ASTROBIOLOGY
Volume 24, Number 2, 2024

DOI: 10.1089/ast.2023.0067

Rapid Communication

Open camera or QR reader and scan code to access this article and other resources online.

Enabling Data Discovery with the Astrobiology Resource Metadata Standard

Shawn R. Wolfe,[1] Barbara Lafuente,[2] Richard M. Keller,[1] Angela M. Detweiler,[3] Thomas F. Bristow,[1] Mary N. Parenteau,[1] Kevin Boydstun,[1] Christopher E. Dateo,[1] David J. Des Marais,[1] Linda L. Jahnke,[1] Sara Rojo,[1] Nathan Stone,[4] and Mark Vorobets[1]

Abstract

As scientific investigations increasingly adopt Open Science practices, reuse of data becomes paramount. However, despite decades of progress in internet search tools, finding relevant astrobiology datasets for an envisioned investigation remains challenging due to the precise and atypical needs of the astrobiology researcher. In response, we have developed the Astrobiology Resource Metadata Standard (ARMS), a metadata standard designed to uniformly describe astrobiology ''resources,'' that is, virtually any product of astrobiology research. Those resources include datasets, physical samples, software (modeling codes and scripts), publications, websites, images, videos, presentations, and so on. ARMS has been formulated to describe astrobiology resources generated by individual scientists or smaller scientific teams, rather than larger mission teams who may be required to use more complex archival metadata schemes. In the following, we discuss the participatory development process, give an overview of the metadata standard, describe its current use in practice, and close with a discussion of additional possible uses and extensions. Key Words: Metadata—Open science—Long tail. Astrobiology 24, 131–137.

1. Introduction

In response to a White House directive for open and equitable research (White House, 2023), the National Aeronautics and Space Administration (NASA), along with other federal agencies, has declared 2023 to be the Year of Open Science (NASA TOPS, 2023). Open Science entails the sharing of scientific data, software, and other research products (Fecher and Friesike, 2014). This is important because reuse and repurposing of legacy data is increasingly commonplace, with over half of the investigations funded by NASA's Science Mission Directorate (SMD) based on archival data (Big Data Task Force, 2017), for example. However, discovery of relevant astrobiology datasets remains difficult in part due to the lack of precise dataset descriptors tailored toward the needs of researchers (Aydinoglu *et al.,* 2014). Hence, we have developed the Astrobiology Resource Metadata Standard (ARMS) (Keller *et al.,* 2019), an evolving comprehensive standard for the description, access, and discovery of information related to all areas relevant to astrobiology. ARMS goes beyond just datasets and can describe any product of astrobiology research, including physical samples, software, publications, and so on. From an astrobiology viewpoint, the advantage of using ARMS over a generic metadata standard comes from the inclusion of metadata specific to astrobiology, which allows the researcher to describe their information more precisely and in much greater detail.

ARMS is not a static metadata standard, but one that will evolve based on community feedback, emerging trends in scientific focus, and changing data management needs. Developed within the context of the Astrobiology Habitable Environments Database (AHED) (AHED Team, 2023a), ARMS is nonetheless a standalone metadata standard

[1]NASA Ames Research Center, Moffett Field, California, USA.
[2]The SETI Institute, Mountain View, California, USA.
[3]Bay Area Environmental Research Institute, Moffett Field, California, USA.
[4]Open Data Repository, Gray, Maine, USA.

independent of AHED, fully available to be used in novel contexts. The ARMS standard is described online (AHED Team, 2023b) and available as an Extensible Markup Language Schema (World Wide Web Consortium, 2004). By making ARMS publicly available, we support NASA's vision of an Open Science ecosystem that will lead to a transformation in science, increasing accessibility to knowledge and accelerating scientific discoveries (Strategic Data Management Working Group, 2019; NASA Science Mission Directorate, 2022).

