Summary Report

Climate and Environmental Data Listening Session | July 26, 2024

The Climate and Environmental Data Listening Session of the BIOFAIR Data Network Project held on July 26, 2024 was led by BIOFAIR Data Network Steering Committee members Barbara Thiers (New York Botanical Garden), Dori Contreras (Perot Museum of Nature and Science), David Kunkel (Oklahoma State University), and Anna Monfils (Central Michigan University Herbarium), in collaboration with Key Domain Representatives Ben Halpern (National Center for Ecological Analysis & Synthesis or NCEAS) and Ty Tuff (Environmental Data Science Innovation & Inclusion Lab or ESIIL).

Among the 32 session participants were representatives from the American Institute of Biological Sciences (AIBS), the University of Maryland's GEDI Lab, the National Science Foundation (NSF), the Environmental Data Initiative, The University of British Columbia Faculty of Forestry, Cooperative Institute for Research In Environmental Sciences (CIRES) Earth Lab at the University of Colorado Boulder, Michigan State University, University of Illinois Urbana-Champaign, Climate Hazards Center at the University of California Santa Barbara, University of Wisconsin-Madison, U.S. Geological Survey (USGS), Florida International University, Association of Ecosystem Research Centers, Central Michigan University, National Oceanic and Atmospheric Administration (NOAA), National Institute of Standards and Technology (NIST), HydroSHEDS/Confluvio, USGS Southwest Climate Adaptation Science Center, NCEAS, ESIIL, and Montana State University Bozeman, and the Biodiversity Collections Network (BCoN).

Summary

The Climate and Environmental Data (CED) Listening Session brought together data providers and users whose products and research have a primary focus in the CED data domain to discuss challenges and potential solutions to making different environmental data more interoperable, as well as the interfacing of these data with biological collections. The session began with an overview highlighting work of the Biodiversity Collections Network with a particular focus on the National Science Foundation (NSF) funded BIOFAIR Data Network project (Award No. 2303588). Additional presentations throughout the program included overviews of using CED in biological research and data integration efforts at NCEAS and ESIIL. The program also included discussions regarding the definition of common terms (environmental data and interoperability), the challenges associated with the interoperability of different environmental data sources and their initial development, the integration of these data sources with biological collections, and a hypothetical scenario about constructing a common

dashboard where a wide variety of environmental data would be freely accessible globally to address critical questions.

Key Topics Discussed

Introduction to the BIOFAIR Data Network Project

The session started with an introduction to the BIOFAIR Data Network project, aimed at fostering collaborations towards an integrated open data network that can expand the use of biological collections for research and education. As a conceptual framework for data integration, BCoN has promoted the extended specimen network initiative, which aims to make biodiversity data more accessible for a wider range of research projects through linkage to related biological data. The BIOFAIR Data Network project requires greater engagement with data communities beyond biological collections.

Scope of Environmental Data

The first discussion topic focussed on the definition and scope of environmental data. Participants argued that environmental data should include both abiotic and biotic components of the Earth system, as well as changes caused by human activities. It was suggested that environmental data is not defined by scale, but includes data that can be utilized to research a full spectrum of questions from broad global patterns to unique local phenomena. The value of sensory data in providing information on human activity and land use change was highlighted.The discussion concluded with the observation that the consensus among LTER sites was that data needed to develop research theories in other disciplines (e.g., social sciences) are not considered environmental data.

Data Interoperability

During a discussion about data interoperability, its future considerations, participants highlighted the importance of good quality metadata in current best practices. In cloud optimization efforts, streaming data is becoming more common now than downloading data for analysis, significantly reducing processing time and streamlining workflows. Participants also discussed the challenges of balancing diverse user needs with practical data sharing limitations, particularly in developing countries with bandwidth issues. The discussion concluded with the agreement on the need for greater integration between diverse data sets and improved user training and knowledge products.

Using Climate and Environmental Data in Biological Research

David Kunkel presented on the use of CED in biological research using his dissertation as a lens for discussing the topic. He opened with a broad overview of his work focused on characterizing the contemporary niche space of American *Asclepias*, more commonly known as milkweeds, using that information to understand the group’s evolutionary history, and predicting

how the distributions of these species may, or may not, shift under climate change scenarios. The presentation primarily focused on the contemporary niche space component by discussing the integration of species occurrence data and CED for answering important biological questions, what climatic and edaphic variables contribute most to the differentiation in niche space for species in this group, the utility of predicting species distributions, and utilizing these species distribution models for understanding how much species overlap in their niche space.

