# Towards Connecting Molecular Data and the Biodiversity Research Community: An ENA and ELIXIR biodiversity community perspective

Joana Paupério[‡], Josephine Burgin[‡], Toni Gabaldón[§], Jerry Lanfear[|], Robert M Waterhouse[¶], Guy Cochrane[‡]

‡ European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, CB10 1SD, United Kingdom
§ Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Barcelona, Spain
| ELIXIR, Wellcome Genome Campus, Hinxton, Cambridgeshire, United Kingdom
¶ Department of Ecology and Evolution and Swiss Institute of Bioinformatics, University of Lausanne, Lausanne, Vaud, 1015, Switzerland

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

## Abstract

Global and regional efforts for generating molecular sequencing data are fundamental to characterise and monitor the Earth's biodiversity. However, exploiting the full potential of molecular data for biodiversity monitoring and conservation is still a challenge. There is still the need to fully connect the generation and archiving of sequence data with other biodiversity infrastructures, thereby promoting Findability, Accessability, Interoperability and Reusability (FAIR) of data.

Here we present the ongoing activities and future plans of the European Life-Science Infrastructure (ELIXIR) and the European Molecular Biology Laboratory European Bioinformatics Institute's (EMBL-EBI) European Nucleotide Archive (ENA, the European node of the International Nucleotide Sequence Database Collaboration - INSDC) towards an enriched set of sequence data connected to the wider biodiversity research community.

ELIXIR has an emerging Biodiversity Community that was originally created as a focus group in 2019, to better align the work in biodiversity across the ELIXIR Nodes and with global initiatives in the biodiversity domain. This group has been working on understanding the capabilities, interests and ongoing projects that exist across the Nodes, developing connections with external partners in the biodiversity area (e.g. Global Biodiversity Information Facilitiy, GBIF; LifeWatch Eric) and developing a longer term strategy for support of biodiversity by ELIXIR. A recent opinion piece by the group ( Waterhouse et al. 2021) highlights opportunities for infrastructure developments in the area of biodiversity and provides recommendations for closer integration of molecular

data with biodiversity research. These recommendations include the alignment of taxonomies across domains and the general adoption of standardized metadata.

ELIXIR and EMBL-EBI are involved in several biodiversity genomics initiatives, including the Earth BioGenome Project (EBP), the Darwin Tree of Life Project (DToL), the European Reference Genome Atlas (ERGA), and the BIOSCAN Europe, where support is being provided to data curation, submission and visibility and in the definition of standards for the associated metadata (e.g. Lawniczak et al. 2022). Moreover, EMBL-EBI is a partner of UniEuk, an initiative that is working towards building a flexible universal taxonomic framework for eukaryotes. ELIXIR and EMBL-EBI are also part of the Biodiversity Community Integrated Knowledge Library (BiCIKL), an Horizon 2020 project that is working towards establishing FAIR practices in the biodiversity domain, and thereby developing tools and workflows for connecting data along the biodiversity research cycle (Penev et al. 2022).

These projects and community efforts are contributing to improving metadata standards and pushing the development of tools and workflows to support enriched metadata and increased linkage with other biodiversity infrastructures. Overall, we need to continue to work towards a strong foundation of interlinked knowledge to be able to effectively respond to global challenges such as biodiversity loss and ecosystem change.


## Keywords

data linkage, data management, sequencing data


## Presenting author

Joana Paupério

## Presented at

TDWG 2022


## Funding program

This work was funded by ELIXIR, the research infrastructure for life-science data. BiCIKL project receives funding from the European Union's Horizon 2020 Research and Innovation Action under grant agreement No 101007492

## Grant title

BiCIKL - Biodiversity Community Integrated Knowledge Library

## Conflicts of interest

## References


- Lawniczak MN, Davey R, Rajan J, Pereira-da-Conceicoa L, Kilias E, Hollingsworth P, Barnes I, Allen H, Blaxter M, Burgin J, Broad G, Crowley L, Gaya E, Holroyd N, Lewis O, McTaggart S, Mieszkowska N, Minotto A, Shaw F, Richards T, Sivess LS (2022) Specimen and sample metadata standards for biodiversity genomics: a proposal from the Darwin Tree of Life project. Wellcome Open Research 7: 187. https://doi.org/10.12688/wellcomeopenres.17605.1
- 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
- Waterhouse R, Adam-Blondon A, Agosti D, Baldrian P, Balech B, Corre E, Davey R, Lantz H, Pesole G, Quast C, Glöckner FO, Raes N, Sandionigi A, Santamaria M, Addink W, Vohradsky J, Nunes-Jorge A, Willassen NP, Lanfear J (2021) Recommendations for connecting molecular sequence and biodiversity research infrastructures through ELIXIR. F1000Research 10: 1238. https://doi.org/10.12688/f1000research.73825.1