2. Background

General guidance exists on what characteristics a metadata standard should have, as well as its usage. The FAIR principles recommend optimizing specific qualities of data (Wilkinson *et al.,* 2016), such that they should be *F*indable, *A*ccessible, *I*nteroperable and *R*eusable, hence the acronym. Many of these principles relate to design of metadata. TRUST (Lin *et al.,* 2020) is another set of principles that are more geared toward properties of repositories, advocating for *T*ransparency, *R*esponsibility, *U*ser Focus, *S*ustainability, and *T*echnology. NASA's SMD requires that all newly SMD-funded research shall make its publications, data, and software publicly available and that its data should follow the FAIR principles (NASA Science Mission Directorate, 2022). Greenberg *et al.* (2009) argue that a scientific metadata architecture should be easy to use, interoperable with other standards, and suitable for machine processing. However, a survey of 16 metadata standards found that these aims are rarely met (Qin and Li, 2013). Although there are many semantic commonalities among the standards, differences in terminology create barriers to interoperability and tools reuse. Complexity in some standards also creates an undue burden on the creators of the metadata.

Several standards for scientific metadata have informed and influenced our development of ARMS. Darwin Core (Wieczorek *et al.,* 2012) was developed for biodiversity informatics, influenced by the seminal Dublin core metadata standard (Weibel *et al.,* 1998). In turn, the Dryad repository developed their own metadata standard (Greenberg *et al.,* 2009), based in part on the Darwin Core, with a focus on evolutionary biology. The Investigation/Study/Assay Metadata Framework (Johnson *et al.,* 2021) is concentrated on life science and structured around the three concepts in its name. Similarly, the Core Scientific Metadata Model (Matthews *et al.,* 2010) was developed to describe data from large-scale facilities, with its core organizing concept being a scientific study. NASA's Planetary Data System (Planetary Data System, 2022a) has its own metadata standard for encoding the data (Planetary Data System, 2022b), with its obvious focus on planetary missions. Finally, the Site-Based Data Curation Project developed a metadata standard to describe scientifically significant data from Yellowstone's hot springs research (Palmer *et al.,* 2017), a valuable site for a variety of astrobiological investigations.

As a multidisciplinary field, describing astrobiology metadata is challenging due to the varied data sources, measurements, and formats in use (Detweiler *et al.,* 2019). We found no existing metadata standard had both the scope and detail needed—either the standard failed to cover the breadth of astrobiology or was so high-level as to be of limited use. Moreover, we wanted to make ARMS available quickly, keeping the standard lightweight initially. This would ease early adoption, as metadata would initially be created largely by hand, with tools and automation allowing us to add more semantics in later years. Thus, we made the pragmatic choice of creating ARMS as an independent standard. The downside of initially avoiding some complexity is that future integration with existing standards and improving semantics becomes more difficult.

3. Development Methodology

We used a participatory design approach (Spinuzzi, 2005), where users of the desired product (in this case, ARMS) actively participate in the design process. The development team consisted both of computer scientists well-versed in metadata and standards as well as astrobiologists who would be the eventual end-users. In addition, we regularly met with scientists outside our team to vet our modeling choices and to ensure ARMS represented the larger field. We also consulted the literature, particularly the standards mentioned in Section 2. To develop the astrobiology-specific elements, specifically a set of keywords, we reviewed astrobiology journal keywords from the prior 10 years as well as keywords independently developed by Taşkın and Aydinoglu (2015) and Miller *et al.* (2014). We validated the ARMS keywords by searching for their occurrences in the abstracts of the 2019 Astrobiology Science Conference (Meadows, 2019). Over 98% of these abstracts contained at least one of our keywords, indicating good coverage; conversely, about one-third of the keywords were not found in any abstract (Lafuente *et al.,* 2019). The extensive coverage from our initial set of keywords could mask a lack of specificity in certain areas, where a general keyword is found in the text, but a more appropriate, specific keyword is missing from ARMS. To discover new potential keywords, we identified groups of similar publications and then identified distinguishing keywords of each, mimicking an artificial intelligence technique known as topic modeling (Blei, 2012). We used abstracts from the 2019 Astrobiology Science Conference as the corpus for our study. We used the *partitioning around medoids* (PAM) algorithm (Schubert and Rousseeuw, 2019) to organize the abstracts into disjoint clusters, with the number of clusters chosen empirically to yield clusters that were neither too broad nor too narrow. From each cluster, we automatically identified terms that were rare in other clusters but prevalent within the cluster as potential keywords to add to ARMS. From this, we identified 36 keywords that we added into ARMS, including new areas such as education and outreach.

