A Strategy for Managing NASA's Long Tail of Planetary Research Data

Insights from the Development of the Astrobiology and Habitable Environments Database (AHED)

Lead:

Thomas Bristow, NASA Ames Research Center
thomas.f.bristow@nasa.gov, 650.604.4665

Co- investigators:

Barbara Lafuente, SETI Institute
Nate Stone, Open Data Repository
Mary Parenteau, NASA Ames Research Center
Shawn R. Wolfe, NASA Ames Research Center
Sara Perez Rojo, NASA Ames Research Center
Kevin Boydstun, NASA Ames Research Center
Robert Downs, University of Arizona
David Blake, NASA Ames Research Center
Linda Jahnke, NASA Ames Research Center
David Des Marais, NASA Ames Research Center
Christopher Dateo, NASA Ames Research Center
Mark Fonda, NASA Ames Research Center

Endorsers:

Kennda Lynch, Lunar and Planetary Institute|USRA
Leslie Bebout, NASA Ames Research Center
Svetlana Shkolyar, USRA|NASA
Valerie Tu, Jacobs
Jennifer G. Blank, NASA ARC | Blue Marble Space
Tom McCollom, LASP, University of Colorado, Boulder
Roger Summons, Massachusetts Institute of technology
Shaunna M. Morrison, Carnegie Institution for Science
David Summers, NASA Ames Research Center
Xiaogang Ma, University of Idaho
Jolyon Ralph, Hudson Institute of Mineralogy
Robert M. Hazen, Carnegie Institution for Science

Co-signer sign-up spreadsheet:
https://docs.google.com/spreadsheets/d/1-l2JFkNgmiPurhzVxjwipk-iT-98QUzBKZATkEbA7I0/edit?usp=sharing

Introduction

The importance of sharing scientific data is increasingly recognized by the general public, scientists, publishers, as well as the NGOs and government agencies that direct research worldwide[1]. NASA-funded scientists work in collaborative and interdisciplinary research projects. Sharing and mining of data is an integral part of their workflow. Over half of new science sponsored by NASA's Science Mission Directorate (SMD) is sourced from data archives. A number that is set to grow[2]. Efforts to improve the accessibility and discoverability of NASA data are important in empowering traditionally disadvantaged countries and people to get involved and contribute to NASA science[1]. As a result, policies and mandates that require public data archiving of NASA data have been implemented.

Despite the benefits, changing work practices and policies, significant barriers to data archiving and sharing remain, including lack of acknowledgment, time, money, guidance, expertise and trust in available platforms[3]. These challenges are disproportionately felt by the 'long tail' of research in planetary science performed by individual PIs, and small research teams. The long tail often lacks data management resources available to larger groups and missions, and tend to collect heterogenous datasets (a variety of formats stemming from multiple analytical techniques, coupled with contextual information about samples and field areas) needed to pursue science questions, not necessarily suited to large-scale homogenous repositories (e.g. GenBank). This is compounded by growing user expectations in terms data accessibility and ease of use. The pool of users, and the way data is used is also becoming more diverse[4]. With an increasing array of analytical techniques and volume of data being collected by PI-led NASA-funded research in planetary sciences, strategies for streamlining the management, preservation and utilization of this data are needed to optimize the scientific return from NASA's past and ongoing research programs.

The Strategic Data Management Working group (SDMWG) recently outlined visions and goals to NASA SMD in the final report[5] 'Strategy for Data Management and Computing for Groundbreaking Science 2019-2024.' The report makes a broad set of recommendations – many of which are relevant to the long tail of planetary research. *The purpose of this white paper is to detail how insights gained during the development of the* ***Astrobiology and Habitable Environments Database (AHED – Fig. 1)*** *can be applied to other NASA-funded scientific disciplines where long tail research is performed. We make recommendations for the implementation of SMD's data management vision and goals with the aim of supporting NASA's planetary science over the next decade.*

Summary of Recommendations

1) To optimize scientific returns, NASA should guide and invest in exploration and expansion of data management approaches in planetary science. These approaches should be compatible with relevant efforts from the international science community.
2) Metadata is foundational element of the type of open science ecosystem envisioned by the SDMWG. NASA's Planetary Data System (PDS) has developed an internationally recognized standard for archiving planetary science data (PDS4). PDS4 is focused around targets, missions, and specific planetary instrument. We see an opportunity for NASA, with guidance from the scientific community, to extend and publicize metadata standards describing resources relevant to areas of science it intends to pursue in the next decade.

