A STRATEGY FOR MANAGING NASA'S LONG TAIL OF PLANETARY RESEARCH DATA: INSIGHTS FROM THE DEVELOPMENT OF THE AHED REPOSITORY. T. F. Bristow[1], B. Lafuente[2], S. Wolfe[1], N. Parenteau[1], N. Stone[3], S. Rojo[1], K. Boydstun[1], D. Blake[1], R. Downs[4], C. Dateo[1]. [1]NASA Ames Research Center, Moffett Field, CA USA (thomas.f.bristow@nasa.gov); [2]SETI Institute, Mountain View, CA USA; [3]The Open Data Repository, Main, MA, USA; Geoscience Department, University of Arizona

Introduction: The importance of sharing scientific data is increasingly recognized by the 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 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. The Astrobiology Habitable Environments Database (AHED) is a long-term, open-access repository and productivity platform for the storage, search, and analysis of diverse data relevant to the field of astrobiology [5]. Here we detail the components and architecture of the AHED system to help guide strategies for data management in other NASA-funded scientific disciplines where long tail research is performed.

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. 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.

AHED is currently a conceptually mature and functional system of software [6] (Fig. 1), built around an astrobiology specific standardized metadata framework (called ARMS – Astrobiology Resource Metadata Standard). The AHED Portal 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.

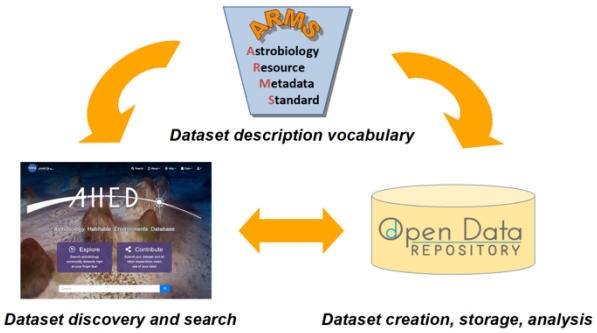

Figure 1: Components of the AHED system.

ARMS: 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. However, astrobiology-specific pieces of metadata are essential in making datasets discoverable in a variety of use-cases. We think extending metadata standards to other scientific disciplines prioritized by NASA is essential to future data management efforts.

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 (Fig. 2) 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 [8]) being rapidly adopted by stakeholders in scientific publishing. The AHED web portal also provides an interactive, multifaceted search interface for AHED datasets.

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.

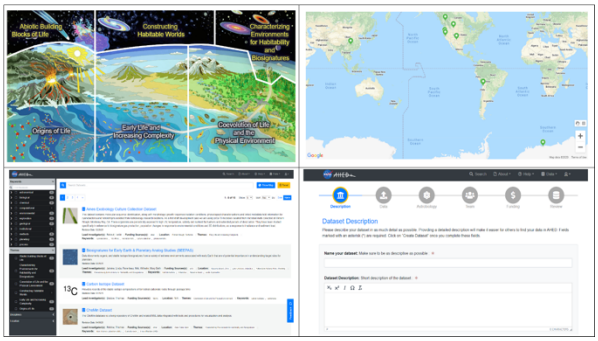

Figure 3: Screenshots of AHED search tools (top), search page (bottom left) and contribution wizard (bottom right).

Summary: 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, the AHED Web Portal and ODR are examples of 'modular open service(s)' identified by the Strategic Data Management Working group (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 across a range of disciplines.

Acknowledgments:
We gratefully acknowledge the support for this work by the NASA Planetary Science Division's Science Enabling Research Activity (SERA) and the NASA Mars Science Laboratory (MSL) Program.

References:
[1] Open Data in a Big Data World – An International Accord (2016) *Chem. Int.*, 38, 17. [2] 6th & Final Report of the Big Data Task Force. [3] Aydinoglu, A.U. et al. (2014) *Astrobiology*, 14, 451–461. [4] Planetary Data System Roadmap Study for 2017– 2026. [5] Bristow, T.F. (2020) *PSADS white paper*. [6] Detweiler, A. et al. (2019) *AbSciCon*. Abstract# 319-213. [7] Keller, R.M. et al. (2019) *AbSciCon*. Abstract# 401-9. [8] Wilkinson, M.D. et al., (2016) *Scientific Data,* 3, 160018.