ARPHA Preprints

PENSOFT

PREPRINT

Author-formatted, not peer-reviewed document posted on 29/10/2026

DOI: https://doi.org/10.3897/arphapreprints.e221240

Proposal for a federated landscape of curated reference libraries for enhancing aquatic (e)DNA biomonitoring in Europe

Filipe Costa, Frederic Rimet, Arne J. Beermann, Ioannis Kavakiotis, Jorge Emanuel Moutinho, Spiros Papakostas, Peter Woollard, Michal Grabowski, Emilie Boulanger

ARPHA Preprints *Author-formatted, not peer-reviewed document posted on 29/10/2026.* DOI: https://doi.org/10.3897/arphapreprints.e221240

Title: Proposal for a federated landscape of curated reference libraries for enhancing aquatic (e)DNA biomonitoring in Europe

Authors: Filipe O. Costa[1,2], Frédéric Rimet[3], Arne J. Beermann[4,5], Ioannis Kavakiotis[6], Jorge Moutinho[1,2], Spiros Papakostas[6], Peter Woollard[7], Michal Grabowski[8], Emilie Boulanger[9]

Running title: Federated curated reference libraries for aquatic eDNA

Affiliations:

[1]Centre of Molecular and Environmental Biology (CBMA) and ARNET-Aquatic Research Network, Department of Biology, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal
[2]Institute of Science and Innovation for Bio-Sustainability (IB-S), University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal

[3]Université Savoie-Mont Blanc, INRAE, CARRTEL, 74200 Thonon-les-Bains, France

[4]Aquatic Ecosystem Research, University of Duisburg-Essen, Universitätsstraße 5, 45141 Essen, Germany
[5]Centre for Water and Environmental Research (ZWU), University of Duisburg-Essen, Universitätsstraße 2, 45141 Essen, Germany

[6]Department of Science and Technology, School of Science and Technology, International Hellenic University, 57001 Thessaloniki, Greece

[7]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, United Kingdom

[8]Department of Invertebrate Zoology and Hydrobiology, Faculty of Biology and Environmental Protection, University of Lodz, Lodz, Poland

[9]Intergovernmental Oceanographic Commission of UNESCO, Ocean Biodiversity Information System, IOC Project Office for IODE, InnovOcean Campus, Jacobsenstraat 1, Oostende 8800, Belgium

Abstract

DNA and environmental DNA (eDNA) metabarcoding are transforming aquatic biodiversity monitoring in Europe. Still, the reliability of taxonomic assignments depends critically on the quality, completeness, and interoperability of the underlying DNA reference libraries. Here, we combine a landscape assessment of the current state of DNA reference libraries for aquatic organisms in Europe with a set of recommendations for a federated system of taxonomically curated reference libraries, following community consultation. We surveyed 26 major reference libraries across 40 attributes spanning taxonomic scope, metadata provision, curation practice, and technical accessibility, and complemented this with a questionnaire completed by 33 institutions actively generating new reference sequence data. Our assessment revealed a markedly heterogeneous landscape: databases vary widely in taxonomic and geographic coverage, molecular markers, and update frequency; most lack programmatic (API) access; and curation practice ranges from minimal taxonomic annotation to full phylogenetically-informed validation along with review of bibliographic references, with no collectively agreed minimum standard applied across databases or declared by data contributors. Building on these findings and on extensive stakeholder consultation within the eDNAqua-Plan project, we propose a federated system resting on three pillars: a minimum set of essential, interoperable metadata; standardised curation workflows culminating in a transparent, traceability-based quality ranking; and technical interoperability grounded in existing (meta)data standards (e.g., GSC MIxS, Darwin Core). We outline two feasible data-architecture options for implementation, together with a phased roadmap addressing governance and long-term sustainability. These recommendations aim to strengthen the reliability, reproducibility, and interoperability of eDNA-based aquatic biomonitoring across Europe and beyond.

