

Liberate the power of biodiversity literature as FAIR digital objects

Donat Agosti[‡], Laurence Bénichou[§], Ana Casino^l, Lars Holm Nielsen[¶], Patrick Ruch^{#,□}, Puneet Kishor[‡], Lyubomir Penev[«], Patricia Mergen[»],[^], Christos Arvanitidis[∨]

[‡] Plazi, Bern, Switzerland

[§] Muséum Nationale d'Histoire Naturelle (MNHN), Paris, France

^l CETAF, Brussels, Belgium

[¶] European Organization for Nuclear Research (CERN), Geneva, Switzerland

[#] HES-SO - HEG Geneva, Geneva, Switzerland

[□] SIB Swiss Institute of Bioinformatics, Geneva, Switzerland

[«] Pensoft Publishers & Bulgarian Academy of Sciences, Sofia, Bulgaria

[»] Meise Botanic Garden, Meise, Belgium

[^] Royal Museum for Central Africa, Tervuren, Belgium

[∨] e-Science European Infrastructure for Biodiversity and Ecosystem Research (LIFEWATCH) Biodiversity and Ecosystem Research - LifeWatch ERIC, Seville, Spain

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

Abstract

Knowledge about biodiversity is largely embedded in a daily growing corpus of over 500 million pages of biodiversity literature that is not machine-actionable. It is thus not open to building a biodiversity knowledge graph, or facilitating the use of artificial intelligence tools. This hinders the completion of a much-needed taxonomic name reference system, prevents the discovery of the biotic interactions underpinning the prediction and understanding of global change trends and consequences, viral spillovers, annotation of genes with their respective phenotypes, and their citations in various domains dealing with biological species such as conservation, agriculture, medicine, life sciences and industry, necessary to achieve the objectives of the Green Deal and address the targets identified in the Global Biodiversity Framework. This Policy Brief highlights key actions that can liberate the scientific data published, exploit their use, promote an enhanced way to publish, and ultimately foster excellence and innovation in biodiversity science, monitoring and conservation.

Keywords

biodiversity, literature, FAIR, data retrieval, publications, access, interoperability, make available, discover, knowledge, standards, integration, research, innovation, technology

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 there is a 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 and harmonised 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 information 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 by high-level policy-making, beyond isolated initiatives and projects.

This Policy Brief (Suppl. material 1) calls for urgent action to liberate data contained in non-machine actionable formats and non-interoperable platforms, and for that, is addressed to the policy actors at national and European levels, including 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 funders.

Key advances in accessing biodiversity literature

The knowledge about biological diversity is embedded in a daily growing corpus of over 500 million pages of scientific literature. Since Linnaeus's times (1753 onwards), each species description and later re-descriptions use the format of taxonomic treatments and contain dedicated sections of text providing data on all described species of the world. This includes, among others, the entire catalogue of life including synonyms, biotic interactions, distribution, traits and references to collection, and specimens used in subsequent studies. This structured text lends itself perfectly as input and training data for artificial intelligence applications.

An estimated 10% of the literature has been digitised by the Biodiversity Heritage Library (BHL), and a growing, but largely inaccessible, digital corpora is created by taxonomists and kept off-line mainly due to copyright restrictions. More than 50% of ongoing taxonomic publications are of closed access.

The EU-funded project Biodiversity Community Integrated Knowledge Library (BiCIKL) (2021-2024) (Penev et al. 2022; cf. the BiCIKL collection of all project outputs at: <https://>

doi.org/10.3897/rio.coll.105) created several key advancements to access and disseminate data in literature that resulted in new and innovative workflows, linkages and integrative mechanisms and services:

