Uniting FAIR data through interlinked, machine-actionable infrastructures

Lyubomir Penev[‡], Quentin Groom[§], Ana Casino[|], Boris Barov[¶]

‡ Pensoft Publishers & Bulgarian Academy of Sciences, Sofia, Bulgaria
§ Meise Botanic Garden, Meise, Belgium
| CETAF, Brussels, Belgium
¶ Pensoft Publishers, Sofia, Bulgaria

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

Abstract

A new community of research infrastructures has joined forces to provide scientists with seamless access to the plethora of data, services and tools in biodiversity research. New levels of technological innovation and interoperability between infrastructures foster unprecedented access to biodiversity data across all data domains and the entire research lifecycle, thus advancing open science practices and strengthening Europe's position in the global biodiversity research landscape. This policy brief highlights the potential benefits derived from enhanced connectivity and interoperability among various types of biodiversity data, fostering innovation and advancements in biodiversity science, monitoring, conservation, and policy development.

Keywords

biodiversity, data, infrastructures, interoperability, FAIR, policy, access, standards, technology, collaboration, research, innovation, integration, open science, digital

Policy Context

The EU Biodiversity Strategy for 2030 frames the efforts to protect, preserve and restore biodiversity across the EU and beyond as the living-world pillar of the Green Deal. With the Global Biodiversity Framework (GBF) agreed at COP15, the ambition level has increased, so the need to support these goals with world-class research and innovation. The European Nature Restoration Law alongside the existing EU legislation (namely the Birds, Habitats, Water and Marine Framework Directives) include binding targets to be pursued, monitored and evaluated. They require precise data to underpin the design of effective measures for restoration and conservation. Access to such reliable and comprehensive data at a European scale is urgently required. Such data translates into evidence on which policy decisions must be taken. Furthermore, accessibility, FAIRness

and interoperability among data and knowledge holders are instrumental and rooted in open science principles. This data must meet high standards of integrity and reproducibility, so it can be reused beyond isolated initiatives and projects by high-level policy-making (Suppl. material 1).

This policy brief is addressed to: the European Commission's DG RTD & ENV; European Environment Agency; the Joint Research Centre; science and policy interface platforms such as EUBP; organisations and programmes (e.g. Biodiversa+, EuropaBON) engaged in biodiversity monitoring, protection and restoration; and to the Member States research funds.

Key advancements in the access to biodiversity data

The recommendations in this policy brief are based on the experience and the key advancements of the BiCIKL project (2021-2024, Penev et al. 2022) which resulted in (cf. the BiCIKL collection of all project outputs at: https://doi.org/10.3897/rio.coll.105):

- Bi-directional linking of biodiversity data among 15 world-class research infrastructures. Biodiversity data across disciplines and domains is now accessible through harmonised procedures for which the new **Biodiversity Knowledge Hub** (BKH) portal serves as a knowledge broker. The new services presented by the BKH are being registered in EOSC.

- BKH tools and services support **direct interoperability** across multiple infrastructures and cater for the wider research community and their specific scientific demands. This approach encompasses multiple infrastructures within and beyond the existing **ESFRI Roadmap**.

- The infrastructures provide access to data that is not only **FAIR** (Findable, Accessible, Interoperable, and Reusable) but also interlinked and capable of meeting computer-driven demands. This enhanced access is the result of technological innovations and protocols rigorously tested through open-call projects and by multiple users of virtual access facilities.

- Comprehensive **workflows for text and data mining** and **semantic publishing** unravel data from published and historical literature. This process includes annotation, semantic enhancement, quality assurance, and dissemination as FAIR data across knowledge brokers.

Recommendations to policy-makers

1. **Improved and widened access to biodiversity data**
    - **Support research infrastructures** to join BKH by ensuring compliance with the FAIR data criteria and following the interoperability guidelines developed for BKH.

- **Encourage** data managers to use interoperable infrastructures and avoid scientific knowledge creation in isolation.
- **Continue to promote open access to data** by making research infrastructures follow the established policy for open access to publicly funded research outputs, including publications, datasets and methodologies.

