Storage of Physical Sample Metadata in the Astrobiology Habitable Environments Database (AHED)

Rich Keller, Ph.D.
Intelligent Systems Division
NASA Ames Research Center

Physical Samples and Digital Libraries Workshop
June 22-23, Newark NJ USA
Outline

• Introduction
  – What is Astrobiology?
  – What is AHED?

• Motivation
  – Data Sharing in Astrobiology
  – Perceived benefits

• AHED Project background
  – Open Data Repository
  – Data sharing problems

• AHED Metadata Standardization Effort

• Challenges for data sharing
Astrobiology is **MULTIDISCIPLINARY in content and INTERDISCIPLINARY in its execution**

("Astrobiology is MULTIDISCIPLINARY in content and INTERDISCIPLINARY in its execution" (The NASA Astrobiology Roadmap, Des Marais et al., 2008))
What is AHED?

Astrobiology Habitable Environments Database

– A repository containing scientific datasets from multiple astrobiology project teams

– Use cases:
  • As a science team’s private repository
  • As a repository for sharing data with other specified individuals or teams
  • As a field-wide repository for sharing data with the entire community
  • As an educational outreach portal

– (But can it serve all these purposes??)
The AHED Team

– Consolidated group of astrobiologists from different active research teams at NASA Ames Research Center
  • Brad M. Bebout, Leslie E. Bebout, Thomas F. Bristow, David J. Des Marais, Angela M. Detweiler, Michael D. Kubo, Barbara Lafuente, Niki Parenteau

– Assisted by:
  • data scientist: Rich Keller
  • database developer: Nate Stone

– Funded by:
  • NASA Science-Enabling Research Activity (SERA) Project of the NASA Science Mission Directorate
Motivation for AHED Development

- Federal government has a responsibility to share data gathered with taxpayer money (Office of Science and Technology Policy Memo: February 22, 2013)
  - NASA has increasingly required internal and external PIs applying for NASA funding to formulate data management plans.
  - Uneven application of these requirements:
    - Planetary science missions (yes)
    - Earth science missions (yes)
    - Human exploration missions (mixed)
    - Aeronautics (no)
    - Astrobiology Program (no)
- Premise: synergy and information sharing propels the science forward
  - Inspired by successes with genome databases, biodiversity databases, mineralogy databases, others
Data sharing in Astrobiology

Benefits of data sharing/archiving:
- Reanalysis of data to verify results
- Reinterpretation of data with a different approaches
- Data integrity and preservation
- Eliminates data redundancy
- Training tool for future researches

Barriers to data sharing/archiving:

<table>
<thead>
<tr>
<th>Barrier</th>
<th>Number of Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>People don't need them</td>
<td></td>
</tr>
<tr>
<td>Insufficient time</td>
<td></td>
</tr>
<tr>
<td>Lack of funding</td>
<td></td>
</tr>
<tr>
<td>No place to put them</td>
<td></td>
</tr>
<tr>
<td>Lack of common standards</td>
<td></td>
</tr>
<tr>
<td>They shouldn't be available</td>
<td></td>
</tr>
<tr>
<td>Don't have the rights</td>
<td></td>
</tr>
<tr>
<td>Don't have technical skills &amp; knowledge</td>
<td></td>
</tr>
<tr>
<td>Sponsor doesn't require it</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
</tr>
<tr>
<td>Not complete/finished</td>
<td></td>
</tr>
</tbody>
</table>

(n=194)

[Aydinoglu et al., 2014]

AHED
A Platform Form Sharing Astrobiology Datasets

[Aydinoglu et al., 2014]
AHED Pilot Databases

PLRP - Pavillon Lake Research Project DB
- 536 records
- Microbialite and water samples

Direct molecular evolution sequence DB
- 43 records
- Sequence data from in-Vitro evolution experiments

CheMin DB
- 12 records
- Data from CheMin instrument (MSL-Curiosity)

ES Culture Collection DB
- 9 records
- Data from isolated cyanobacteria and heterotrophs

CROMO: Serpentinizing System DB
- 106 records
- Drill samples from serpentinizing systems

Lipid Biomarker DB
- 2 records
- Lipids from pure cultures of microbes
AHED Databases built upon Open Data Repository

• A scientific data repository system developed by the University of Arizona
• Allows scientific teams to design, develop, and deploy web-based data repositories
• Simple drag-and-drop database template authoring & web layout, coupled with Excel-based data uploading capabilities
Database Templates

- Each template encodes the set of metadata fields to be stored in an AHED database record.
Problems with AHED

- Each AHED database developed independently
- Even though some similar types of data are being stored in the different AHED databases, it is not possible to:
  - Search across databases
  - Discover data in other databases

Lack of standardization in:
- Field naming
- Data typing
Standardization issues mirrored in Astrobiology data cataloguing practices

• Astrobiology researchers conduct both field-based and laboratory-based research, during which physical samples are collected, processed, and catalogued.

• great disparity in practices employed by different teams or individuals

• no specific standards available to guide the collection and recording of astrobiology sample data.
AHED Metadata Standardization Effort

Develop standardized:

• Template types
• Template field naming and field datatypes
• Template metadata values

Standardized Metadata Model
Template Types

• Metadata previously mixed in a single template was segregated into multiple logical groupings describing:
  – Site
  – Sample
  – Measurement/Observation
  – Image
  – Instrument
  – Culture
  – Person
  – Organization
Metadata Model

- Image
- Site
- Sample
- Culture
- Measurement/Observation
- Organization
- Person
- Instrument

Relations:
- hasImage
- collectionSiteFor
- hasMeasurement
- measured
- imagedBy
- collectedBy
- culturedBy
- memberOf

Example:
- Instrument measured Using Sample
- Sample culturedBy Culture
- Culture hasImage Image
Template development process

• Weekly meetings to develop consensus on template definitions

• For each data field, must define:
  – Field name
  – Field type
    • Text, integer/float, choice, file, pointer
  – Field Inclusion Status
    • Required
      – sample type
      – sample label
      – sample collection date
    • Recommended: desirable
      – sample collection time
    • As needed: project-specific needs
      – sample collection method
  – Choice values
    • e.g.: site characterization: lacustrine, marine intertidal, marine, coastal, open ocean, hot spring, arid, hyperarid, cave, well, hypersaline, estuarine, evaporite, mine, subsurface, deep subsurface, acid mine drainage, riverine, spring, poza
Challenges to Data Sharing

• Astrobiology is broad and multi-disciplinary
• Teams studying many different phenomena
• Teams collecting many different types of samples
• Will one team’s very specific data be useful to another?
• How to standardize data collection so that it will be of value across such heterogeneous teams?
• How to motivate scientists to share data when it may advance a competitor’s research?
• Need to develop consensus with broader Astrobiology community!