The Astrobiology Data Ecosystem, Open Science, and the AI Era

Diana Gentry, NASA ARC diana.gentry@nasa.gov
Paul M. Bremner, NASA MSFC
Nathalie Cabrol, SETI Institute
Victoria Da Poian, Microtel LLC / NASA GSFC
Ryan Felton, ORAU / NASA ARC
Jian Gong, University of Wyoming
Ashley M. Hanna, University of Maryland, College Park
Adrienne Hoarfrost, University of Georgia
Floyd Nichols, Virginia Tech
Conor A. Nixon, NASA GSFC
Tejas Panambur, University of Massachusetts
M. Joseph Pasterski, ORAU / NASA GSFC
Sunanda Sharma, Carnegie Institution for Science
Brenda Thomson, Rensselaer Polytechnic Institute
Caleb Scharf, NASA ARC
Hamed Valizadegan, KBR / NASA ARC
Kimberley Warren-Rhodes, SETI Institute / NASA ARC
Michael L. Wong, Carnegie Institution for Science
Anastasia Yanchilina, SETI Institute

DARES Alignment

- Primary Topic: Identify Emerging Themes and Technologies
- Secondary Topic(s): Review Recent Advancements, Strengthen Community

Motivation

Astrobiology, perhaps more than any other field, is inherently cross-disciplinary. It studies the origin, history, status, and fate of habitable planets, life, ecosystems, and civilizations. This requires understanding complex and deeply intertwined physical, chemical, biological and social phenomena integrated across vastly different scales of time, space, and energy. Machine learning (ML), and the broader set of related models and techniques that are currently called artificial intelligence (AI), offer a rapidly expanding and unparalleled ability to reveal, connect, and model relationships – including non-linear and context-dependent ones – between large numbers of features in data of many different types [1, 2]. Astrobiology, therefore, may have a uniquely high potential to benefit from these tools.

Recent astrobiology AI/ML implementations include identifying mineral types associated with habitability from Raman and LIBS spectra [3], classifying the transit signals of Kepler and TESS to find new exoplanets [4], and distinguishing mass spectra [5, 6], XRF spectra [7], or isotopic signatures [8] of biogenic and abiogenic organic compounds. The greater benefit is likely to come from use of *multi-modal data* to explore the boundaries that separate life from non-life, the processes that characterize them, and the path(s) towards life's emergence. Multi-modal data can be any combination of visible imagery, reflectance spectroscopy, mass spectrometry, fluorescence spectrometry, micrography, isotope ratio analysis, Raman spectroscopy, X-ray diffraction, morphology, topography, metagenomics, and many more. However, to tap into the full potential of ML, there is a requirement for sufficient depth and breadth of cross-compatible data in order to create robust models – and these requirements scale upwards with the complexity and diversity of the systems under study. If astrobiology is to take advantage of these tools, we must invest in making both the data we already have and the data we take in the future suitable for AI/ML applications.

These data requirements are synergistic with the emerging shift within the scientific community towards open science, typified by NASA's Open Science Initiative. Open science practices, such as compatibility with the FAIR principles [9], can reduce unneeded duplication of effort, improve scientific reproducibility, increase the value of historical and future data, and lower barriers to innovation, especially in cross-disciplinary fields. Astrobiology particularly benefits from open data and sample sharing, as it relies on measurements and materials from difficult-to-access field sites, rare laboratory facilities that simulate hypothetical or real off-world environments, or unique and irreplaceable planetary exploration operations. Many fields within or overlapping with astrobiology are making progress building local open data ecosystems, such as the Planetary Data Ecosystem, Astrobiology Habitable Environments Database, the Metabolomics Workbench, and the NASA Open Science Data Repository. Yet even if these efforts were integrated, major gaps in the quantity, quality, and types of data and samples currently covered would remain.

This white paper **represents a collection of ground-level observations on the current state of the astrobiology data ecosystem and anticipated needs for the emerging era of open data and AI applications**, drawn from the authors' recent experiences and conversations. While they skew towards field and laboratory work, several apply to mission and exoplanet observation data as well. They are grouped into four areas, emphasizing cross-disciplinary and multi-modal applications, with accompanying recommendations:

1. finding and unifying the data we already have;
2. getting the data we need but don't have;
3. improving access to unique resources; and
4. lowering barriers to implementation with streamlining and support.