2. **International cooperation among infrastructures**
   - **Invest in** and **enable** European researchers and infrastructures to engage in global collaborations to support **interoperability with global research infrastructures**, and align operational frameworks.
   - **Support** the implementation of the **all-European system for persistent identifiers (PIDs)** for biological specimens through the alignment between the concepts of "digital specimen" (developed by the EU DiSSCo ES-FRI) and "extended specimen" (developed in the USA) as machine actionable FAIR Digital Objects within the International Partner Group for Digital Extended Specimen (IPDES).
   - **Strongly encourage** biodiversity infrastructures to continue operating and expanding under full compliance with the **international standards for taxonomic data** developed in the frame of the Biodiversity Information Standards Organisation (TDWG).
3. **Increased interaction of research infrastructures with industry.**
   - **Boost** the engagement of private companies such as IT developers and publishers dealing with biodiversity data and information. For example, the ARPHA Writing Tool publishing platform developed by Pensoft and the text and data mining tools and workflows developed by Plazi are widely used by several leading natural history institutions, journals and publishers. Any journal or publisher can use the Biodiversity Literature Repository (BLR) at Zenodo to deposit their articles and data. The tools and workflows developed in biodiversity genomics become part of the EMBL-EBI Industry Programme.
   - **Support** bi-directional exchange and work with the private sector and, more generally, the overall framework of EOSC expanding to the private and public sector activities.
   - **Promote further** engagement of the private sector in biodiversity research and conservation efforts through **incentives** for companies **to invest** in biodiversity-related research and data management.
   - **Encourage partnerships** between infrastructures and innovation companies including publishers, data management companies, and technology firms that can contribute to the dissemination and application of biodiversity data.
4. **Use of compatible data tools and services to EOSC**
   - **Ensure** the adoption of data interoperability standards and protocols as a condition to facilitate integration into EOSC of biodiversity data coming from diverse sources.

- **Foster** the endorsement of the EOSC principle of "Data as open as possible, as closed as necessary" to rely on data made open and FAIR through the services onboarded on the EOSC Marketplace.

5. **Contribution to other research areas and broader EU priorities**
   - **Support** linkage of data across domains through BKH services. Combined data from molecular biology resources will link to natural history collections, taxonomy and literature thanks to tools and workflows for accurate reporting of source annotations and facilitated curation of available data in collaboration with other biodiversity genomics projects.
   - **Create** strong technical and operational interfaces as in BiCIKL among infrastructures that operate in adjacent but different ESFRI domains (e.g. Environment, Health and Food).
   - **Encourage** cross-sectoral cooperation on the basis of data to support specific scientific needs of high priority.

Key action points

1. **Enhanced biodiversity research accessibility.**
   - The collaboration between 15 research infrastructures started by the BiCIKL project has significantly improved the accessibility of biodiversity data by establishing the **Biodiversity Knowledge Hub (BKH).** This hub enables seamless access to data, services, and tools, fostering open science practices and supporting Europe's position in global biodiversity research.
2. **Policy alignment with biodiversity goals**
   - The policy brief emphasises the alignment and contribution of technological advancements by BKH to the EU Biodiversity Strategy for 2030 and the Global Biodiversity Framework. It highlights the critical role of data, especially FAIR data, in achieving the legal obligations set in existing and emerging EU environmental legislation.
3. **Importance of interoperability and standardisaiton**
   - The recommendations call for continued integration and interoperability among research infrastructures. The adoption of common standards for persistent identifiers (PIDs) is crucial for seamless data discoverability and interoperability across various biodiversity domains.
4. **Role of technology in data processing**
   - The policy brief underscores the importance of technological innovations, including AI and semantic analysis tools, in unravelling and interlinking biodiversity data.
5. **Semantically enhance Linked Open Data (LOD)**
   - The integration among research infrastructures, tools and services they provide in the field of biodiversity data and knowledge will continue. The long-term vision towards which this process leads is a LOD-based, AI-assisted **Overarching Biodiversity Supergraph**: a single access point to

distributed data resources across multiple providers, knowledge domains and tackling data acorss their entire lifecycle. Such a knowledge graph provides curated and trustworthy data that is indispensable for the successful development and application of AI tools.

