BIOFAIR
DATA NETWORK

Summary Report

Biodiversity Informatics Listening Session | August 26, 2024

The biodiversity informatics-focused listening session of the BIOFAIR Data Network Project held on August 26, 2024, was led by BIOFAIR Data Network Steering Committee members Andrew Bentley, Dori Contreras, Mike Webster, and Cameron Pittman, in collaboration with Key Domain Representatives Tim Robertson (Global Biodiversity Information Facility or GBIF), Sharif Islam (DiSCCo), and Jose Fortes (iDigBio).

Among the 59 session participants were representatives from Agri-Food Canada, Atlas of Living Australia (ALA), Berkeley, Biodiversity Heritage Library, CNRI, CONABIO, DiSCCo, GBIF, Global Biodata Coalition, iDigBio, Naturalis, NEON, NIST, NSF, Plazi, Smithsonian, Specify Collections Consortium, Symbiota, US Geological Survey, Vertnet, Yale Center for Biodiversity and Global Change, and more.

Summary

The BIOFAIR Data Network project aimed to create an integrated data network for biological data, focusing on biodiversity informatics and the need for automated data integration. Participants discussed various ongoing projects, including the Extended Specimen Network, GBIF, DiSSCo, and the iSamples project, and the challenges they faced, such as interoperability, data ownership, and standardization. Participants also explored potential solutions, including developing a data model for digital specimens, using platforms like Zenodo for digital preservation, and the need for centralized support and collaboration.

Key Topics Discussed

Biodiversity Informatics Listening Session

Dori Contreras initiated the biodiversity informatics listening session as part of the BIOFAIR Data Network project, which aims to create an integrated data network for biological data. Andrew Bentley presented on the BCoN BIOFAIR Project and the Extended Specimen Network. He highlighted BCoN activities, key initiatives such as the Extended Specimen Network, and the importance of community engagement and international cooperation. The project, which involves six virtual listening sessions, aims to foster collaboration among data communities to create and manage an integrated data network, with specific deliverables including a statement of engagement and a roadmap. The current listening session focused on biodiversity informatics, aiming to explore existing tools and infrastructure and organize efforts for a synergistic, extended specimen infrastructure.

Integrating Data Into Digital Extended Specimen Architecture

Participants indicated that the most commonly discussed integrations with specimen data involve genetic information and climate or environmental data but that differences in methodologies and scale of data can be challenging. Other promising data integrations include living collections (zoos and aquaria) and ecological plot data. Relationships between specimen and species data to assist in visualization of e.g. pollination and virus transmission, and spatial metadata integrations to highlight biomes and intersections with protected areas were also highlighted as important. Integration at the science-policy boundary is equally important in informing policy of direct influence and application for biological data. They acknowledged that some of these integrations are already happening but are not yet seamless, discussed the need for more automated solutions to integrate data into a digital extended specimen architecture, and emphasized the importance of data integration. At the same time, participants suggested learning from current efforts in the humanities, publishing, and industry sectors. Representatives of publishers proposed integrating links from specimens to related publications and engaging with natural history libraries and publishers. One participant shared challenges their team faced while integrating archives with their biological collections database, Arctos, and highlighted the need for collaboration with stakeholders from different disciplines. Problems were also noted in interoperability across different data types and sources in academia and industry. The importance of operationalizing disparate data was also highlighted.

GBIF

GBIF showcased how specimens are managed and linked within their system. GBIF integrates data from various sources, including natural history institutions, projects liberating content from literature, and DNA-related databases. It is common for data related to the same specimen to be shared in GBIF from different sources. GBIF runs a daily data clustering process to identify and link related records, such that material citations, sequences, and specimens connect.

The core GBIF data structure has been focused on species occurrences; the record documenting the existence of an organism at a place and time, often representing the gathering event for a specimen. This data structure is inadequate to properly capture the more complex relationships between specimens, parts of specimens, the sampling protocols by which the specimen was collected and other measurements and media connected to a specimen. GBIF is in the process of revising the data model to address this better.

Another recent advancement in GBIF has brought the ability to attach collection descriptors for partially or non-digitized specimens to the collections registered in GRSciColl. This improves the ability to discover the existing collections that might be of interest even when the specimens are not yet digitized, by improving the richness of the descriptive metadata about the collection. Possible integration with the newly developed Latimer Core extension was also discussed.