3) Any sustainable open science ecosystem model must empower participation by individual researchers and small teams. In addition to developing metadata standards, NASA must encourage the development of tools and software systems that enable the adoption and use of standards in labeling data by scientists that do not have backgrounds in computer or data science. Data sharing and archiving should be incentivized by designing systems that:
    a) provide intuitive and discipline specific tools for data discovery and navigation,
    b) describe the data set in sufficient detail so that other researchers may make effective use of it, and understand its limitations,
    c) allow contributed and utilized data sets to be cited,
    d) provide integrated analysis and collaboration capabilities,
    e) can adapt and scale with changing needs and directions inherent in research.

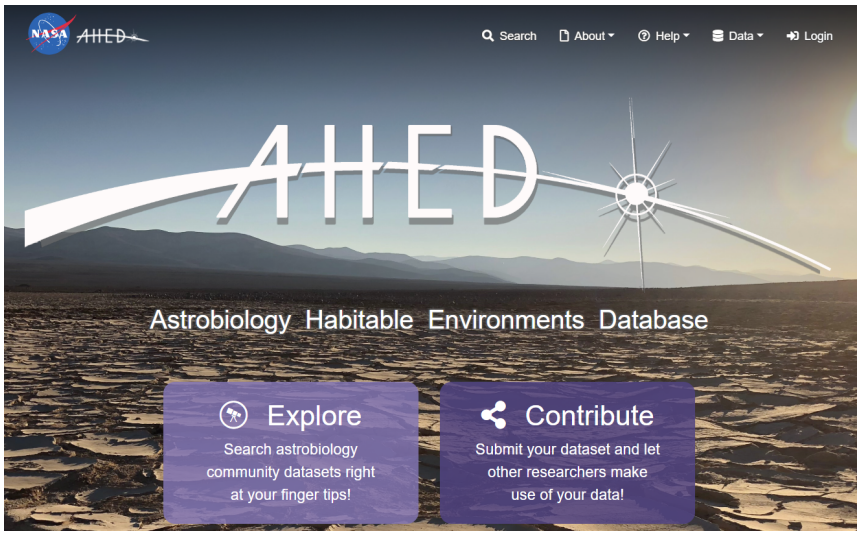

Figure 1: Screenshot of the landing page of the AHED portal.

The Value of AHED as a Case Study

Astrobiology is an inherently multidisciplinary field with proportionally large contributions from long tail research. High impact science requires integration of disparate sets of data (often complex and specialized) that may extend beyond traditional scientific disciplines, the expertise of a single team member, or even a team of scientists. Online data sharing and integration platforms within astrobiology are in their infancy, with archiving, when performed, currently relying on a patchwork of commercial and agency-run repositories and databases such as the Planetary Data System (PDS), Zenodo, Harvard Dataverse and others. This reflects the relatively recent addition of requirements to produce a data management plans (DMP) in ROSES research proposals and the complexities and challenges of archiving noted previously.

NASA's model planetary science repository is the PDS. It has been archiving planetary mission data for decades. The PDS provides curated products to scientists and to the public without charge. Rigorous standards (currently PDS4) for describing and storing data have been designed to enable future scientists who are unfamiliar with the original experiments to find and analyze the data. These standards are internationally recognized and adopted. Archived datasets are subject to peer-review and revisions before acceptance, and PDS staff work closely with data providers to ensure they

conform to data formatting and labeling requirements. The PDS is on the front-line of serving increasingly broader groups of users with high expectations and ever more complex data archiving needs, all within the constraints of mismatched funding profiles. The PDS is charged with fulfilling a variety of functions[4] but has prioritized protection of NASA's data for the future, and historically focused on archiving data and providing search tools and 'front end' capabilities for missions and larger projects in planetary sciences.