6. **Global collaboration in biodiversity research**
    - To strengthen European biodiversity research efforts, the policy brief recommends fostering international collaboration and aligning with global standards. The consultation with international partners and compliance with global standards for taxonomic data demonstrate the commitment to global cooperation.
7. **Engagement of the private sector in biodiversity research**
    - The policy brief advocates for stronger interaction between research infrastructures and private companies, emphasising the use of tools and workflows. Encouraging partnerships with IT developers, publishers, and technology firms is seen as vital for the management and use of biodiversity data.
8. **Integration with European Open Science Cloud (EOSC)**
    - The BKH and participating infrastructures aim to contribute to EOSC in line with the EU open science principles.
9. **Cross-domain collaboration for holistic insights**
    - The BiCIKL project demonstrates the potential for infrastructures operating in different ESFRI domains(e.g. Environment, Health, and Food) to collaborate effectively. **This cross-domain collaboration supports specific scientific needs, showcasing the broader impact of biodiversity research on multiple research areas and EU prioritites.**

Project identity

Project name: Biodiversity Community Integrated Knowledge Library (BiCIKL)

Coordinator: Prof. Lyubomir Penev, Pensoft Publishers, Bulgaria, l.penev@pensoft.net

Consortium:

- Pensoft Publishers (PENSOFT), Bulgaria
- Stichting Naturalis Biodiversity Center (NATURALIS), Netherlands
- Plazi GMBH (Plazi), Switzerland
- Agentschap Plantentuin Meise (MeiseBG), Belgium
- European Molecular Biology Laboratory (ELIXIR/EMBL-EBI), Germany
- European Organization for Nuclear Research (CERN), Switzerland

- Consortium of European Taxonomic Facilities (CETAF), Belgium and Muséum national d'Histoire naturelle (MNHN, associated party to CETAF), France
- Institut Suisse De Bioinformatique (SIB), Switzerland
- Tartu Ülikool (UTARTU), Estonia
- E-Science European Infrastructure for Biodiversity and Ecosystem Research (LIFEWATCH), Spain
- Freie Universität Berlin (FUB-BGBM), Germany
- Global Biodiversity Information Facility (GBIF), Denmark
- SPECIES 2000 (sp2000), United Kingdom
- Stichting International Working Group On Taxonomic Database (TDWG), Netherlands

Funding scheme

Call: Integrating and opening research infrastructures of European interest (H2020-INFRAIA-2018-2020)

Topic title and ID: Integrating Activities for Starting Communities (INFRAIA-02-2020)

Project: Biodiversity Community Integrated Knowledge Library (BiCIKL), grant agreement No 101007492.

Duration: 1 May 2021 - 30 April 2024 (36 months)

Budget: EU Contribution: € 4 995 158,50

Website: https://bicikl-project.eu/

Acknowledgements

The Biodiversity Community Integrated Knowledge Library (BiCIKL) project, funded by the European Union Horizon 2020 Research and Innovation Action under grant agreement No 101007492, has supported the publication of this work.

Conflicts of interest

The authors have declared that no competing interests exist.
Disclaimer: This article is (co-)authored by any of the Editors-in-Chief, Managing Editors or their deputies in this journal.

References

- 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 https://doi.org/10.3897/rio.8.e81136

Supplementary material

Suppl. material 1: Uniting FAIR data through interlinked, machine-actionable infrastructures

Authors: Lyubomir Penev, Quentin Groom, Ana Casino, Boris Barov
Data type: PDF file
Brief description: This file represents the policy brief in a PDF format, designed for printing and distribution. The file is published under Creative Commons CC-BY 4.0 license.
Download file (2.55 MB)