4D Workshop Abstract, Carnegie Institution for Science, June 4-6, 2018, Washington DC

Steps Toward Improved Integration, Search, and Analysis of Heterogeneous Data in the Astrobiology Habitable Environments Database

Richard M. Keller[1], Angela M. Detweiler[2], Barbara Lafuente Valverde[3], David F. Blake[4], Thomas F. Bristow[4], George W. Cooper[4], Christopher E. Dateo[4], David J. Des Marais[4], Linda L. Jahnke[4], Michael D. Kubo[5], Niki Parenteau[4], Leslie E. Prufert-Bebout[4], and Nate Stone[6]

NASA Ames Research Center
Intelligent Systems Division[1] Bay Area Environmental Research Institute[2] Universities Space Research Association[3] Exobiology Branch[4] SETI Institute[5] Open Data Repository[6]

The Astrobiology Habitable Environments Database (AHED) is a new data system being developed as a long-term, open-access repository for astrobiology data. AHED is intended to store user-contributed results from NASA or externally-funded research in astrobiology, and to encourage sharing and synergy within the astrobiology community. However, the interdisciplinary nature of astrobiology presents some specific challenges to data management, integration, and analysis within AHED. In some disciplines (e.g., genomics), open databases thrive because the contributed products are fairly uniform and standardized (e.g., sequence data). In astrobiology, each investigation produces a unique set of data products; this makes it difficult to search across different datasets to find similar data, or to combine results from separate investigations.

With AHED, we are taking steps to ensure there is adequate metadata – both at the dataset and record levels – to facilitate search, integration, and analysis. At the dataset level, we are developing a new metadata standard for describing astrobiology datasets, with detailed information about content, funding source, and scientific relevance, along with a set of topical keywords for characterizing datasets. At the record level, we are encouraging users to provide more structured content and finer-grained metadata. In many user-contributed science data repositories, few restrictions are placed on the uploaded data format, and minimal or no record-level metadata is required; thus users are unburdened when it comes to data preparation. The tradeoff is that deep integration and search across datasets is almost impossible without standardized structures and metadata. Although AHED users are free to upload minimally-described datasets, they will be encouraged to use database authoring tools (supplied by the underlying platform – Open Data Repository's *Data Publisher*) plus a set of customizable astrobiology-specific templates to help structure their data and provide standardized metadata. In reward for their extra effort, AHED will be able to deliver enhanced search, discovery, and analysis capabilities.