The PDS supports creation of ROSES DMPs and archiving materials from PI-led research. However, the PDS Roadmap Study for 2017-2026 raised concerns whether 'PDS nodes will have the resources to serve the data archiving requirements of individual ROSES investigations.' Searching the PDS reveals that archives stemming from PI led projects in the field of astrobiology remain uncommon. There are a variety of possible reasons for this; several have been documented in the PDS Roadmap Study[4]. Many of the criticisms PDS faces stem from the growing number of functions and customers it is charged with serving, all with a limited budget.

All this argues for exploration and guided expansion of other data management approaches in planetary science to serve growing needs, in ways that support the PDS and are compatible with other relevant international efforts (**Recommendation 1**). Some examples were highlighted in the PDS Roadmap Study and include developing approaches for archiving NASA-funded software and archiving laboratory analyses of astromaterials collected by NASA missions – a recommendation that has been adopted (see http://www.astromat.org). We recommend this list should be extended to include terrestrial analogue campaigns[6], science instrument development and capabilities and strategies for life detection. Here we describe the development and implementation of a data management strategy for PI-led research in astrobiology.

AHED Project Status and Background

The AHED project started as a NASA Science Enabling Research Activity (SERA) based at Ames Research Center. Members of the Space Science and Astrobiology, and the Intelligence Systems Divisions at Ames work with developers and scientists affiliated with the University of Arizona. AHED is envisioned as a long-term repository and productivity platform for the storage, discovery and analysis of data relevant to the field of astrobiology. The goals of AHED are to:

1) serve as a centralized digital library of NASA funded research relevant to the Astrobiology Program,
2) enable proposers to fulfill mandated data management plan (DMP) archiving requirements, and
3) serve as resource for the broader scientific community promoting the advancement of astrobiology through data sharing and standardization – including non-NASA funded research data.

The importance of leveraging NASA's institutional reputation for supporting long-term projects was identified at the project's inception. A perquisite to users' willingness to adopt a new software system and invest time sharing data hinges on assurances of continued availability, reliability, quality and ease of use.

AHED is currently a conceptually mature and functional system of software built around an astrobiology specific standardized metadata framework (called ARMS – Astrobiology Resource Metadata Standard). The AHED Portal (Fig. 1-3) provides a web-based home to the project allowing new and returning users

to create new ARMS compliant datasets, learn more about AHED and ARMS, and search for relevant datasets using a range of search tools designed around the needs of astrobiologists. Behind the scenes, the Open Data Repository (ODR) provides a powerful and flexible platform for the publication of datasets.

The project is about to launch an external pilot study in which NASA-funded users from outside Ames are asked to create datasets and provide feedback on their experiences with the system. For the pilot, software is being deployed to and cloud computing provided by Amazon Web Services (AWS). The system will be accessible to pilot users through https://astrobiology.nasa.gov. This approach will allow us to scale the service as AHED is opened up to the entire astrobiology community and follows SDMWG guidance to 'use commercial cloud environments for open science to make data accessible by diverse set of academic and commercial users.'

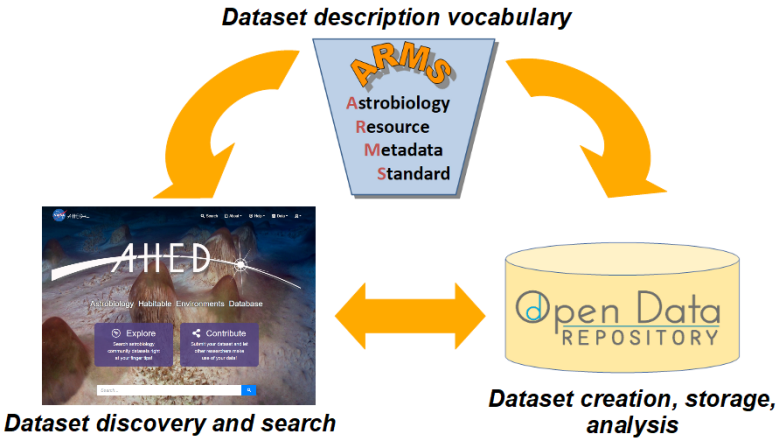

Figure 2: Components of the AHED system.

