Linking Between Molecular and Biodiversity Data: A BiCIKL Perspective

Joana Paupério[‡], Vikas Gupta[‡], Vishnukumar Balavenkataraman Kadhirvelu[‡], Kessy Abarenkov[§], Wouter Addink[|], Donat Agosti[¶], Olaf Bánki[#], Josephine Burgin[‡], Marcus Ernst[¤], Tobias Guldberg Frøslev[«], Quentin Groom[»], Anton Güntsch[¤], Suran Jayathilaka[‡], Sam Leeflang[|], Urmas Kõljalg[ˆ], Joe Miller[˅], Guido Sautter[¶], Lyubomir Penev[¦,ʔ], Guy Cochrane[‡]

‡ European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, United Kingdom
§ University of Tartu Natural History Museum, Tartu, Estonia
| Naturalis Biodiversity Center, Leiden, Netherlands
¶ Plazi, Bern, Switzerland
# Catalogue of Life, Amsterdam, Netherlands
¤ Freie Universität Berlin, Berlin, Germany
« Global Biodiversity Information Facility, Copenhagen, Denmark
» Meise Botanic Garden, Meise, Belgium
ˆ University of Tartu, Tartu, Estonia
˅ GBIF, Copenhagen, Denmark
¦ Pensoft Publishers & Bulgarian Academy of Sciences, Sofia, Bulgaria
ʔ Institute of Biodiversity & Ecosystem Research - Bulgarian Academy of Sciences and Pensoft Publishers, Sofia, Bulgaria

Corresponding author: Joana Paupério (joanap@ebi.ac.uk)

Abstract

Molecular sequencing data generation is being driven by global and regional efforts to discover, understand and monitor biodiversity. To fully explore this data in biodiversity research we need a network of connected data resources, linking sequence data with natural history collections, taxonomy and literature. The BiCIKL project (Biodiversity Community Integrated Knowledge Library, Penev et al. 2022) has set the groundwork towards creating this network of linked data and fostering FAIR (Findable, Accessible, Interoperable and Reusable) practices in the biodiversity domain.

Connecting biodiversity and molecular data along the biodiversity research cycle requires a foundation of well-structured and rich metadata in the molecular sequence databases. Referencing the physical specimens is important as this provides context about the source of the material that was used for generating the molecular sequence data, including information about origin and species identification. To connect biodiversity and molecular data, we developed tools and workflows for improving and standardising metadata, federated searches and validations for specimen reference in sequence data, such as the SpASe tool, which enables the discovery of links between natural history collections and sequences, and the European Nucleotide Archive Source Attribute Helper API, which facilitates the construction of specimen attributes in a structured format. This work was done in close collaboration with DiSSCo (Distributed System of Scientific

Collections) and some biodiversity genomics projects (e.g. Biodiversity Genomics Europe, BGE).

Furthermore, we enabled community curation of biological source annotations such as specimen references in sequence data through the PlutoF platform and the ELIXIR Contextual Data Clearinghouse (Abarenkov et al. 2021, Balavenkataraman Kadhirvelu et al. 2022) and increased bidirectional linking from sequences in the European Nucleotide Archive (ENA) to collections, taxonomy and literature services (e.g., Plazi TreatmentBank, OpenBioDiv). We also worked closely with the community to enable the structured publication of environmental DNA data, promoting and engaging in the definition of standards and developing tools to facilitate data deposition and retrieval.

Overall, the project has contributed significantly to strengthen the connections between the biodiversity and genomics communities towards higher data integration and interoperability. Structured, enriched, accessible and linked sequence data will provide a strong foundation for the application of biodiversity knowledge in the response to global challenges, such as biodiversity loss, ecosystem change and food security. Beyond BiCIKL, we will continue our work as a community to promote a culture of FAIR linked molecular data, towards a fully integrated biodiversity knowledge ecosystem.

Keywords

sequence data, specimens, taxonomy, literature, FAIR, biodiversity community

Presenting author

Joana Paupério

Presented at

SPNHC-TDWG 2024

Funding program

The BiCIKL project received funding from the European Union's Horizon 2020 Research and Innovation Action under grant agreement No 101007492.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Abarenkov K, Zirk A, Põldmaa K, Piirmann T, Pöhönen R, Ivanov F, Adojaan K, Kõljalg U (2021) Third-party Annotations: Linking PlutoF platform and the ELIXIR Contextual Data ClearingHouse for the reporting of source material annotation gaps and inaccuracies. Biodiversity Information Science and Standards 5: e74249. https://doi.org/10.3897/biss.5.74249
- Balavenkataraman Kadhirvelu V, Abarenkov K, Zirk A, Paupério J, Cochrane G, Jayathilaka S, Bánki O, Lanfear J, Ivanov F, Piirmann T, Pöhönen R, Kõljalg U (2022) Enabling Community Curation of Biological Source Annotations of Molecular Data Through PlutoF and the ELIXIR Contextual Data Clearinghouse. Biodiversity Information Science and Standards 6: e93595. https://doi.org/10.3897/biss.6.93595
- Penev L, Koureas D, Groom Q, Lanfear J, Agosti D, Casino A, Miller J, Arvanitidis C, Cochrane G, Hobern D, Banki O, Addink W, Kõljalg U, Copas K, Mergen P, Güntsch A, Benichou L, Benito Gonzalez Lopez J, Ruch P, Martin C, Barov B, Demirova I, Hristova K (2022) Biodiversity Community Integrated Knowledge Library (BiCIKL). Research Ideas and Outcomes 8: e81136. https://doi.org/10.3897/rio.8.e81136