BIOFAIR DATA NETWORK

Listening Sessions Reveal Broad Consensus on Building an Integrated, Open, Findable, Accessible, Interoperable, and Reusable (BIOFAIR) Data Network: A Summary

This report summarizes discussions held during six virtual listening sessions hosted by the Biodiversity Collections Network (BCoN) as part of the National Science Foundation-funded project, Building an Integrated, Open, Findable, Accessible, Interoperable, and Reusable (BIOFAIR) Data Network (DBI Award No. 2303588), in the summer of 2024. The BIOFAIR Data Network project is designed to further one of the group's key initiatives, the Digital Extended Specimen (DES) Network (1). The project intends to create a forum for discussion among relevant data communities about building the infrastructure needed to support data initiatives such as the DES that seek to integrate a wide range of biological and environmental data into an expanded data network.

The BIOFAIR Data Network Listening Sessions (June – August, 2024)

The Listening Sessions represent the first phase of the BIOFAIR Data Network project. Increasing the value of biological specimens by creating durable linkages with derivative and associated data requires technical infrastructure solutions and the development of social norms and contracts as yet not fully implemented. The Listening Session engagement was intended as a first step toward creating an overarching data community to facilitate the sharing of biological and environmental data, including genomic, ecological, climate, geological, biodiversity, behavioral, and human health data.

These themed sessions brought together students, and emerging and established professionals from a wide range of backgrounds, expertise, and perspectives for two-hour discussions led by BCoN Advisory Committee members and key domain representatives who were selected because of their involvement in that domain. The themes were: Federal Agency Data (14 June), Genetic and Genomic Data (26 June), One Health Data (2 July), Ecological Data (July 12) Climate and Environmental Data (July 2), and Biodiversity Informatics (August 26). 199 people (besides the BIOFAIR steering committee) participated in the six listening sessions, representing 139 projects or institutions. (See the list of participants and the list of projects and institutions represented by the participants)

This report focuses on the areas of unity across the listening sessions; the wide range of perspectives shared during the Listening Sessions are documented in the individual session summaries accessible through the BCoN BIOFAIR Data Network website (2). Fundamentally, listening session participants indicated strong support for the vision of a globally accessible data network that would serve the data needs of a broad range of potential users and help improve and sustain the individual data resources that constitute the network. Participants recognized that significant barriers exist to the realization of this vision, however. Broadly categorized, these barriers are data access, discovery, standardization, insufficient training, and insecure data sources. Many of the recommendations to overcome these barriers will require the development of a collective impact model for FAIR (Findable, Accessible, Interoperable, and Reusable) biological and environmental data. A collective impact model is a structured approach to tackling complex social issues by bringing together different organizations to work toward a common goal (3). The perceived barriers and recommendations for overcoming them are described in detail below.

Barriers to Data Integration and Implementation of a BIOFAIR Data Network and Recommendations to Overcome these Challenges

Address equity in data access. Bandwidth and access issues in developing countries prevent data access and sharing. This means that we are not only missing data from climatically sensitive, biodiverse regions of the world but also excluding the participation of scientists in these regions from documenting the impacts of global environmental change. Participants across the six listening sessions assigned a high priority to addressing global inequities in access to raw and synthesized data. Essentially all of the recommendations outlined in this document would help to democratize data access. However, concerted efforts will be needed to ensure these recommendations are effective in all countries.

Establish standards and protocols for ensuring ethical use of data. Data access strategies must balance FAIR data principles with necessary data protections. We must ensure that legal and ethical considerations (e.g., data sovereignty, attribution, and permissions surrounding traditional knowledge) are incorporated into models for data sharing and integration across different data jurisdictions. For example, researchers and institutions should engage with indigenous communities and incorporate mechanisms like Local Context labels and notices (4) to recognize and protect traditional knowledge associated with data. Database providers should explore options for data embargoes or redactions to comply with regulations and protect sensitive information, and the community should hold further discussions around third-party ownership of genetic sequence submissions to allow for editing and correction in perpetuity.

Incentivize best practices. We must collect data following established protocols such that they can be integrated with other datasets and be used in studies with very different objectives than the study for which they were originally collected. Documentation of datasets should include the generation of complete and standardized metadata with details on data cleaning methods, analysis, and synthesized data storage, while clean or augmented data should be shared back to data providers. To help comply with such recommendations, we should create resources to better describe data types and tools to integrate across domain-specific data types. A roadmap or concept map of data integration efforts, including users, contributors, repositories, and data aggregators, would guide future data integration. A system of credit or recognition for submitting data and metadata that adhere to best practices might also incentivize more attention by researchers to this aspect of their work.