AHED Development

Astrobiology Resource Metadata Standard (ARMS):

ARMS is a key component of the AHED system and was developed in response to insights gained from two exercises performed early in the project. The first exercise involved curating a diverse set of PI generated datasets spanning the breadth of topics in astrobiology within a single online repository system (ODR), with the aim of identifying common aspects that could be used for navigation, search and as building blocks of templates for rapid creation of new datasets. A software engineering methodology called 'use-case analysis' was also carried out to anticipate the required functionality of the system and needs of users.

ARMS is intended to uniformly describe astrobiology 'resources', i.e. virtually any product of astrobiology research – including datasets, physical samples, software (modeling codes and scripts), publications, websites, images, video, presentations, etc[7]. The current version of ARMS defines 16 different metadata properties used to describe a given resource. A number of these properties are fairly generic, and cover aspects such as resource identification, personnel, funding, and publications. The true power in ARMS for search and discovery will come from three astrobiology-specific pieces of metadata used to annotate a resource:

- Research theme: The broad research area most relevant to the resource (as identified in the 2015 NASA Astrobiology Strategy Document[7]).
- Astrobiology keywords: The set of topical keywords that best characterizes the resource. A structured dictionary of ~800 keywords was developed by Ames scientists for this purpose. The keyword dictionary has been validated independently using data science techniques applied to all 2019 Astrobiology Science conference abstracts (AbSciCon 2019).

- Field location: The field site place name or geographic coordinates associated with the resource (for field campaigns or missions).

ARMS was developed by Ames scientists, but it is intended to evolve based on community feedback (starting with the upcoming external pilot study) and changing data management needs. We plan on publishing ARMS in human and computer readable formats as a community reference with detailed change log and documentation. Although the AHED system has been designed with ARMS-based search and dataset publication capabilities- the development of ARMS as a standalone metadata standard will allow astrobiologists to use alternative archiving and data publishing platforms as they wish, or develop new search and data mining capabilities, based on emerging technologies and approaches. Thus, ARMS is an example of a 'foundational component(s) of an SMD open science ecosystem' envisioned by the SDMWG. Our experience with astrobiology datasets leads us to the conclusion that additional metadata standards are required to describe data resources in other areas of science and exploration prioritized by NASA, as a foundation for the 'transformational open science' promoted by the SDMWG (**Recommendation 2)**.

ARMS and the AHED Web Portal:

Based on the importance of labeling datasets with appropriate metadata (such as ARMS) for discoverability and easy navigation between similar resources, the AHED Web Portal hosts an online dataset creation tool. The tool lets users rapidly and intuitively archive ARMS-labeled files or links to other online resources hosted by the ODR (Fig. 3). In the future, permanent identifiers such as Digital Object Identifiers (DOI) will be provided for each dataset in AHED to facilitate dataset discovery and citation. The AHED project is taking additional steps to conform with community data archiving standards (FAIR principles) being rapidly adopted by stakeholders in scientific publishing.

The AHED web portal also provides an interactive, multifaceted search interface for AHED datasets. Based on our use-case analysis we designed a variety of tools for data discovery to suit a wide audience – from amateur astrobiologists exploring the discipline to focused research scientists (Fig. 3).

Open Data Repository (ODR) Data Publisher:

ODR is being developed in parallel with the AHED system to provide a framework for managing and publishing data without the need for a programming team or specialized training. The objective of ODR is to provide an accessible end-to-end solution for data management from collection, to archiving, and analysis, focused on the needs of long tail researchers. Although ODR is the platform used by the AHED system for archiving datasets, it is designed to work with datasets and metadata standards from all scientific disciplines and is currently used to publish:

- CheMin Database (https://odr.io/chemin): living repository of CheMin and related MSL data integrated with tools and procedures for visualization and analysis.
- CheMin Analog database (https://odr.io/CheMinAnalog): Database of X-ray diffraction profiles collected on instruments that are similar to the CheMin instrument of the Mars rover, Curiosity
- Lunar Regolith XRD database (https://odr.io/lunar-regolith-xrd): Dataset containing X-Ray Diffraction (XRD) patterns and Rietveld refinement of mineralogy of 118 lunar regolith samples (<150 µm size fraction) from all landed Apollo missions.

