Title: Standardizing sample information to support efficient tracking, data integration, and reuse in Environmental Systems Science

Joan Damerow[1], Deb Agarwal[1], Kristin Boye[2], Eoin Brodie[1], Shreyas Cholia[1], Hesham Elbashandy[1], Kim Ely[3], Amy Goldman[4], Valerie Hendrix[1], Christopher Jones[5,6], Matthew Jones[5,6], Zarine Kakalia[1], Kenneth Kemner[7], Annie Kersting[8], Kate Maher[9], Nancy Merino[8], Fianna O'Brien[1], Zach Perzan[9], Emily Robles[1], Cory Snavely[10], Patrick Sorensen[1], James Stegen[4], Pamela Weisenhorn[7], Karen Whitenack[1], Mavrik Zavarin[8], and Charuleka Varadharajan[1]

[1]Lawrence Berkeley National Laboratory, Berkeley, CA; [2]SLAC National Accelerator Laboratory, Menlo Park, CA;
[2]Brookhaven National Laboratory, Upton, NY;
[4]Pacific Northwest National Laboratory, Richland, WA;
[5,]National Center for Ecological Analysis and Synthesis (NCEAS), Santa Barbara, CA
[6]DataONE, Santa Barbara, CA
[7]Argonne National Laboratory, Lemont, IL
[8]Lawrence Livermore National Laboratory, Livermore, CA
[9]Department of Earth Systems Science, Stanford University, Palo Alto, CA
[10]National Energy Research Scientific Computing Center (NERSC), Berkeley, CA;

Contact: (JoanDamerow@lbl.gov)
Project Lead Principle Investigator (PI): Deb Agarwal, Charuleka Varadharajan
BER Program: CESD Data Management
Project: Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE)
Project Website: https://ess-dive.lbl.gov/
Project Abstract: Physical samples are foundational entities for research in earth and environmental sciences; they are not only the basis of individual studies but could also be integrated with other data to inform new and broader-scale questions. Data contributors to the Department of Energy's (DOE) Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE) repository often work in large, interdisciplinary teams and send samples to multiple facilities for analyses. This community needs an efficient system for persistent sample identification to enable efficient sample tracking, data integration, and reuse.

We therefore conducted a community pilot test on the use of persistent identifiers for physical samples– specifically, International Geo Sample Numbers (IGSNs). Eight projects with a variety of sample types registered samples for IGSNs, standardized sample collection metadata, published sample metadata in the System for Earth Sample Registration (SESAR) sample catalog and have or will publish related datasets in ESS-DIVE. We compare existing sample-related standards and evaluate the experience of users to develop practical recommendations for sample identification and documentation. We gathered information for the pilot test through discussions with project teams and documented several components, such as the efficiency of the process (i.e. use of templates, labeling, registering samples, and updating metadata) and any apparent problems. We resolved uncertainties in the allocation of related sample identifiers, use of metadata elements, and added standard terms as needed. Throughout the pilot test, we also gathered feedback on desired use cases, which include: improvements in data management, advanced search capabilities, ability to link identifiers, promote interoperability of biological and geological samples, and ability to integrate and reuse sample data.

The pilot test has informed community-driven standards and tools for sample identifiers, tracking, and metadata in the ESS-DIVE repository. Our overall goal is to provide practical recommendations for efficient sample data management while also preserving and maximizing the potential value of samples into the future.