4. Description

To be ARMS compliant, the astrobiology resource (*e.g.,* dataset, software, physical sample) must be accompanied by a standard set of metadata that describes the resource and characterizes its content. These metadata fall into two categories: *resource identification metadata* (describing provenance, points of contact, versioning information, funding/support, and associated geospatial collection information) and *content metadata* (describing the actual content of the resource and its relation to the broader astrobiological

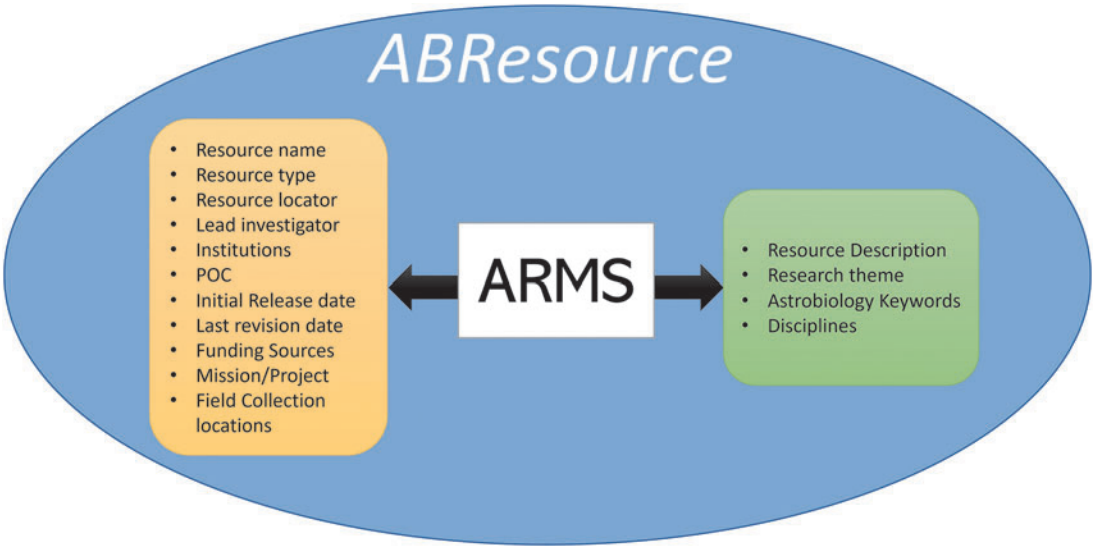

FIG. 1. Top-level elements of an ARMS AstroBiology Resource (ABResource), with identification metadata on the left and content metadata on the right.

context), as shown in Fig. 1. Some metadata elements are free text, while others are restricted or complex types.

Of particular interest are the metadata elements that are especially useful as search parameters, namely

- *Research Theme*, the broad research area(s) most relevant to the resource, focused on those identified in the 2015 NASA Astrobiology Strategy Document (Hays *et al.,* 2017);
- *Project or Mission* during which the resource was collected;
- *Discipline*(s) most relevant to the creation or use of the resource; and
- *Keyword*(s) that best characterizes the astrobiology resource described by the ARMS metadata.

The keywords are organized into a four-tier hierarchy, under a top-level structure of 11 categories (AHED Team, 2023b). These top-level categories (in italics) contain keywords that describe the resource in several ways: further detailing the scientific discipline (*astronomical*, *biological*, *chemical*, or *geological*) or physical *process*; characterizing the collection (*environmental* [in situ], *exploration* [of unusual environments], *planetary* [beyond Earth]); the analysis *methods* or *computational* methods employed; and the *institutional* support for acquiring the resource. An example of some of these metadata is given in Fig. 2.