Database providers should develop mechanisms for data curation and annotation by institutions and communities without altering original data submissions and should give clear guidance on how to cite material from their collection in genetic repositories, including unique identifiers and linking protocols. Data users should cite data sources so that providers are aware of how their data is being used. Community and participatory science data and the complexities of these disciplines should be included in discussions about biological data integration.

Journals, publishers, and their editorial board members should establish common standards for citing specimen and sequence data in publications, including supplemental materials with specimen and sequence IDs, submitting authors, institutions, and usage details. Funding agencies should require that data and specimen management plans submitted with research proposals should include plans for data sharing, citation, and deposition in appropriate repositories, and should provide guidance to grant applicants on how to develop and budget for such activities. Data repositories such as the Barcode of Life Data System (BOLD, 5), the Global Biodiversity Information Facility (GBIF, 6) and the International Nucleotide Sequence Database Collaboration (INSDC, 7) should continue to develop user-friendly tools and templates to facilitate standardized metadata submission by researchers and should consider implementing translation layers among different data standards and formats. Ecological data communities, including the National Ecological Observatory Network (NEON, 8) and the Long Term Ecological Research Network (LTER, 9), in conjunction with the biodiversity informatics community, need to continue developing and adopting data and research metadata standards, enabling harmonization of data across networks and interoperability of ecological data repositories with biodiversity informatics systems.

Increase data availability. We are still missing the data needed for a comprehensive network of biological and environmental data. Gaps in the data identified by listening session participants include datasets that underlie research projects, especially small and focused datasets from individual researchers and monitoring programs, “non-standard” data types such as acoustic data, and data gathered in participatory or community science projects. Despite recent massive efforts to digitize biological specimens, participants noted that information on current (as opposed to historical) species occurrence data and species absence data are still needed. In some cases, data may exist but are not shared due to concerns about cybersecurity, embargos for publication priority, and/or over-sampling of sensitive species. In addition, many providers restrict the use of their data by the private sector, even though the private sector is a major provider and user of environmental data; similarly, there are restrictions within the private sector to the wide sharing of their data and analyses.

Participants suggested that institutions, projects, and individuals holding data resources should continue to prioritize the digitization of these resources as well as continue to improve these resources through augmentation and standardization. Community leadership, perhaps in the form of a sustainably funded data action center, could address data access barriers as described above, and ensure that a biological and environmental data network is compatible with existing infrastructure projects, for example, the Distributed System of Scientific Collections (DISSCo, 10), GBIF, and the Internet of Samples (iSamples, 11), and promotes interdisciplinary research, particularly incorporating aspects of the social sciences when human impacts are a major component of a research project.

Improve data integration*.* Currently, we do not have common language, ontologies, or data models that are flexible enough to support a BIOFAIR data network. Some components of published biological data are underutilized because they have insufficient or non-standardized metadata. We need to create resources to better describe data types and facilitate integration across domain-specific data types. Such a resource could take the form of a catalog of data resources that could be used to find needed datasets and also could be leveraged to determine gaps in current data availability. A gap analysis would reveal where most of our biodiversity information comes from, where it is utilized, and how the development of missing resources should be prioritized.

Data integration would be facilitated by enhancements to the Darwin Core model (12) such as the Humboldt Core Extension for Ecological Inventories (13) designed to improve the integration of ecological monitoring data (e.g., acoustic monitoring, camera trap, and animal movement data) and other extensions, including a single system of persistent, resolvable identifiers required for a Digital Extended Specimen architecture. Similarly, collaboration with emerging initiatives like the Biodiversity Digital Twin project (https://biodt.eu/, 14, 15) would be beneficial. Differences in methodologies and scale of data represent another barrier to integration, and we need clarification regarding applicable standards and limitations of historical versus newly gathered data and a plan for each as it relates to data types to make these fully interoperable. Researchers, database providers, publishers, and funding agencies, should explore the use of persistent identifiers such as Digital Object Identifiers (DOIs) for linking specimen data across different databases and repositories, considering the challenges of low- and middle-income countries by devising brokering services to provide the necessary linking infrastructure.

Provide adequate training*.* Participants in multiple sessions commented on the need for more robust education or training resources to ensure the effective use and maintenance of biological and environmental datasets. The quantity and breadth of interdisciplinary research are limited not only by data availability but also by a lack of skills for appropriate and innovative data use by potential users. Participants called for the development of human resources that focus on data sharing and integration. Research communities should prioritize hiring data managers, and they should provide training opportunities in data management skills as career advancement for current personnel. Likewise, educational institutions should prioritize data management and standards training for early career scientists. We need improved user training and knowledge products. We should also provide and disseminate the necessary attribution metrics for data collection and curation to highlight human participation in the data lifecycle.