Interoperability of CED Data and Biological Collection Data

Participants were divided into two groups for discussions regarding what critical questions could be addressed by making CED and biological collection data more interoperable, the current impediments to meeting this goal, and ways that species occurrence data specifically could be made more accessible and interoperable with CED. Following the breakout sessions, the ideas brought forward by the breakout groups were summarized. The discussion by the participants primarily focused on current impediments of making CED and biological data more interoperable with a distinction between the social and technological aspects of this problem. The social aspects included whether people are actually utilizing the data and/or have the skills to use it appropriately, highlighting a greater need for training in this regard, as well as whether there should be a prioritization for the generation of data based on big questions currently being asked. Discussions in this realm also focused on the need for better metadata curation and an incentive for researchers to generate and contribute that metadata.In addition, there was some discussion about the utilization of AI to identify cutting edge questions in order to allow data providers to better tailor their data products and potentially justify their development to funding agencies more effectively. The technological aspect of this discussion focused mainly on the need for standards within and across data sources. There is also a need for infrastructure to maintain data sources long term. Participants also discussed a need for better tools for the locating, cataloging, and curating of data sources to make the wide variety of currently available data products discoverable and accessible to a broader audience of end users rather than only research scientists. One participant highlighted the need for species absence data as well as species occurrence data for monitoring and modeling projects. Finally, it was noted that the private sector is a major provider and user of environmental data, but many providers restrict use of their data by the private sector, and likewise there are restrictions within the private sector to the wide sharing of their data and analyses.

Examples of Integrating Environmental Data

Ben Halpern discussed the focus at NCEAS on combining disparate data to gain new insights into the environment and biodiversity. He highlighted examples of how social, economic, and cultural factors can influence the collection and representation of data, and how interoperability can help leverage existing data to fill gaps and make inferences. Ty Tuff then presented on the work of ESIIL and its efforts of making data more interoperable. He highlighted several working groups within the organization focused on combining various data types to answer key questions in biology and environmental science. He emphasized the center's focus on

environmental science with computer science and their goal of promoting more data streaming and cloud-specific ways of working.

Unified Metrics Resource for Earth's Environment

Participants discussed a hypothetical scenario about the creation of a set of globally available data resources intended to link a wide variety of data sources describing aspects of Earth's abiotic and biotic environments. Discussion focused on suggestions for short-term (within 5 years) and long-term (10 years or more) goals for this initiative and how such a resource might be utilized.

Participants identified several short-term goals for the creation of this hypothetical resource:

1. Create equity for data providers and users at a global scale and address existing global inequities relative to technology, access to the internet, and training.
2. Compile a catalog of environmental data resources that could be leveraged to determine gaps in current data availability by focusing on gaps in the catalog and community needs more broadly. A gap analysis would reveal where the majority of our biodiversity information comes from and where it is utilized.
3. Promote good data standards, foster a culture of good data management practices, and promote the sharing of data.
4. Provide targeted training to allow users to fill data gaps and adhere to good management practices.
5. Prioritize resources based on projected impacts of climate change globally.

Long-term goals that were discussed included:

1. Connect ecosystem functions and processes that occur in different data domains.
2. Promote interdisciplinary research, particularly incorporating aspects of the social sciences when human impacts are a major component.
3. Utilize growth hacking strategies for ecological transitions.
4. Work toward the creation of digital twins for Earth's ecosystems.

Additionally, the discussion returned to the need for standardized protocols and metadata (within and across domains) for the integration of environmental and biological data across data sources. Participants discussed looking for commonalities in data storage across different taxa (plants, mammals, etc.), clear reporting of geodetic systems, and utilizing environmental data integration protocols and frameworks modeled after those in the public health sector. Participants asked about the role of AI in the short- and long-term goals of the hypothetical initiative. The consensus was that AI would surely play a role, although participants cautioned that the use of AI will not obviate the need to address the short-term goals.

Recommendations

- Environmental data providers should explore ways to make their data more widely accessible and interoperable. These could include:

    - Produce complete and standardized metadata with details on data cleaning, analysis, etc.
    - Share clean or augmented data back to data providers
    - Promote the establishment of data standards by domain type
    - Create resources to better describe data types and tools to integrate across domain-specific data types
- Environmental data users should cite data sources so that providers are aware of how their data is being used.
- Providers of species occurrence data should continue to identify commonalities across data storage methods, as well as provide a clear reporting of what geodetic system is used for the reported coordinates.
- Gaps in data availability should be identified to determine future directions for what data resources should be developed.
- Prioritize training in the use of environmental data in research.
- Hold future discussions about how to create a culture of data sharing and metadata preparation in science.
- Learn from public health data integration efforts to inform data integration efforts related to environmental data.