5. ARMS in Practice

ARMS was initially developed to standardize the metadata contained in AHED (AHED Team, 2023a), a long-term repository and productivity platform for the storage, discovery, and analysis of data relevant to the field of astrobiology (Bristow *et al.,* 2021). The consistent structure established by ARMS greatly facilitates precise querying and rapid interpretation of search results. The AHED Web Portal (AHED Team, 2023a) hosts an online dataset creation tool, letting users rapidly and intuitively archive ARMS-labeled files or links to other online resources (Fig. 3). AHED also provides an interactive, multifaceted search interface for AHED datasets based on ARMS metadata.

Beyond AHED, ARMS can be used to inform the astrobiology portion of a comprehensive science modeling effort, or even be directly incorporated into such. For instance, we have participated in the NASA Science Mission Directorate's Data Catalog effort, which seeks to build a cross-science search capability (Bugbee *et al.,* 2022) and provided ARMS to inform their modeling efforts. In a similar vein, portions of ARMS may be useful for indexing astrobiology text documents, particularly the extensive keyword hierarchy. Examples of this include indexing astrobiology research papers, conference submissions, or proposals for funding. Finally, ARMS could be used to support cross-system communication and representation of astrobiology resources. In fact, the AHED system uses the Open Data Repository (Lafuente *et al.,* 2018), a general data repository, as its backend.

6. Conclusions and Future Work

The transformation to Open Science cannot happen by fiat; it must receive the necessary institutional and technological support. Toward this end, we have created ARMS to describe products of astrobiology research. Standardizing the metadata aids the discovery and interpretation of astrobiological datasets, supporting the aims of Open Science. Our approach can also serve as a blueprint for similar endeavors in other science disciplines. As discussed above, ARMS is not the only metadata standard with relevance to astrobiology. Efforts should be made to harmonize ARMS with other relevant standards, such as Darwin Core and PDS-4, potentially translating ARMS to or from these standards. As the discipline and focus of astrobiology evolves, so must ARMS; in particular the keywords, funding sources, missions/projects, and science disciplines will need to be updated over time. The methods outlined in Section 3 to identify keywords can be used to find new keywords in the future, but these too should evolve. Ultimately, this