Maintain existing data resources. Many key data resources lack a plan for sustained funding. Data resources that underlie research analyses and conclusions must remain available so they can be used to reproduce this research. Therefore, a biological and environmental data community must explore sustainable funding and infrastructure models for long-term data preservation and access through increased engagement with public and private funding sources. We should also take steps to unify the landscape of multiple aggregators such as GBIF, the Global Genome Biodiversity Network (GGBN, 16), iDigBio (17), the Ocean Biodiversity Information System (OBIS, 18), and Vertnet (19) to provide a simpler, more user-friendly experience while also exploring more efficient publishing mechanisms to link specimens to citations, e.g., Biodiversity Literature Repository (20), CETAF e-publishing EJT (21) Pensoft ARPHA publishing tool (22).

Next Steps

In addition to providing feedback to the listening session participants, this summary will inform the development of a preliminary roadmap for the creation of an integrated, open, and FAIR data network that will be further elaborated during a virtual workshop to be held in February 2025 as the second phase of the BIOFAIR Data Network project. This workshop will engage a subset of listening session participants plus additional relevant data community members to refine the roadmap, identify key intermediary objectives or milestones, and suggest impactful use cases for an integrated data network.

References

(1) Hardisty, A R, ER Ellwood, G Nelson, B Zimkus, et al. 2002. Digital Extended Specimens: Enabling an Extensible Network of Biodiversity Data Records as Integrated Digital Objects on the Internet. BioScience, Volume 72, pages 978–987. doi: 10.1093/biosci/biac060

(2) BCoN BIOFAIR website (https://bcon.aibs.org/biofair/)

(3) Flood J, M Minkler, S Hennessey Lavery, et al. 2015. The Collective Impact Model and Its Potential for Health Promotion: Overview and Case Study of a Healthy Retail Initiative in San Francisco. Health Education & Behavior. 42:654-668. doi:10.1177/1090198115577372

(4) Local Contexts https://localcontexts.org/#

(5) Ratnasingham S, P DHebert. 2007. BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). Mol Ecol Notes. May 1;7:355-364. doi:10.1111/j.1471-8286.2007.01678.x.

(6) Global Biodiversity Information Facility, GBIF (www.gbif.org)

(7) International Nucleotide Sequence Database, INSDC (https://www.insdc.org/)

(8) National Ecological Observatory Network, NEON (https://www.neonscience.org/)

(9) Long Term Ecological Research Network LTER (https://lternet.edu/about/)

(10) Distributed System of Scientific Collections (DISSCo)https://www.dissco.eu

(11) Davies, N. , J.Deck, E. C. Kansa, S. W. Kansa. 2021. Internet of Samples (iSamples): Toward an interdisciplinary cyberinfrastructure for material samples, GigaScience, Volume 10, Issue 5, https://doi.org/10.1093/gigascience/giab028

(12) Wieczorek, J., D. Bloom, R. Guralnick, S. Blum.. 2012. Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLOS One https://doi.org/10.1371/journal.pone.0029715.

(13) Sica Y, K Ingenloff , Y-M Gan, Z Kachian. 2022. Application of Humboldt Extension to Real-world Cases. Biodiversity Information Science and Standards 6: e91502. https://doi.org/10.3897/biss.6.91502

(14) National Academies of Sciences, Engineering, and Medicine. 2024. *Foundational Research Gaps and Future Directions for Digital Twins*. Washington, DC: The National Academies Press. https://doi.org/10.17226/26894.

(15) National Academies of Sciences, Engineering, and Medicine. 2024. Foundational Research Gaps and Future Directions for Digital Twins. Washington, DC: The National Academies Press. https://doi.org/10.17226/26894.

(16) Global Genome Biodiversity Network, GGBN https://www.ggbn.org/ggbn_portal/

(17) Integrated Digitized Biocollections, iDigBio https://www.idigbio.org/

(18) Ocean Biodiversity Information System (OBIS) https://obis.org/

(19) Vernet (http://www.vertnet.org/)

(20) Biodiversity Literature Repository (https://biolitrepo.org/)

(21) Bénichou, L, I Gérard, I, É Laureys, M Price. 2018. Consortium of European Taxonomic Facilities (CETAF) best practices in electronic publishing in taxonomy. European Journal of Taxonomy, (475). https://doi.org/10.5852/ejt.2018.475

(22) ArPHA Writing Tool https://arpha.pensoft.net/