Needs and Recommendations

1. Finding what we have: data labeling, indexing, and search

Observations: Consider a simple project to compare images, Raman spectra, and elemental composition of two leeward cold basaltic rock faces, one from Earth and one from Mars. All the data necessary to do this exists within resources like RRUFF, the USGS Spectral Library, and the PDS Geosciences Node. However, finding it requires days of reading through extensive annotations to identify and extract the correct data files, and those files often require (or have been subjected to) processing that is only documented in publications several links away – if at all. This difficulty substantially lowers the value of past mission, field, and laboratory data, and is currently one of the largest barriers to assembling a multi-modal astrobiology data set large enough to be useful for AI/ML work.

Many disciplines within astrobiology's broad umbrella have established their own ontologies, ranging from naming schemes like IUPAC nomenclature, to keyword thesauri like MeSH, to frameworks like BLAST. The Life Detection Knowledge Base is a broader-scale attempt with a different focus. However, even repositories using the same ontologies frequently have incompatible or poorly implemented APIs for automated data retrieval and interaction. This proliferation of incompatible standards may only make it more difficult in the long run to assemble the necessary breadth of unified data needed for effective and efficient ML work.

Recommendations: **A working group of subject-matter experts from across the discipline could set the scope and implementation plan for an astrobiology-wide ontology and API standard**. It would have to encompass descriptors for microbiological, plant, animal, ecological, fossil, rock, sediment, chemical, water, air, geological, mineralogical, planetary, orbital, and astronomical systems as well as the diversity of instruments and techniques used to measure them. It would also need to be robust to cross-disciplinary terminology differences, such as *species*, *particle*, *cell*, *plant*, *habitat*, *quadrant*, *mesoscale*, *cloud*, or *matrix*. This is not a trivial task. It could be assisted, though not automated, by text-based AI tools capable of rapidly parsing the existing literature and databases and recommending commonalities [10]. Its adoption could then be supported through providing a list of repositories and archives compatible with it.

Beyond organizational recommendations, there is also substantial work to do in standardizing and unifying existing bodies of astrobiology-relevant data to make them more useful to the community. Subject matter experts with data science support are best placed to do this work; however, such efforts do not easily fit into existing ROSES solicitations. **A new ROSES call, similar to PDART but scoped to include all astrobiology-relevant data as well as standardization and API improvements**, would provide much-needed support for work with existing data; alternately, or in parallel, **a dedicated line in existing ROSES programs for making data open science and AI-ready** would support for future data.

2. *Getting what we don't have: data breadth and coverage gaps*

Observations: Most research work performed on non-living systems uses different handling methods, techniques, scales, instruments, and assumptions than on living systems, even when the description of the methods is similar. For instance, geological samples are commonly baked out to remove volatiles before analysis, but biological samples are not. Aseptic technique, or lack thereof, hampers cross-project use of physical samples. Very few samples have been assessed with both biological and physicochemical techniques outside of work where biosignature or prebiotic signature detection was an explicit goal. Satellite imagery is a counterexample only because the entire Earth's surface has been mapped several times for independent applications– but even in this data, temporal, resolution, scale, and instrument differences can be confounding [11].

Similarly, field expeditions rarely, if ever, take the same sets of measurements across environmental, biological, and physicochemical contexts as each other. This lack of simultaneous, comparable measurements for biological and non-biological systems is compounded by the low level of detail provided in most published protocols, where phrases like "spectra were baseline corrected" or "instrument artifacts were removed" are common. Examples of data and metadata frequently not covered in OSDMPs include duplicate field samples for archiving alongside those to be analyzed; coupons or readings from intermediate steps in analysis; as-run protocols in sufficient detail to allow replication; raw instrument files; and any processing steps, whether manual or scripted, to reproduce the "final" data from the raw instrument readings.

Recommendations: The overhead of assembling and standardizing past data can be reduced by **developing clear and common AI-ready standards for data generators and principal investigators** (PIs). **A working group could establish a recommended minimum set of context measurements and metadata to be taken in all future astrobiological field work or sample analysis**. A more expansive solution would be **a "library" available for checkout of core equipment for field measurements, including mission-analog instruments** – such arrangements are now usually made through networks of individual PIs, which creates barriers to entry and exclusivity. In parallel, **a wiki-like collection of standard protocols in sufficient detail to allow replication** (a level beyond what is included in most publications) could be created for proposing PIs to reference and future PIs to be required to contribute to. AI/ML tools may even be able to use such a collection to help identify ways to improve comparability among various projects. Lastly, **repositories recommended for use should be verified to provide support for physical samples and both raw and processed data, as well as specific protocols or scripts** for all processing that takes place.