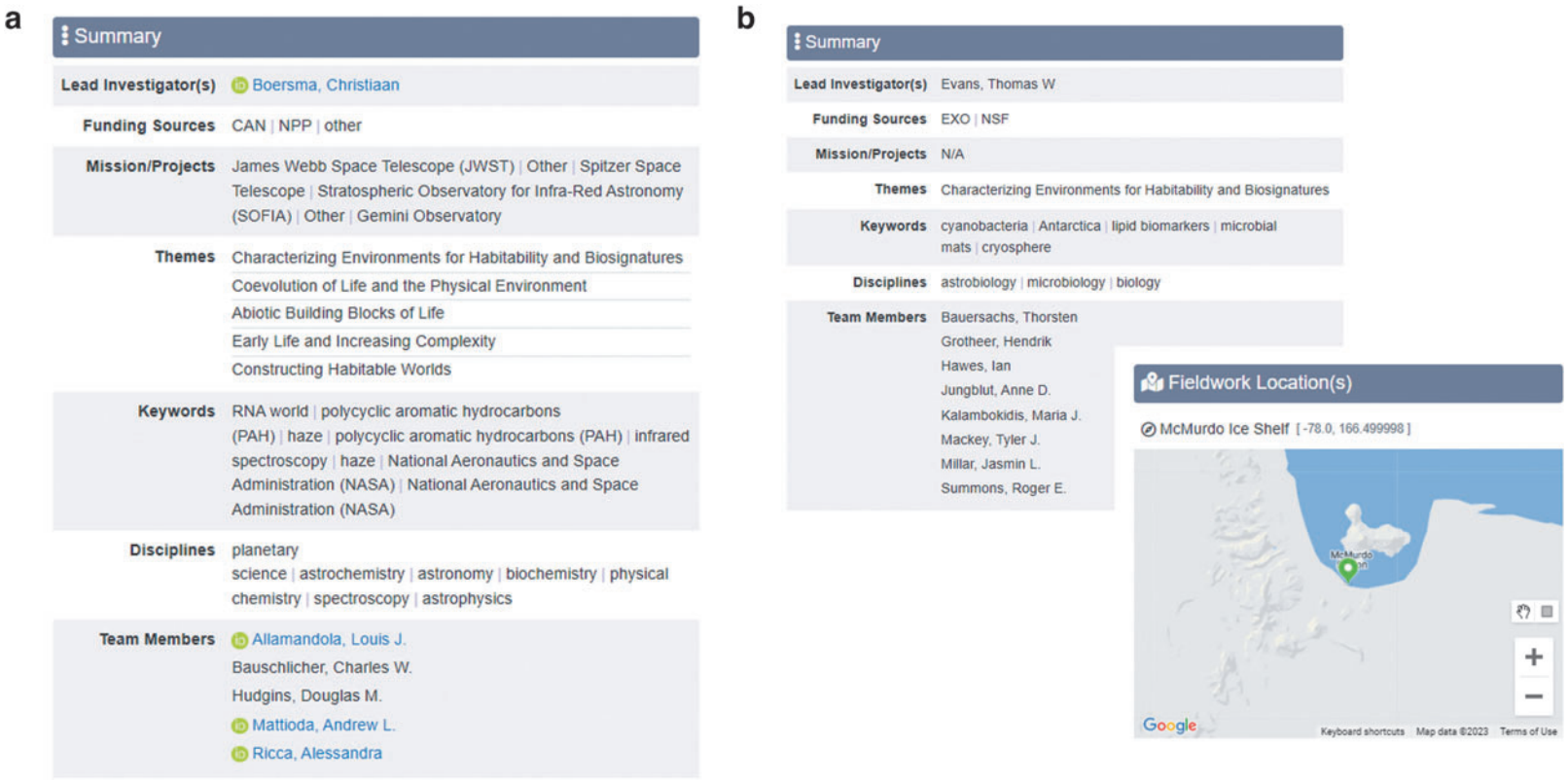

FIG. 2. Subset of ARMS metadata for (**a**) The NASA Ames PAH IR Spectroscopic Database (Allamandola *et al.,* 2022) and (**b**) Lipid Biomarkers from Microbial Mats on the McMurdo Ice Shelf, Antarctica: Signatures for Life in the Cryosphere (Bauersachs *et al.,* 2022), with a detail of the Fieldwork Location map. Only the leaf (terminal) nodes are displayed for the keywords.

