Summary Report

Genetic & Genomic Data | June 26, 2024

The Genetic & Genomic Data listening session of the BIOFAIR Data Network Project was held on June 26, 2024 and led by BIOFAIR Data Network Steering Committee members Breda Zimkus (Museum of Comparative Zoology, Harvard University), Andrew Bentley (Biodiversity Institute, University of Kansas), John Bates (Field Museum), Mike Lomas (NCMA, Bigelow Laboratory for Ocean Sciences), and Nimanthi Abeyrathna (Clarkson University), in collaboration with Key Domain Representatives Conrad Schoch (National Center for Biotechnology Information, NCBI), and Kevin Kerr (Centre for Biodiversity Genomics, University of Guelph).

Among the 28 session participants were representatives from the American Institute of Biological Sciences (AIBS), Bigelow Laboratory for Ocean Sciences /NCMA, Kansas University Biodiversity Institute, Black in Genetics, Boston University, CDC, Centre for Biodiversity Genomics, Denver Museum of Nature & Science, EMBL-European Bioinformatics Institute, GFBio e.V. / Leibniz Institute DSMZ, Global Biodiversity Information Facility (GBIF), Global Invertebrate Genomics Alliance, Louisiana State University Agricultural Center Aquatic Germplasm and Genetic Resources Center, Memphis Zoo, NCBI/NLM/NIH, Pennsylvania State University, Phoenix Bioinformatics, UCLA/California Conservation Genomics Project, University of California Santa Cruz, University of Southern California, US Department of Agriculture (USDA)- Agricultural Research Service (ARS), The University of Texas at Arlington, and the Biodiversity Collections Network (BCoN).

Summary

Attendees discussed the development of an integrated data and sample network for biodiversity research, with a focus on genetic and genomic data and the involvement of various collections institutions, agencies, and other stakeholders. They explored the challenges and potential solutions for managing and standardizing data submission, sequence information, and data sharing, with a focus on the importance of spatial and temporal metadata and standardizing metadata for effective searching. Issues surrounding reporting and withholding data, particularly in relation to endangered species, invasive species, and threatened populations, were also discussed.

Presentations

1. Welcome (Breda Zimkus)
2. BIOFAIR Data Network: Building an Integrated, Open, Findable, Accessible, Interoperable, and Reusable (BIOFAIR) Data Network (Andrew Bentley)

Synopsis: The BIOFAIR project and the Biodiversity Collections Network (BCoN) is a partnership dedicated to promoting new uses of biological collections and their derivative data, including progress on the creation of an extended specimen network.

3. International Sequence Database Collaboration (Conrad Schoch and Joana Paupério)
   Synopsis: The INSDC, which includes partners from the US (GenBank), Japan (DDBJ), and Europe (EBI/ENA), is analyzing database usage and developing tools for better linking of sequence data with specimen information and has recently agreed on new requirements, such as the need for location data.
4. Barcode of Life Data Systems (Kevin Kerr)
   Synopsis: The Barcode of Life Data Systems (BOLD) facilitates the generation and application of DNA barcode data for species identification. The FAIRness of data on BOLD was reviewed, highlighting how sequences are identified with unique, persistent identifiers and coupled to rich metadata (Findable); data is stored in common file formats and a stable structure (Interoperable); public sequences are free of copyright and image data can be shared under a Creative Commons license (Reusable); and BOLD data can be freely accessed, though some platform operations will require a user account (Accessible).
5. Global Biodiversity Information Facility (Tobias Frøslev)
   Synopsis: The Global Biodiversity Information Facility (GBIF) provides free and open access to biodiversity data, including an experimental clustering algorithm used to identify likely related or duplicated records.

Discussion Summary

NCBI Metadata, Registration, and Data Verification

The group discussed the importance of spatial and temporal metadata in INSDC databases. Country-level information is now mandatory; more detailed information is encouraged, and reasons for not providing the country must be given. Data connectivity and access have been issues for the community, with participants expressing concerns about difficulties in linking data to individual specimens. The group agreed that the accuracy of the data depends on the submitter and discussed the challenges in verifying data authenticity. Institutions currently cannot correct or amend data submitted to NCBI, as the ownership and responsibility for changes lie with the original submitter. There was a concern raised that duplication of specimen data in these databases may cause problems with synchronization; records would become out of date when changes are made to the original specimen record (e.g., re-identifications, georeferencing). The participants suggested that it would be better to link to the original specimen record than duplicate the data.

Standardizing Data Submission and Sequence Integration

Participants discussed the challenges and potential solutions for managing and standardizing data submission and sequence information. Standardizing metadata is necessary for effective searching and accessibility, and FAIRness would be enhanced if communities could promote these standards. This is currently a social issue, but there is a need for a technological solution so that specimen information can be integrated rather than duplicated. The duplication of metadata (and the importance of data synchronization) in disparate databases was highlighted as a problem. Ideally, source data should be linked rather than duplicated to ensure that data

remains in sync. The discussion touched on the ongoing issues with specimen voucher information, and the need for a solution that allows for corrections without altering the original record. The creation of an international sample database was proposed, and the use of unique, global identifiers for samples would ensure consistency across databases. In addition, the establishment of an intermediary broker database to aggregate data might be a good solution because data management formats that work well for natural history museums might not apply equally well to many of the users of platforms like BOLD from other communities.

