Terms***— Data stewardship, cloud technologies, metadata, algorithm documentation, reproducibility

## 1. INTRODUCTION

Algorithm Theoretical Basis Documents (ATBDs) are documents which accompany Earth observation data products generated from algorithms. Raw data from a remote sensing instrument requires some processing in order to provide data in standardized sensor units or geophysical variables. ATBDs describe the physical theory, mathematical procedures and assumptions made for the algorithms that convert the raw data received by remote sensing instruments into geophysical quantities or measurements.

ATBDs are of importance to the scientific community for a number of reasons. First, ATBDs are key to ensuring scientific reproducibility and repeatability by allowing scientists to examine the fundamental scientific theory and assumptions made when creating higher level data products. ATBDs, when accompanied by source code and raw data inputs, also make it possible to reproduce a data product. As concerns continue to grow about the reproducibility crisis in scientific research [1], ATBDs ensure that the methods used for generating key data products are transparent and also provide clarity on the provenance of output data products. Second, ATBDs are necessary to understand whether a data product is fit for use for a given research question. As a wide range of users begin leveraging Earth observation data for a variety of applications, it has become increasingly important that users understand how the data was generated and whether the use of the data is appropriate for a given use case.

While ATBDs are clearly important to the scientific community, writing these documents in the past has been cumbersome for a number of reasons. First, a standardized ATBD template does not exist and guidance on what information is required within an ATBD is minimal. This leads to a number of content issues including the duplication of information within the document itself or the repetition of information provided elsewhere in data or instrument documentation. Lack of guidance leads the community to one-off sharing of previously written ATBDs that may or may not be of a given quality in order to meet the documentation requirement. Second, ATBDs have typically been written manually in a static format on a local machine. Since ATBDs are often written by multiple authors within a science team, collaboration on a local document can be challenging. Lastly, since some authors view the ATBD as a 'once and done' task, the ATBDs are often not updated despite being associated with living and dynamic data and code that may be periodically updated or improved.

Once an ATBD is written and made publicly available, finding these key documents is often difficult. Links to the ATBD are sometimes provided in the collection level metadata of the output data products but to date this practice has not been universally adopted and, where it has been adopted, links are often not updated as needed. Additionally, ATBDs themselves are not available in a centralized location for discovery. Since ATBDs are not currently provided in machine readable formats, the content within the ATBDs is also not discoverable.

Clearly, there are technical, organizational, social and informational issues and needs behind many of these ATBD challenges. There is a need to provide consistent, standardized, human-readable documents that provide insights to scientists and users while also providing machine-readable documents that are discoverable and available for building tools and services. There is also a need for a consistent governance process within NASA's Earth science division and data system on the creation and maintenance of these essential documents.

In this paper, we present our efforts to address many of these needs by developing the prototype Algorithm Publication Tool (APT) for NASA's Earth science data systems. The APT represents two new contributions to the Earth observation community. These contributions include the development of a standardized information model for ATBDs and the creation of a cloud-based ATBD authoring and editing tool built on the model. Lessons learned from prototype development and future work for the project, including a more formalized governance process, are also described.

## 2. ATBD STANDARDIZED INFORMATION MODEL

In order to build the APT, a standardized ATBD template and information model was developed to improve communication with ATBD authors by setting clear expectations on what information is required within the documents. The standard information model was used for database implementation in the APT. The APT information model approach was informed by both the Journal Article Tag Suite [2] and the work being done on machine actionable data management plans [3].

To begin standardizing the ATBD document template and the underlying information model, a community review of ATBDs throughout the Earth observation community was conducted. This review included documents from NASA, ESA, NOAA and NEON. The 'Algorithm Description' documents, created by the Ocean Biology Processing Group (OBPG) [4], were also reviewed to inform both the information model and the design of final ATBD HTML pages. Once the review was completed, a core set of information was identified that is common to all ATBDs. Examples of core information include authors, contact information, version number and title.

The community review also revealed a need to accommodate different types of ATBDs while still maintaining a certain level of standardization. To accommodate the different information needs, extensions of the model have been developed. The primary model extension developed to date for ATBDs is geophysical variables with other extensions, such as instrument ATBDs, planned.

The information model also leverages reusable component metadata [5] in order to ensure consistency within the metadata and to reduce the amount of information input required by authors. Reusable component metadata within APT currently includes citation information, author information and scientific equations.

In the end, capturing information in a structured model means that information is provided in a machine-readable form that enables interoperability between tools and services [3]. These include a tool for authoring and updating ATBDs as well as a tool for displaying and discovering ATBDs. Longer term, an ATBD structured information model also makes establishing connections between data and code metadata possible. These connections are desirable for scientific reproducibility.

## 3. APT PROTOTYPE IMPLEMENTATION

The prototype APT is an ATBD authoring and editing service built off of the ATBD information model using cloud-based components (Figure 1). The tool streamlines the ATBD creation process and makes it easier to update existing ATBDs.

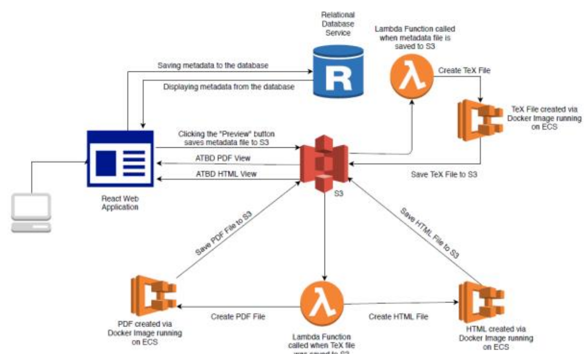


Fig. 1: The prototype APT cloud-based components and associated workflows for generating ATBDs. The APT back end workflow leverages several AWS services including Lambda functions, the relational database service, S3 and Elastic Container services.