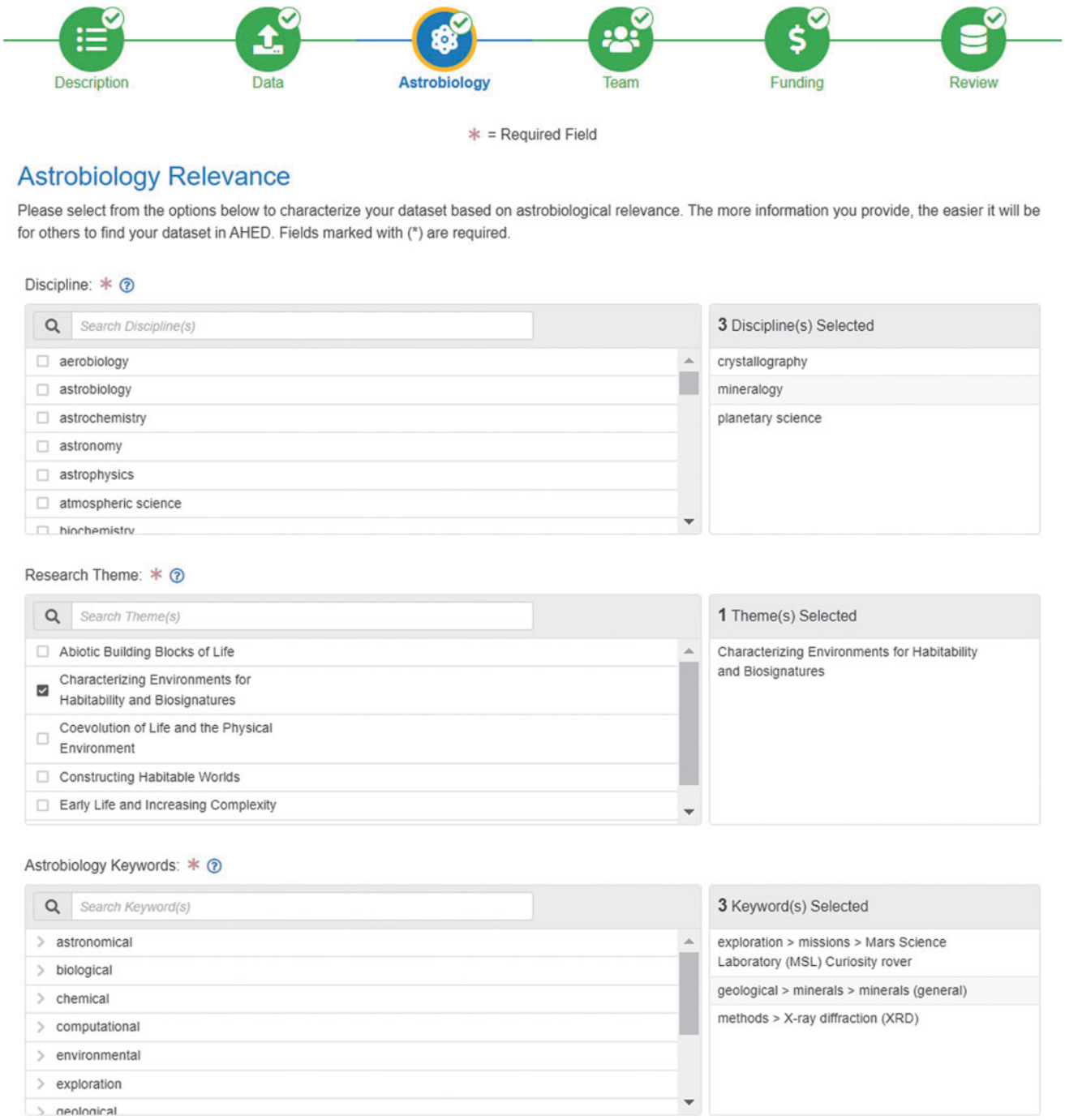

FIG. 3. Excerpt of AHED online tool to create ARMS metadata.

approach could be expanded to not only update the existing structure but to facilitate generating initial keyword structures, so that our approach can be applied to new science domains. Finally, more tools should be developed to assist the labeling of astrobiology datasets with ARMS metadata, leveraging recent advances in natural language processing as appropriate. Within AHED, we have strived to streamline this process, but nonetheless it is a multistep process that can take tens of minutes. Choosing appropriate keywords can be particularly daunting. However, much of this information can be easily gleaned from available sources, for instance associated publications that can be provided as part of the submission process. The challenge is to extract this information automatically from such documents.

Acknowledgments

The authors would like to thank Caleb Scharf and Daniel Berrios for their reviews and comments.

Author Disclosure Statement

No competing financial interests exist.

Funding

This work was funded by NASA SMD Planetary Science Division's Science Enabling Research Activity (SERA) program.

References

AHED Team. *Astrobiology Habitable Environments Database*. NASA, Washington, DC; 2023a. Available from: https://ahed.nasa.gov [Last accessed 3/10/2023].

AHED Team. *Astrobiology Resource Metadata Standard*. NASA, Washington, DC; 2023b. Available from: https://ahed.nasa.gov/help/help-arms [Last accessed 3/10/2023].

Allamandola L, Bauschlicher C, Boersma C, *et al. The NASA Ames PAH IR Spectroscopic Database. Astrobiology Habitable Environment Database*. NASA, Washington, DC; 2022; Available from: https://doi.org/10.48667/6p1n-w007 [Last accessed 10/3/2023].

Aydinoglu AU, Suomela T, Malone J. Data management in astrobiology: Challenges and opportunities for an interdisciplinary community. *Astrobiology* 2014;14(6):451–461; doi: 10.1089/ast.2013.1127.