Facilitating Data Compliance

The importance of documenting metadata, especially the methods used for generating consensus sequences, and creating tools to facilitate compliance with community standards was discussed. The group agreed to further explore these ideas and continue the discussion on gaps and current infrastructure. It was noted that researchers should take advantage of GBIF's infrastructure to improve data quality and compliance, particularly with the Darwin Core standard. The group discussed incentives to encourage compliance among researchers, especially those from lower resource settings. Some institutions have a policy of not closing loans until submissions to NCBI are formatted correctly to ensure that sequences are identified correctly by researchers and can be linked from NCBI records to the collection management system. Better metadata could lead to more reuse and citations, and community standards for data citation. One of the Key Domain Representatives highlighted the challenge of error prevention without automation and human intervention.

Data Sharing Challenges

The issue of data sharing and reuse was discussed with participants. It is important that researchers share datasets that attract attention and citations, and there is a need for a more collaborative approach to promoting their work. The issue of data embargoing for compliance with regulations was raised, and the need to balance this with ensuring fair sharing of benefits was discussed. The group agreed on the importance of data being recorded in the database, even if embargoed, to maintain the integrity of the database. It was identified that data integration is as much a social issue as a technological one and that social incentives to share and integrate FAIR data may be useful. It was suggested that if incentives are offered to researchers submitting metadata, data acquisition might be more attractive. Unfortunately, no matter how many resources (e.g., training, templates) are provided for metadata acquisition, it is not enough to attract people to submit it. If scientists from lower resource backgrounds could get some public endorsement when they share data fairly, for example on their resume/ CV or in their publications, it may encourage them to share data more often. One of the participants also discussed the challenge of sharing data for the reproducibility of the analysis. As a scientific community, we should be able to carry on novelty in scientific research by reusing a data set that is available through publication.

Reporting Challenges and Data Solutions

Participants shared their experiences regarding reporting and withholding sensitive information, such as locations of endangered species, invasive species, and threatened populations. The group agreed on the importance of respecting the reasons for data withholding and suggested solutions such as fuzzy geolocation and the use of unique identifiers for indigenous data. Some

participants have used Local Context labels and notices to streamline access to restricted data while giving Indigenous communities agency over their data. There was also a discussion about adding identifiers for the controls (positive/negative) that researchers use to avoid confusion when reusing data. As a community, we need to think of ways by which we can share the benefits of the shared data with its contributors.

Funding Options

There are many challenges associated with funding the implementation of a global solution using DOIs, and it was acknowledged that not all may afford the cost. Potential solutions include establishing norms for data citation and implementing a Specimen Management Plan for funded research. The importance of proper citation was emphasized, and one participant noted that attribution in lower and middle-income countries could be addressed by establishing norms while working with the journals as a lower-cost solution. Concerns were raised about the National Science Foundation's inability to address funding issues in lower-middle-income countries (LMICs), suggesting that other countries could replicate their funding structures and initiatives. Other suggestions included creating modular training courses and engaging academic institutions to create a more effective pipeline. It was suggested that the National Science Foundation (NSF) provide guidance for other funding agencies because, for now, NSF is creating or adding data management plans more often with their guidelines. There are also websites that provide mechanisms for creating data management templates, like DMPtools.org and ezDMP.org. Allowing these 3rd party data management agencies to manage this can save time for researchers. Training students in data management is also effective and sustainable.

Recommendations for Implementation

- Funding agencies should require and provide guidance to researchers, especially students, on data and specimen management plans for funded research, including plans for data sharing, citation, and deposition in appropriate repositories, and the funding of these activities.
- Funding agencies and organizations should explore mechanisms for funding the minting of persistent identifiers (e.g., digital object identifiers or DOIs) for data objects, particularly in low- and middle-income countries as well as brokering services to provide the necessary linking infrastructure.
- Researchers need to provide accurate and detailed metadata when submitting sequence data, including location, time of collection, and specimen information. Alternatively, integration tools (APIs, etc.) should be developed to link this data to the original source to keep it up to date.
- Database providers (INSDC, BOLD, GBIF) should continue to develop user-friendly tools and templates to facilitate standardized metadata submission by researchers.
- Database providers should develop mechanisms for data curation and annotation by institutions and communities, without altering original data submissions. Further discussions around third-party ownership of sequence submissions to allow for editing and correction in perpetuity are encouraged.
- Journals, publishers, and their editorial board members should establish common standards for citing specimen and sequence data in publications, including in

supplemental materials with specimen and sequence IDs, submitting authors, institutions, and usage details.

- The community, including researchers, database providers, and publishers, should explore the use of persistent identifiers (e.g., DOIs) for linking specimen data across different databases and repositories.
- Database providers should explore options for data embargoes or redactions to comply with regulations and protect sensitive information.
- The community should identify incentives for data sharing and attribution, such as recognition for FAIR data submission on researchers' CVs, in publications and annual reviews.
- Institutions should provide clear guidance on how to cite material from their collection in genetic repositories including unique identifiers and linking protocols. They should also consider requesting co-submission privileges to allow for editing.
- Database providers and researchers need to develop guidelines and best practices for handling and reporting positive controls and potential false positives in metabarcoding and environmental DNA data.
- Researchers and institutions should engage with indigenous communities and incorporate mechanisms like Local Contexts labels and notices (https://localcontexts.org/#) for recognizing and protecting traditional knowledge associated with data.

A final workshop is scheduled for Spring 2025 to develop recommendations and a roadmap for a collaborative network. Conversations on the above recommendations need to continue because finding implementation solutions that work for all partners is a shared and evolving responsibility, specifically in the areas of data synchronization and the creation of an intermediary broker database to aggregate data. Continued collaboration between the major players in the landscape of genetic and genomic data will provide opportunities for solutions to these issues.