3. Improving access: unique samples and instruments

Observations: The best data often comes from work with astrobiological flight-analog instruments, which explicitly target biological or habitability-related signatures. However, access to these instruments is usually tightly controlled and saturated with mission-critical experiments. Additionally, many do not approve work with biological materials for fear of contamination. This significantly limits the community's ability to provide support to planetary exploration efforts and to take advantage of the large investment that developing and ground-truthing such instruments requires.

Guidelines and options for archiving physical samples from field analogue sites are also particularly lacking. The Astromaterials Data System and the NASA Biological Institutional Scientific Collection are very specific in their coverage. There are no dedicated archives for field samples, despite the expense and effort associated with planetary analogue expeditions. Synthetic laboratory materials and most biological materials (derived cell cultures, DNA extracts, etc.) are rarely even addressed in OSDMPs.

An enormous wealth of expensive, valuable, and difficult-to-obtain field samples, derived materials, images, and other data languishes at the bottom of various lab freezers and on old hard drives, and has been and will continue to be lost when the responsible PI passes away, moves institutions, or simply loses funding or interest.

Recommendations: One mitigation could be **a requirement in instrument development calls to provide a "community access" version, which would then be archived for long-term access under a proposal model similar to that used for astromaterials**. Though this would require an additional funding line, the community version could be a significantly cheaper build (not miniaturized, not hardened, etc.) as long as it produces equivalent data. Alternatively, there could be **a requirement to produce sufficient documentation to allow other researchers to build an "open-source clone" of the core functionality** where appropriate commercial-off-the-shelf parts exist. At a minimum, **a public list of field samples collected through funded work and contact information** should be created to which future PIs must add their entries. This could be accompanied by **a set of recommended protocols for long-term storage of different sample types**. Providing these resources would enhance the science return of these instruments by allowing more scientists to replicate results, make improvements, or conduct new foundational research.

4. Making it happen: lowering barriers to implementation with streamlining and support

Observations: The best recommendations are useless if they aren't followed. More researchers than will admit choose what to write in their OSDMPs based on a last-minute Google search for "example of a data archive". This leads to an over-reliance on general repositories and archives such as Github and Zenodo, even though more specialized ones with better support and integration may exist. Combined with the aforementioned lack of standardization in tagging, formatting, and APIs, the end result is that the final data products end up difficult to find and difficult to use, undoing much of the intended benefit.

There is also simply a lack of expertise and resources. Preparing data for upload, documenting protocols, aliquoting samples for storage all take substantial time and effort. However, they are not treated equivalently to "getting publications out" or even "getting the final report in".

Recommendations: Providing researchers **a list of self-identified astrobiology-specific repositories and archives** would be a good start. The development of an astrobiology ontology and a minimum standard of metadata would make it further possible to generate **a list of "certified" repositories and archives** compatible with it. While changing the paradigm for open science credit entirely is beyond programmatic reach, **providing equal acknowledgment, publicity, and outreach for data releases** would be a start. In a best-case world, there would be one or more **program staff data scientists available to work with the PI after award to approve the details of OSDMP implementation and provide support** for the actual upload and release process.

References

[1] van der Maaten, L. and Hinton, G. 2008. *Journal of Machine Learning Research*.

[2] McInnes, L. et al. 2018. *Journal of Open Source Software*. DOI:10.21105/joss.00861.

[3] Lamm, S.N. et al. 2025. *56th Lunar and Planetary Science Conference*.

[4] Valizadegan, H. et al. 2022. *The Astrophysical Journal*. DOI:10.3847/1538-4357/ac4399.

[5] Cleaves, H.J. et al. 2023. *Proceedings of the National Academy of Sciences*. DOI:10.1073/pnas.2307149120.

[6] Pasterski, M.J. et al. 2025. *Journal of the American Society for Mass Spectrometry*. DOI:10.1021/jasms.4c00300.

[7] Buckner, D.K. et al. 2025. *56th Lunar and Planetary Science Conference*.

[8] Clough, L.A. et al. 2025. *Earth and Space Science*. DOI:10.1029/2024EA003966.

[9] Wilkinson, M.D. et al. 2016. *Scientific Data*. DOI:10.1038/sdata.2016.18.

[10] Bhattacharjee, B. et al. 2024. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, EMNLP 2024*.

[11] Warren-Rhodes, K. et al. 2023. *Nature Astronomy*. DOI:10.1038/s41550-022-01882-x.