Bauersachs T, Evans T, Grotheer H, *et al. Lipid Biomarkers from Microbial Mats on the McMurdo Ice Shelf, Antarctica: Signatures for Life in the Cryosphere. Astrobiology Habitable Environment Database*. NASA, Washington, DC, 2022; https://doi.org/10.48667/r1zs-v785 [Last accessed 10/3/2023].

Big Data Task Force. *6th and Final Report of the Big Data Task Force*. NASA, Washington, DC; 2017. Available from: https://science.nasa.gov/science-committee/subcommittees/big-data-task-force [Last accessed 3/10/2023].

Blei DM. Probabilistic topic models. *Commun ACM* 2012; 55(4):77–84 ; doi: 10.1145/2133806.2133826.

Bristow TF, Lafuente B, Wolfe SR., *et al.* A strategy for managing NASA's long tail of planetary research data: Insights from the development of the AHED repository. In *5th Planetary Data Workshop & Planetary Science Informatics & Analytics*. Lunar and Planetary Institute, Houston, 2021; abstract 7093.

Bugbee K, Ramachandran R, Acharya A, *et al.* Approaches for enabling interoperable enterprise data search: Insights from NASA's Science Mission Directorate (SMD) catalog project. In *EGU General Assembly Conference Abstracts*. European Geosciences Union: Vienna, Austria, 2022; doi: 10.5194/egusphere-egu22-5940.

Detweiler AM, Lafuente B, Keller RM, *et al.* Enhancing data sharing, discovery, and analysis in the astrobiology community. In *Proceedings of the 2019 Astrobiology Science Conference (AbsSciCon 2019)*. American Geophysical Union: Washington, DC, 2019; abstract 319-213.

Fecher B, Friesike S. Open Science: One term, five schools of thought. In *Opening Science*. (Bartling S, Friesike S. eds.) Springer: Cham, Switzerland, 2014; pp 17–47.

Greenberg J, White HC, Carrier S, *et al.* A metadata best practice for a scientific data repository. *J Libr Metadata* 2009; 9(3–4):194–212; doi: 10.1080/19386380903405090.

Hays LE, New MH, Voytek MA. 2015 NASA Astrobiology Strategy Document and the Vision for Solar System Exploration. In *Planetary Science Vision 2050 Workshop*. Lunar and Planetary Institute, Houston; 2017.

Johnson D, Batista D, Cochrane K, *et al.* ISA API: An open platform for interoperable life science experimental metadata. *GigaScience* 2021;10(9):giab060; doi: 10.1093/gigascience/giab060.

Keller RM, Blake DF, Bristow TF, *et al.* ARMS: A developing metadata standard for describing astrobiology research products. In: *Proceedings of the 2019 Astrobiology Science Conference (AbsSciCon 2019)*. American Geophysical Union: Washington, DC, 2019; abstract 401-9.

Lafuente B, Stone N, Bristow TF, *et al.* The Open Data Repository's (ODR) data publisher. In *AGU Fall Meeting Abstracts 2018*. American Geophysical Union: Washington, DC, 2018; abstract #0653.

Lafuente B, Detweiler AM, Keller RM, *et al.* The Astrobiology Habitable Environments Database (AHED) and the Astrobiology Resource Metadata Standard (ARMS): Community-driven tools for astrobiological data. In: *AGU Fall Meeting Abstracts 2019*. American Geophysical Union: Washington, DC, 2019; abstract #P21E-3423.

Lin D, Crabtree J, Dillo I, *et al.* The trust principles for digital repositories. *Sci Data* 2020;7(1):144; doi: 10.1038/s41597-020-0486-7.

Matthews B, Sufi S, Flannery D, *et al.* Using a core scientific metadata model in large-scale facilities. *Int J Digit Curation* 2010;5(1):106–118; doi: 10.2218/ijdc.v5i1.146.

Meadows V. (ed.) *Proceedings of the 2019 Astrobiology Science Conference (AbsSciCon 2019)*. American Geophysical Union: Washington, DC; 2019.

Miller LJ, Gazan R, Still S. Unsupervised classification and visualization of unstructured text for the support of interdisciplinary collaboration. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. Association for Computing Machinery, New York, 2014; pp 1033–1042; doi: 10.1145/2531602.2531666.

