Title: *CIF21 DIBBs: Beyond Data Discovery: Shared Services for Community Metadata Improvement*
Award ID#: 1443062
Principal Investigator: Ray Habermann, The HDF Group
Co-Principal Investigator: Matt Jones, National Center for Ecological Analysis and Synthesis

Metadata is a fundamental part of the research documentation, reuse, and collaboration process that helps ensure that science data are discoverable, accessible, understandable and, ultimately, reproducible and trustworthy. Diverse communities in geoscience, biology, hydrology, ecology, and biogeochemistry use many different metadata standards and dialects for describing their datasets, such as the Ecological Metadata Language (EML), the Federal Geographic Data Committee Content Standard for Digital Geographic Metadata (FGDC CSDGM), Attribute Convention for Data Discovery (ACDD), and standards developed by the International Organization for Standardization (ISO). In addition, community members are faced with a formidable array of recommendations about what needs to be included in metadata and what standard(s) should be used.

Historically, many scientific communities have addressed questions about metadata recommendations and representations by creating community-specific "standards" and then making recommendations about how they should be used. This approach may be expedient, but is typically results in interoperability barriers between communities. The familiar "not invented here" syndrome only exacerbates these barriers. We have separated metadata recommendations from the community standards (termed dialects) they are associated with by identifying documentation concepts included in the recommendations and mapping these concepts across dialects. This approach opens up cross-community communication about shared documentation needs regardless of the dialects that are being used in each community.

Creating mappings with a useful level of detail was our first challenge. It is easy to get buried in the weeds. We are publishing information about concepts, recommendations, dialects, and crosswalks between them on the ESIP Wiki (http://tinyurl.com/hfg6ve4) and describing our results during the upcoming AGU and ESIP meetings. Our mappings are starting to be used by collaborators on this project for creating new mappings aimed at their specific documentation needs.

We have developed several tools for evaluating metadata records and collections in multiple dialects for completeness with respect to various recommendations. An Excel dashboard facilitates comparison of metadata collections (up to ~1000 records) to various recommendations (the Recommendation Analysis Dashboard (RAD)). This dashboard presents four different views of the results: 1) Dialect Suitability compares the dialect used in the collection and the recommendation to identify conceptual gaps between them, 2) Signature Scores identifies groups of records that are missing the same elements, 3) Results Summaries show concept occurrence percentages in the collections, and 4) Concept Guidance provides links to information in the wiki. We are currently sharing this dashboard with collaborators and working with them to increase metadata completeness.

Second, we are developing a set of web services for evaluating records and collections remotely. These evaluations are built around sets of specific checks developed initially by our team, and eventually partners, in several common programming languages. These checks are generally very specific (e.g. a dataset title must exist and must have more than 7 words), and are implemented for multiple dialects (initially EML, ISO, DIF, and ECHO). An evaluation engine uses selected sets of checks (recommendations) to evaluate metadata and data and presents the results in an easily understandable and informative web interface. The evaluation engine is being developed for modularity and flexibility and will soon be ready for tests with collaborators. The first live implementation is currently available at the NSF Arctic Data Center (http://arcticdata.io) and implements the Arctic Data Center recommendations against metadata records written in EML.