- The highly automated workflow in the TreatmentBank Research Infrastructure (RI) to convert Portable Document Format (PDF) documents into Findable, Accessible, Interoperable and Reusable data (open FAIR digital objects) was extended to include alternative input formats (HTML, XML) and more efficient tools to annotate and create bidirectional links; The FAIR data, taxonomic treatments, figures and material citations, preserved in the Biodiversity Literature Repository (BLR), a community in Zenodo, was enhanced with custom metadata referring to standard vocabularies, and bidirectional links;
- TreatmentBank, a dissemination mechanism was developed to dedicated RIs for taxonomic names (Synospecies, ChecklistBank and Catalogue of Life), specimens (Global Biodiversity Information Facilities, GBIF), genes (European Nucleotide Archive, ENA) and annotations (BiodiversityPMC at SIBiLS and OpenBiodiv);
- Connectivity between TreatmentBank and BLR was enhanced as a service to provide virtual access to FAIR data including already 90,000 publications including 910,000 taxonomic treatments, 560,000 figures and 1,510,000 material citations;
- Bulk-upload of large corpora (e.g. 19,500 articles in Taxodros covering all drosophilid taxonomy) of publications to BLR was deployed through the Lycophron tool;
- Persistent Identifiers (PIDs) of data liberated from publications and cited therein were implemented and bi-directional links using PIDs have been established in collaboration with the core data providers;
- The BiodiversityPMC, an extension of the PubMed Central full-text archive of biomedical and life sciences journals, was established to include biodiversity journals, supplements and taxonomic treatments and provide artificial intelligence and annotation tools to explore and enhance the content;
- The new ARPHA Writing Tool 2.0 was developed with enhanced semantic publishing and automated upload of FAIR data in publications to BLR, ChecklistBank, GBIF and ENA;
- An entirely RDF-based biodiversity knowledge graph (OpenBiodiv) was enhanced to include additional semantic mappings, automated RDF conversion workflows and user applications;
- Nanopublications were designed and implemented to allow publication of single scientific statements, data annotations and exchange;

- Open access to all results and data of BiCIKL was provided through publication in open access journals and collected together in a dedicated BiCIKL collection in the *Research Ideas and Outcomes* (RIO) journal;
- More than 15 scientific publications using linked open data and tools created in BiCIKL were published in a dedicated article collection in the *Biodiversity Data Journal*;
- Provenance to the source of taxonomic identifications using PID of the cited taxonomic concept were enabled through these advances, along with immediate access to the knowledge about this taxon, and the biodiversity knowledge graph extended with this identified digital object.

Recommendations to ensure data liberation

Based on the findings and developments accomplished under BiCIKL, important steps have been taken towards the exploitation, linkage and interoperation of different data types captured in biodiversity-related publications. Yet, the move shall remain and further improved beyond these initial achievements to ensure FAIRness and standardisation of the research data published as a way to connect, enhance and exploit the discoveries made, from individual to global scale. Scientific data, and data contained within publications shall be freely extracted, shared and reused once legally accessed. Several recommendations are highlighted here to provide diligent response to the challenge of interconnecting data on reliable and sustainable basis:

1. For biodiversity publishers and literature aggregators

- All publications should be open access and no restrictions for data mining should apply;
- All publications that are not available online should be uploaded to BLR, using a tool such as Lycophon, designed to support checking for DOI availability and bulk uploads;
- All publications should be made available in a machine actionable format (e.g. JATS XML);
- All publications should include persistent identifiers to cited materials and other sub-article research objects such as bibliographic references, figures, gene sequences, taxonomic names, treatments or specimens;
- All taxonomic publications should be converted into JATS Taxpub to feed into globally relevant services like ChecklistBank, Catalogue of Life, BiodiversityPMC and the GBIF;
- Workflows to automatically annotate and curate literature data including quality control by humans should be further developed;

- All publications in BLR converted to JATS should be submitted to biodiversity PubMed Central (PMC);
- A One Health library must emerge, either based on existing infrastructures (e.g., NLM's PMC, Biodiversity PMC, BLR) or others allowing automated data exchange;
- Search engines should be developed to explore supplementary data files.

2. For prospective publications

- Open Access publishing must become a norm to support open science. Authors and publishers should make copyrighted publications as accessible as possible by publishing under a CC BY licence or waiving copyright (CC0);
- Alternative workflows to produce XML-first based structured publications should be developed to cover differing needs of stakeholders;
- PIDs should be assigned to most important sub-article structural metadata and research objects and embedded in the article XMLs, to facilitate machine to machine interaction and save authors time to retrieve information;
- The citation frameworks should be extended with high precision evidence-based citations of research objects, such as treatments, specimens, taxonomic concepts, sequences and other data to allow semantically enhanced publishing.
- Alternative workflows to produce XML-first based structured publications should be developed to cover differing needs of stakeholders;
- PIDs should be assigned to most important sub-article structural metadata and research objects and embedded in the article XMLs, to facilitate machine to machine interaction and save authors time to retrieve information;
- The citation frameworks should be extended with high precision evidence-based citations of research objects, such as treatments, specimens, taxonomic concepts, sequences and other data to allow semantically enhanced publishing.