NASA Science Mission Directorate. *SMD policy document SPD-41a*. NASA, Washington, DC; 2022. Available from: https://ahed.nasa.gov/odr/view/downloadfile/20319 [Last accessed 10/19/2023].

NASA TOPS. *2023: The Year of Open Science*. NASA, Washington, DC; 2023. Available from: https://nasa.github.io/Transform-to-Open-Science/year-of-open-science [Last accessed 3/10/2023].

Palmer CL, Thomer AK, Baker KS, *et al.* Site-based data curation based on hot spring geobiology. *PLoS One* 2017; 12(3): e0172090; doi: 10.1371/journal.pone.0172090

Planetary Data System. *Welcome to the Planetary Data System*. NASA, Washington, DC; 2022a. Available from: https://pds.nasa.gov [Last accessed 3/10/2023].

Planetary Data System. *Planetary data system current version 1.19.0.0*. NASA, Washington, DC; 2022b. Available from: https://pds.nasa.gov/datastandards/documents/current-version.shtml [Last accessed 3/10/2023].

Qin J, Li K. How portable are the metadata standards for scientific data? A proposal for a metadata infrastructure. In: *Proceedings of the 2013 International Conference on Dublin Core and Metadata Applications*. Dublin Core Metadata Initiative, Lisbon, Portugal; 2013; pp 25–34.

Schubert E, Rousseeuw PJ. (2019), Faster k-medoids clustering: Improving the PAM, CLARA, and CLARANS algorithms. In: *Proceedings of the 12th International Conference on Similarity Search and Applications (SISAP 2019)*. Springer International Publishing, Newark, NJ; 2019; doi: 10.1007/978-3-030-32047-8_16.

Spinuzzi C. The methodology of participatory design. *Tech Commun* 2005;52(2):163–174.

Strategic Data Management Working Group. *Science Mission Directorate's Strategy for Data Management and Computing*

for Groundbreaking Science 2019–2024. NASA, Washington, DC; 2019. Available from: https://ahed.nasa.gov/odr/view/downloadfile/20320 [Last accessed 10/19/2023].

Taşkın Z, Aydinoglu AU. Collaborative interdisciplinary astrobiology research: A bibliometric study of the NASA Astrobiology Institute. *Scientometrics* 2015;103(3):1003–1022; doi: 10.1007/s11192-015-1576-8.

Weibel S, Kunze JA, Lagoze C, *et al. Rfc2413: Dublin Core Metadata for Resource Discovery*. (Del Rey M. ed.) The Internet Society, Reston, VA; 1998; doi: 10.17487/RFC2413.

White House. *Fact Sheet: Biden-Harris Administration Announces New Actions to Advance Open and Equitable Research*. The White House, Washington, DC; 2023. Available from: https://www.whitehouse.gov/ostp/news-updates/2023/01/11/fact-sheet-biden-harris-administration-announces-new-actions-to-advance-open-and-equitable-research [Last accessed 3/10/2023].

Wieczorek J, Bloom D, Guralnick R, *et al.* Darwin core: An evolving community-developed biodiversity data standard. *PLoS One* 2012;7(1):e29715; doi: 10.1371/journal.pone.0029715.

Wilkinson MD, Dumontier M, Aalbersberg IJ, *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;3(1):160018; doi: 10.1038/sdata.2016.18.

World Wide Web Consortium. *XML Schema Part 0: Primer Second Edition*. W3C, Wakefield, MA; 2004. Available from: https://www.w3.org/TR/xmlschema-0 [Last accessed 3/10/2023].

Address correspondence to:
Shawn R. Wolfe
Ames Research Center
PO Box 1
Moffett Field, CA 94035
USA

E-mail: shawn.r.wolfe@nasa.gov

Submitted 2 June 2023
Accepted 1 October 2023
Associate Editor: Michael C. Storrie-Lombardi

Abbreviations Used

AHED = Astrobiology Habitable Environments Database
ARMS = Astrobiology Resource Metadata Standard
SMD = Science Mission Directorate