DiSSCo

DiSSCo, the distributed system of scientific collections, aims to bring European natural history collections together into a unified digital infrastructure, working with nearly 200 institutions across 23 countries. The goal is to provide faster, global access to data and make it ready for new research. They highlighted the technical, data operational, and organizational aspects of the project, including shared curation and access policies. They then discussed the concept of a fair digital object for digital specimens. The project aims to create a data infrastructure for digital specimens. They highlighted the project's current status, including the publication of a data schema for community feedback and the assignment of persistent identifiers (PIDs) to digital specimens in collaboration with the DOI Foundation and DataCite.

iSamples

The iSamples project aims to create an aggregation platform for samples and specimen-based data, with a focus on interdisciplinary research. It aims to provide unified access to physical sample records across different collections through iSamples Central. They explained the core metadata model, controlled vocabularies, and approaches for linking samples to contextual data like publications and environmental datasets. The project has developed vocabulary models on GitHub and conducted user studies to improve usability. Future goals include supporting discovery for less curated collections and potential collaborations with other groups.

Biodiversity Data Action Center Community Help Desk

The group discussed the concept of a biodiversity data action center community help desk, with a focus on connecting data frameworks with human capacity. They explored existing repositories, APIs, and software that could facilitate this goal, including projects like BiCIKL (through their Biodiversity Knowledge Hub) and the NFDI4 Biodiversity project together with integrations with the Biodiversity Literature Repository, TreatmentBank, and Zenodo. Participants also addressed the challenges and potential solutions for improving their data collection and sharing processes, emphasizing the importance of standardized rules, best practices, and data accessibility. They agreed on the need to show appreciation for data providers and the importance of human participation while highlighting the need for support to encourage adoption. There was also a general realization that we needed to simplify the landscape of multiple aggregators (GBIF, iDigBio, OBIS, GGBN, Vertnet, etc.) and data end-points to provide a simpler more user-friendly experience and interface for sample use and discovery while also exploring more efficient publishing mechanisms to link specimens to citations (e.g. Biodiversity Literature Repository, CETAF e-publishing EJT, or Pensoft ARPHA publishing tool using XML.

Developing Data Models for Digital Specimens

The group discussed the need to develop and mature a data model for more accurate capture of digital specimen information. Both DiSCCo and GBIF are currently undertaking work to enhance the Darwin Core model for Digital Extended Specimens. The importance of linking data to collection information and the need for resources to implement processes such as merging duplicates and updating data was emphasized. Concerns were raised about the lack of a centralized platform for discussing ongoing projects, while a centralized, distributed system for data storage was proposed. The importance of digitizing and publishing specimen data more efficiently was highlighted while stressing the need to integrate data from various sources and consider issues of risk management, policy development, and data ownership. Collaborations with Digital Twin and similar initiatives were also mentioned. Major challenges, particularly social issues, data integration, and data ownership, were discussed, with progress on improving their data model to integrate different data types and scaling up their data production being highlighted. The need for a single system of persistent, resolvable identifiers as part of the Extended Digital Specimen architecture was highlighted as critical. DOIs are increasingly being touted as the identifier of choice given the existing infrastructure in place and obvious collaborative opportunities, however, the cost of these may be prohibitive for some.

Digital Preservation, Data Ownership, and Ecosystem

Participants discussed utilizing platforms like Zenodo for digital preservation and the need for centralized support from partners. Challenges in managing long-term storage and coordination as well as concerns about protecting community data's commercial value were raised. How do we ensure FAIR data while simultaneously supporting data sovereignty, attribution, and permissions surrounding traditional knowledge (CARE principles and Local Contexts labels)? Discussions covered legal protections (intellectual property rights, copyright), data ownership vs. data stewardship, balancing open and protected data, changing the economic model of biodiversity data, collaborations with biodiversity-aware legal counsel, and the idea of a data ecosystem where providers control access. Plans were made for a virtual workshop in the spring of 2025 to develop a roadmap for an open environmental data network. Of course, all of this requires sustainable, long-term funding and other support (personnel) for the extended digital specimen infrastructure to be realized and maintained, and participants discussed options and models for such infrastructure funding. The short-term, grant-funded model is not supportive of a long-term vision or sustainability. In a distributed system this would also require the individual components of the infrastructure to secure similar long-term funding. The community needs to do a better job of promoting what we consider critical infrastructure to ensure long-term viability and attract funding. Collaboration with the Global Biodata Coalition was mentioned in this regard while the possibility of the Action Center in the US taking over this role was discussed.

Recommendations

- Explore potential collaborations and integrations between existing data infrastructure projects (e.g., GBIF, DiSSCo, iSamples).
- The community needs to develop standardized best practices for data sharing and integration.

- The community needs to address the challenges of data ownership/stewardship, protection, and FAIR use in extended specimen networks.
- The community should explore sustainable funding and infrastructure models for long-term data preservation and access.
- The community should ensure that legal and ethical considerations for data sharing and integration across different jurisdictions are incorporated into any such infrastructure.
- The community should develop strategies for balancing FAIR data principles with necessary data protections.