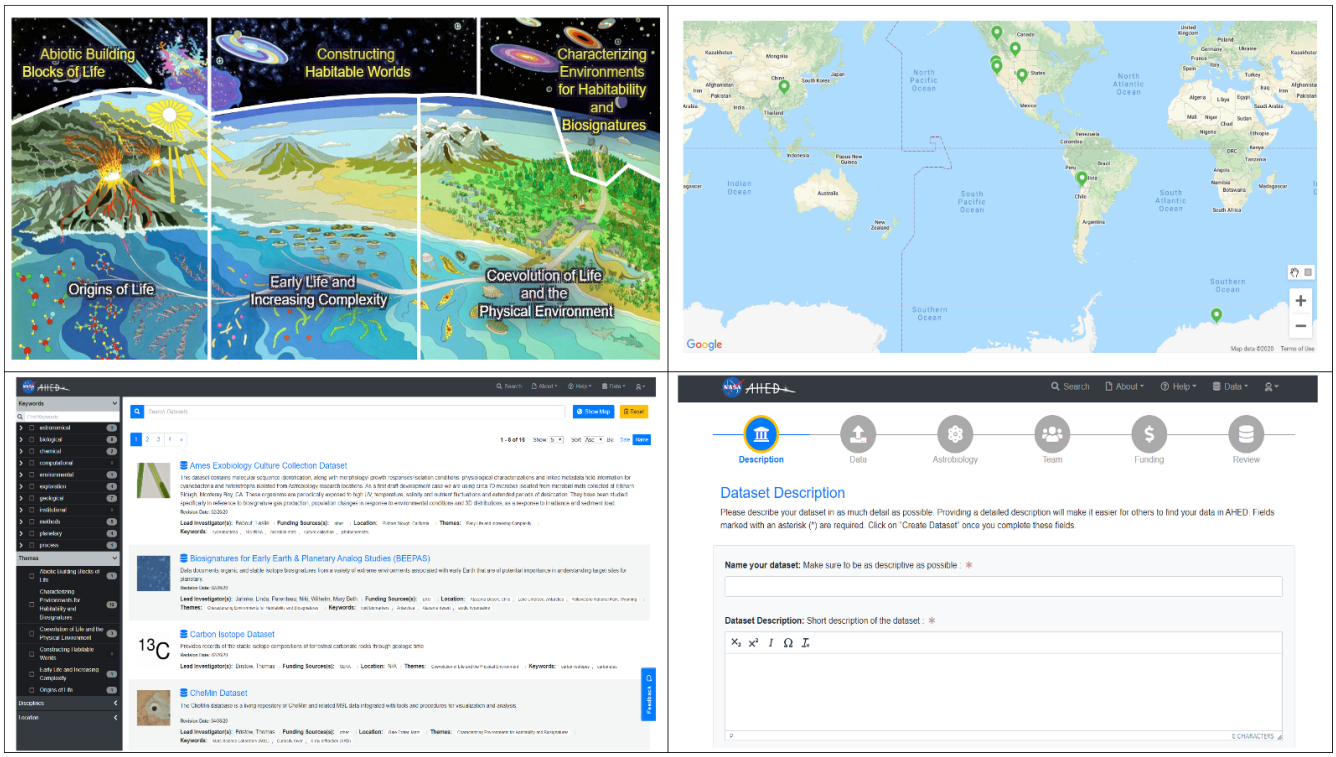

Figure 3: Screenshots of AHED search tools (top), search page (bottom left) and contribution wizard (bottom right).