The prototype front end web application is built using React, a Javascript library for creating user interfaces, and allows authors to write new ATBDs, to edit existing ATBDs and to search for ATBDs that are in the process of being written.

When creating a new ATBD, the APT steps the user through seven document sections including document identifying information, algorithm description, references and more. For scientific text, the APT uses KaTeX to render visually appealing mathematical equations synchronously. The APT also uses slate.js for writing rich text. Together, the KaTeX and slate.js libraries make it possible to write the rich text needed for scientific documents. The APT authoring tool also includes a reference manager that allows the user to upload a Bibtex file containing a list of relevant publications.

Once a draft of an ATBD is completed, the front end application features document previews as either an HTML page or a PDF document. When the 'preview' button is selected in the application, an AWS Lambda function generates a TeX file of the document via a Docker image on an EC2 instance and is saved in S3. Once the TeX file is created and saved, another AWS Lambda function creates either the HTML or PDF file via a Docker image on an EC2 instance depending on the user's output format selection. The file is then stored on S3 where it is available for viewing via the front end web application (Figure 2).

**Direct Irradiance - $\mathbf{E_{dd}(\lambda, 0^+)}$**

$E_{dd}(\lambda, 0^+)$ is computed by

$$E_{dd}(\lambda, 0^+) = F_0(\lambda) cos(\theta) T_r(\lambda) T_{oz}(\lambda) T_o(\lambda) T_w(\lambda) T_a(\lambda) \tag{3}$$

where $F_0(\lambda)$ is the mean extraterrestrial irradiance corrected for earth-sun distance and orbital eccentricity, $\theta$ is solar zenith angle, and $T$ is the transmittance after absorption andor scattering by each atmospheric component. The components $r$, $oz$, $o$, $w$, and $a$ represent Rayleigh scattering, ozone, other gases, water vapor, and aerosols, respectively.*Extraterrestrial solar irradiance* The mean extraterrestrial solar irradiance, $H_0(\lambda)$, is taken from the revised Neckel and Labs [1984] data for the wavelength range of 330 to 700 nm. The extraterrestrial solar irradiance corrected for earth-sun distance is given by Gordon et al. [1983][5] as

$$F_0(\lambda) = H_0(\lambda)[1 + ecc * cos[2 * \pi(JD - 3)/365]]^2 \tag{4}$$

**Fig. 2:** A sample of a portion of an ATBD generated by the APT including rich text and mathematical equations.

On the back end, the ATBD content is stored as metadata in a PostgREST instance of an Amazon Relational Database Service (RDS). The previously described information model serves as the schema for RDS. One advantage of this database implementation is the reuse of content, such as author information, within and across documents which reduces duplication of information while simplifying the writing process for authors.

## 4. NEXT STEPS

Next steps for the APT include providing a centralized discovery repository for ATBDs, completing the development of the authoring tool and establishing and implementing the ATBD governance process. The development of a centralized ATBD discovery repository will primarily focus on search, access and download capabilities through both a search interface and an API. Completion of the authoring tool will focus on dynamically generating ATBD documents using an API, providing schema.org markup for ATBD pages, implementing new information model extensions to support other ATBD types and enhancing existing features. Developing and implementing the ATBD governance process will be a significant effort requiring technical, social and organizational solutions. ATBDs are drafted by algorithm scientists and then typically undergo a final approval process before the document is made publicly available. The APT team will work with stakeholders to formalize an agreed upon approval workflow and associated swim lanes for all those involved in the ATBD process. The APT will also feature a dashboard within the authoring tool to track the ATBD authoring and review process. In the end, the APT should be integrated into the workflows of those developing algorithms and into NASA's data stewardship systems to support reproducibility.

## 5. LESSONS LEARNED

Designing a user-friendly solution for writing rich text scientific documents required careful thought and investigation of various viable solutions. The APT team considered the needs of an entire community with various experiences writing technical documents and balanced those needs with project requirements such as using open source technologies. For example, some viable typesetting libraries and rich text editors are not open sourced which eliminated those options from implementation in APT.

The lack of governance and policy requirements also delayed the implementation of some key features. For example, at this time, a versioning scheme for ATBDs has not yet been determined. Establishing a versioning convention for updates to ATBDs is unprecedented. The range of modifications to ATBDs, from minor text edits to major algorithmic modifications are similar to but do not map exactly to modifications to software libraries. There is a need to gain consensus on ATBD versioning principles which will in turn determine how much of the versioning process can be managed by the tool. These decisions on versioning will affect the governance process for assigned unique identifiers as well.

## 6. REFERENCES

[1] 1,500 Scientists Lift the Lid On Reproducibility. Nature. https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970. Accessed 1/8/2020

[2] Introduction to JATS (Journal Article Tag Suite). XML.com. https://www.xml.com/articles/2018/10/12/introduction-jats/. Accessed 01/08/2020

[3] T. Miksa, S. Simms, D. Mietchen and S. Jones. "Ten Principles for Machine-Actionable Data Management Plans," *PLOS Computational Biology*, vol. 15, no. 3, 2019. https://doi.org/10.1371/journal.pcbi.1006750

[4] Algorithm Descriptions. Ocean Color Web. https://oceancolor.gsfc.nasa.gov/atbd/. Accessed 01/08/2020

[5] T. Habermann. "Metadata Life Cycles, Use Cases and Hierarchies," *Geosciences*, vol. 8, article no. 179, 2018. https://doi.org/10.3390/geosciences8050179