3. For legacy publications

- Enable authors to upload the original papers cited in their manuscript to BLR so that they could be processed for data extraction and linking;
- Digitisation of articles should follow Library of Congress standards optimised for text and image extraction to allow more efficient Optical Character Recognition (OCR);
- OCR tools should be further developed and enhanced to achieve a high level of accuracy;

- PDF based on scanned publications need to include adequate metadata (e.g., image size in mm / cm of the source document) to optimise the results of the data conversion output;
- Publications digitised retroactively must comply with FAIR formatting standards (e.g., JATS) to ensure interconnectivity;
- A minimal set of annotations to convert publications into open digitally accessible knowledge has to be defined and added to describe the domain-specific semantic content of biodiversity works.

4. For integration into sustainable Infrastructures and services

- Tools and workflows to convert legacy literature need to be further developed, integrated into stable infrastructures and supported financially to operate at scale;
- Repositories for FAIR data extracted from the literature, publications with rich metadata and links to other FAIR data sources (e.g. BLR at Zenodo, BiodiversityPMC, OpenBiodiv, Synospecies) need to be further developed, supported in the long-term and linked to permanent platforms.

5. For identification of digitisation priorities and strategies

- A stakeholder-driven approach should be used to define corpora of literature to digitise;
- Tools should be built to monitor progress in digitising as well the usage of digitised data for new scientific insights.

Drivers for policies in publications

The above recommendations set up the scenario where political actors should ground their action and lead to develop policies, standards and instructions on scientific publications that, once endorsed, will produce a harmonised, coherent and consistent landscape where scientific published data will be exploited to its full potential and produce impactful results in research and innovation.

To understand the full complexity of past, recent and future changes in biodiversity and natural environments, researchers, publishers, aggregators and repositories should align and jointly instruct the development towards an “Biodiversity Supergraph”, 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 be assisted by Artificial Intelligence (AI) tools, which however should be based on **adequately curated, semantically structured and interlinked biodiversity data**.

The bulk of this high-quality data comes from the peer-reviewed literature, therefore the future of biodiversity informatics and the building of the Biodiversity Supergraph should be supported by the adequate and endorsed frameworks and rooted on the following critical premises:

- The newly developed data extraction, annotation and dissemination workflows enable **immediate reuse of research data**, already in place with the GBIF, ChecklistBank, BiodiversityPMC, ENA OpenBiodiv, Synospecies, and others;
- Access to FAIR data allows the provenance of taxonomic names used to identify biological materials to be cited and improves evidence-based decision making and FAIR-ification of research, especially in **traceability, transparency and reproducibility**;
- Access to FAIR data provides unhindered use and **speeds up the research process** by avoiding tedious search for individual publications that are often closed access or hard to find;
- Annotated articles serve as training material for AI tools to increase the **precision and accuracy** of responses, allow citation of the source down to a statement and its context and through that to open up the huge corpus of biodiversity literature;
- **Nanopublications** already serve multiple purposes in biodiversity sciences, such as data publication, exchange and annotation in both machine-actionable and human-readable formats, and allow human-curated extension of the biodiversity knowledge graph;
- AI-powered literature services as in BiodiversityPMC enable factoid Question-Answering services, which should significantly enhance both **recall and precision of end-users information requests**;
- Natural Language Processing tools support the building of **large-scale biotic interaction networks**, which should broadly impact biological understanding;
- Bidirectional linking between otherwise disconnected data silos are delivering a **web of knowledge** and multiple ways to discover scientifically important statements and ultimately their source publications;
- **Scaling up the access and interoperability** will have an impact on the creation of new indicators for research assessment.

Liberating published biodiversity data and linking the various data types is critical to allow scientific research enhancement at the scale and the speed required in a challenging technological-driven environment, and becomes instrumental for innovations in data publications and access to facilitate the provision and use of reliable, harmonised and FAIR biodiversity information. This knowledge graph forms the basis to articulate a robust, evidence-based political framework in support of biodiversity.

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
- Sib 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: Liberate the power of biodiversity literature as FAIR digital objects

Authors: Donat Agosti, Laurence Bénichou, Ana Casino, Lars Holm Nielsen, Patrick Ruch, Puneet Kishor, Lyubomir Penev, Christos Arvanitidis, Patricia Mergen

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 licence.

[Download file](#) (5.06 MB)