Beyond BiCIKL: Towards Building an AI-Assisted "Biodiversity Supergraph"

Lyubomir Penev[‡,§], Dimitrios Koureas[|], Quentin Groom[¶], Jerry Lanfear[#], Donat Agosti[¤], Ana Casino[«], Joe Miller[»], Guy Cochrane[^], Olaf Bánki[˅], Urmas Kõljalg[¦], Patrick Ruch[ʔ,ʕ], Anton Güntsch[¢], Jose Benito Gonzalez Lopez[ℓ], Patricia Mergen[¶], Joana Pauperio[৯], Tobias Kuhn[P], Nikos Minadakis[₳], Vincent Stuart Smith[₴], Christos Arvanitidis[₣]

‡ Institute of Biodiversity & Ecosystem Research - Bulgarian Academy of Sciences and Pensoft Publishers, Sofia, Bulgaria
§ Pensoft Publishers & Bulgarian Academy of Sciences, Sofia, Bulgaria
| Naturalis Biodiversity Center, Leiden, Netherlands
¶ Meise Botanic Garden, Meise, Belgium
# ELIXIR Europe, Cambridgeshire, United Kingdom
¤ Plazi, Bern, Switzerland
« CETAF, Brussels, Belgium
» GBIF, Copenhagen, Denmark
^ European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, United Kingdom
˅ Catalogue of Life, Amsterdam, Netherlands
¦ University of Tartu Natural History Museum, Tartu, Estonia
ʔ SIB Swiss Institute of Bioinformatics, Geneva, Switzerland
ʕ HES-SO, Geneva, Switzerland
¢ Freie Universität Berlin, Berlin, Germany
ℓ CERN, Geneva, Switzerland
৯ European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom
P Knowledge Pixels, Zurich, Switzerland
₳ FORTH- Institute of Computer Science, Heraklion, Crete, Greece
₴ The Natural History Museum, London, United Kingdom
₣ LifeWatch ERIC, Seville, Spain

Corresponding author: Lyubomir Penev (l.penev@pensoft.net)

Abstract

BiCIKL (**Bi**odiversity **C**ommunity **I**ntegrated **K**nowledge **L**ibrary) is a European Union (EU) Horizon 2020 project (2021–2024) building a new community of research infrastructures (RIs), researchers and other stakeholders, through improved access to interlinked, open and FAIR (**F**indable, **A**ccessible, **I**nteroperable, **R**eusable) biodiversity data along the biodiversity research cycle (specimens, sequences, taxon names, publications) (Penev et al. 2022). The project's 14 partners developed or substantially improved 16 tools and services currently in process of onboarding to the European Open Science Cloud (EOSC), presented in the FAIR Data Place (FDP) of BiCIKL's flagship product, the Biodiversity Knowledge Hub (BKH). The tools and data were used in Open Call projects, performed by research groups worldwide. A key achievement of BiCIKL is the establishment of several new bi-directional links between the participating RIs through shared and interoperable data standards and web services. The sustainability of the BiCIKL services and especially of the strong collaborative spirit developed through

the project will be ensured by a membership agreement for the BKH maintenance and further development.

The results of BiCIKL are diverse and tackle various aspects of the implementation of open science practices in biodiversity research. The project partners and external collaborators from the Open Call projects published more than 80 papers and conference abstracts (see the article collections in Penev et al. 2022a and Thessen et al. 2023), two policy briefs (Penev et al. 2024, Agosti et al. 2024), three Biodiversity Information Science (TDWG) symposia (2021, 2023, 2024), several videos and factsheets and other training materials, guidelines and best practice recommendations, and so on. In the special focus of BiCIKL was the extraction and liberation of data from the PDFs of several thousands of published biodiversity articles making it accessible and re-usable.

The new BiCIKL community proved to be successful in both technological innovation and long-lasting spirit of collaboration between biodiversity and genomics researchers, data repositories, RIs, publishers and other stakeholders.

Beyond BiCIKL, we envisage our work towards further integration and interoperability between data domains by embracing human-in-the-loop collaborations, enhanced by Artificial Intelligence (AI). The implementation of AI and Large Language Models (LLM) should be possible when considering an important condition: to understand the complexity of past, recent and future changes in biodiversity and natural environments the use of AI tools should be based on adequately curated, semantically structured and interlinked biodiversity data. We see this radical new step as a concerted community effort towards building a “Biodiversity Supergraph” (Fig. 1), understood here as a two-component ecosystem consisting of:

1. centrally orchestrated system of tools and services, and
2. distributed sources of transformed, semantically enhanced FAIR Linked Open Data, supplied by the partnering RIs.

The “Biodiversity Supergraph" will provide integration of the biodiversity data on a scale and operational level that has never been attempted before. It is key for the next decade, to enable a baseline of global, biodiversity-related information serving organisations, academia, industry and society.

Keywords

biodiversity informatics, data interorperability, data integration, Linked Open Data, knowledge graph

Presenting author

Lyubomir Penev

Presented at

SPNHC-TDWG 2024

Acknowledgements

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

Funding program

H2020-INFRAIA-2020-1: Integrating Activities for Starting Communities

Grant title

Biodiversity Community Integrated Knowledge Library (BiCIKL), Grant No 101007492.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Agosti D, Bénichou L, Casino A, Nielsen L, Ruch P, Kishor P, Penev L, Mergen P, Arvanitidis C (2024) Liberate the power of biodiversity literature as FAIR digital objects. Research Ideas and Outcomes 10 https://doi.org/10.3897/rio.10.e126586
- Penev L, Groom Q, Lanfear J, Koureas D (2022a) Towards interlinked FAIR biodiversity knowledge: The BiCIKL perspective. Research Ideas and Outcomes. RIO Collections 105. https://doi.org/10.3897/rio.coll.105
- 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 (2022b) Biodiversity Community Integrated Knowledge Library (BiCIKL). Research Ideas and Outcomes 8 https://doi.org/10.3897/rio.8.e81136
- Penev L, Groom Q, Casino A, Barov B (2024) Uniting FAIR data through interlinked, machine-actionable infrastructures. Research Ideas and Outcomes 10 https://doi.org/10.3897/rio.10.e126588
- Thessen A, Pinedo-Escatel JA, Oliveira J, Barbosa R, Tanalgo K, Borges PV, Aneesh PT, Saeedi H, Smith V, Penev L (2023) Linking FAIR biodiversity data through publications: The BiCIKL approach. Biodiversity Data Journal. https://doi.org/10.3897/bdj.coll.209

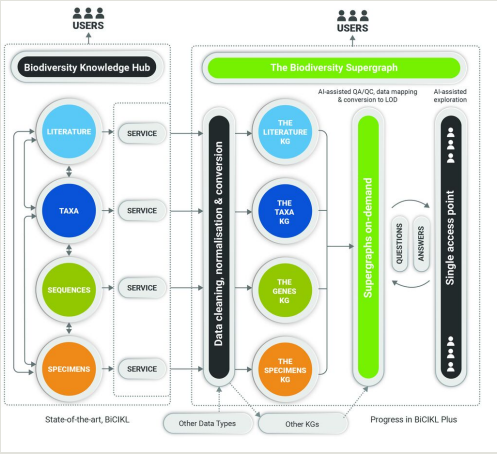

Figure 1.

Illustration of the transition from BiCIKL to the BiCIKL+ concept, showing the various knowledge graphs (KG) and services that will realise an advanced access to the biodiversity data and its reuse in the "Biodiversity Supergraph".