ODR was created to expand upon the success and experiences learned from the creation of the RRUFF minerals database at the University of Arizona (UofA). The RRUFF database aims to collect Raman spectroscopy data, X-ray diffraction data, chemical analysis data, and various other parameters about all terrestrial minerals[8]. Since its debut in 2005, it has greatly influenced the field of mineralogy by providing an open-access resource of mineral data collected by researchers as the UofA and other cooperating scientists. In the intervening years, many scientists have asked how they can publish their data in an open and accessible way. Unfortunately, the RRUFF database is comprised of a singular, non-mutable data schema that fits only mineralogical resources.

After considering the demand from small research groups, the ODR team decided to pursue developing a dataset publication system that facilitates creation of customizable databases using a web-based interface. The ODR system lets independent researchers with small data management budgets easily create a database that will store their data and present it in human-readable web pages as well as in computer-readable formats such as CSV (comma-separated values). Additionally, the ODR system also makes each data record accessible through an API (Application Programming Interface) that facilitates interaction between external computer programs and the database.

In addition to reducing technical barriers to data sharing faced by individual PIs and small teams, ODR is also designed to provide incentives including:

- Customizable templates to facilitate database creation and cross-data searching.
- Support for interactive graphical display of data.
- Dataset citation capabilities.

- A repository of applications for processing and analyzing data within the ODR platform.
- Access control permissions to support collaboration and honor pre-publication sensitivities

The ODR system tracks every revision to a dataset and can produce a snapshot of the dataset at any point in time. Datasets can also be expanded to incorporate data derived from newer research techniques or data formats as the dataset lives and evolves. These tools allow researchers to create, publish, analyze, and maintain living datasets that are more useful and versatile than static repositories of legacy data.

Concluding remarks

The AHED system is designed so that a typical user will not interact directly with the ODR, simplifying the process of data sharing for all users - while nudging motivated teams and individuals to take advantage of powerful data publishing features and tools provided by the ODR. The AHED portal is designed to search for ARMS labeled datasets irrespective of data publishing platform. This means that the AHED system could, in theory, operate with another dataset publication and storage system. Thus, AHED web portal and ODR are examples of 'modular open service(s)' identified by the SDMWG as 'foundational components of an SMD open science ecosystem'. They are designed to empower direct participation by individual researchers and small teams in data sharing and management – something we think is a necessity for developing a sustainable and scalable open science ecosystem **(Recommendation 3)**.

References

1 Open Data in a Big Data World – An International Accord (2016) *Chemistry International*, 38, P. 17.

2 6th & Final Report of the Big Data Task Force. https://science.nasa.gov/science-committee/subcommittees/big-data-task-force

3 Aydinoglu, A.U., Suomela, T., Malone, J. (2014) Data Management in Astrobiology: Challenges and Opportunities for an Interdisciplinary Community. *Astrobiology*, 14, 451-461.
Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E., Manoff, M., Frame, M. (2011) Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE,* 6: e21101.

4 Planetary Data System Roadmap Study for 2017 – 2026. https://pds.jpl.nasa.gov/home/about/PlanetaryDataSystemRMS17-26_20jun17.pdf

5 Science Mission Directorate's Strategy for Data Management and Computing for Groundbreaking Science 2019-2024. Strategic Data Management Working Group. https://science.nasa.gov/science-red/s3fs-public/atoms/files/SDMWG%20Strategy_Final.pdf

6 Stern, J., Weng, M., Graham, H., Bowen, J., Hooker, S. et al. (2020) Building Consensus, Collaboration, and Capability for Ocean Worlds Field Science. White Paper submitted to National Academies.

7 Keller, R.M., Blake, D.F., Bristow, T., Cooper, G., Dateo, C.E. et al. (2019) ARMS: A Developing Metadata Standard for Describing Astrobiology Research Products. Astrobiology Conference, Bellevue, WA, 24-28 June, abstract 401-9.

8 Lafuente B., Downs R.T., Yang H., Stone, N. (2015) The power of databases: the RRUFF project. In: Highlights in Mineralogical Crystallography, T. Armbruster and R M Danisi, eds. Berlin, Germany, W. De Gruyter, p.1-30.