Keywords: biodiversity monitoring; data curation; DNA metabarcoding; eDNA; FAIR data; interoperability; metadata standards; taxonomic assignment

Introduction

The critical role of reference libraries in (e)DNA-based aquatic biomonitoring

DNA and environmental DNA (eDNA) metabarcoding (hereafter we use eDNA metabarcoding as an umbrella term for both, unless specified otherwise) have revolutionised aquatic biodiversity assessments, providing a robust, high-throughput, non-invasive toolkit for detecting species and monitoring ecosystem health. This molecular approach is increasingly used in ecological research, conservation, and regulatory biomonitoring, such as under the EU Water Framework Directive (WFD), and can be pivotal for reaching multiple targets under the Kunming-Montreal Global Biodiversity Framework (GBF). However, the success of eDNA metabarcoding critically depends on DNA barcode reference libraries (Hollingsworth et al. 2026). These DNA sequence collections from taxonomically verified specimens are the indispensable backbone for identifying taxa in environmental samples. The completeness and reliability of these libraries directly impact the accuracy of taxonomic identification and the validity of ecological conclusions (Lavrador et al. 2023; Claver et al. 2023). As eDNA inferences increasingly inform policy, high-quality, comprehensive reference libraries are foundational for the consistency and credibility of DNA-based aquatic biomonitoring (Altermatt et al. 2025; Gauvin et al. 2025).

The diverse and mosaic landscape of DNA barcode reference libraries for aquatic organisms

The current landscape of DNA barcode reference libraries for aquatic organisms is marked by significant diversity in taxonomic scope, geographical coverage, data standards, and accessibility (Weigand et al. 2019). This heterogeneity stems from the vast biodiversity of aquatic environments and the historically decentralised nature of data generation across many countries, institutions and projects (Kaitetzidou et al. 2026). While large public archives like Barcode of Life Datasystems (BOLD) and INSDC (NCBI GenBank, ENA, DDBJ) are central hubs, many datasets reside in smaller, disparate repositories, often with unequal requirements for data quality, metadata, and curation protocols. The diversity of genetic markers used as barcodes, and adapted to the group of organisms studied, further complicates data synthesis. This fragmentation poses substantial challenges to developing a cohesive reference framework for standardised aquatic biomonitoring in Europe.

Pervasive issues of data quality, accuracy, and interoperability

Despite the growing volume of sequence data, significant concerns persist regarding the quality, accuracy, and interoperability of existing DNA barcode reference libraries. The reliability of eDNA taxonomic assignments hinges on the integrity of reference sequences and metadata. Deficiencies like misidentified vouchers, sequencing errors, contamination, inconsistent taxonomy, and incomplete metadata can lead to erroneous species detections, misguiding conservation and management (Radulovici et al. 2021; Keck et al. 2023). Biological complexities such as NUMTs, cryptic species, incomplete lineage sorting, hybridisation/introgression and high intraspecific variation are often inadequately represented in the libraries and further challenge their accuracy. The lack of universally adopted data and metadata standards hampers interoperability, i.e., the ability to integrate and exchange data stored in different reference libraries and repositories. For instance, an evaluation of European aquatic eDNA data repositories revealed substantial variability in metadata reporting and workflow documentation, demonstrating that practical alignment with FAIR (Findable, Accessible, Interoperable,

Reusable; Wilkinson et al. 2016) principles remains uneven across current platforms (Kavakiotis et al. 2026). While FAIR principles aim to address this, their comprehensive implementation is ongoing and vital for enhancing the trustworthiness of DNA-based assessments. Both the assessment by Kavakiotis et al. (2026) and the present paper are complementary outputs of the Horizon Europe eDNAqua-Plan project (https://cordis.europa.eu/project/id/101112800): the former addresses the wider eDNA data ecosystem, while the present paper focuses specifically on the landscape and curation of DNA reference libraries.

Central vs. local reference libraries: a shifting trend

The architecture of DNA barcode reference libraries involves dynamics between large, centralised international databases (e.g., BOLD, GenBank) and others focused on a particular subject (e.g., a particular geographical zone or biological group). While global platforms offer broad coverage, ensuring meticulous curation across all entries can be challenging. For instance, BOLD contains numerous species-labelled sequences that lack independent taxonomic verification, relying instead on automated reverse taxonomy to match new records against pre-existing Barcode Index Numbers (BINs; Ratnasingham and Hebert 2013) and species present in the database. Consequently, there is a growing trend towards locally curated reference libraries, managed by individual labs or national projects (e.g., Lavrador et al. 2023; Appleyard et al. 2025; Gauvin et al. 2025; Kaitetzidou et al. 2026) or international networks of scientists working on the same organism group (e.g., fungi: UNITE, Abarenkov et al. 2024; diatoms: Diat.barcode, Rimet et al. 2019). These offer benefits such as better representation of local genetic diversity and agile expert updates. Nevertheless, a critical shortcoming remains: despite the value of local curation, there is no concerted effort to connect this dispersed taxonomic expertise directly to centralised, global databases/platforms.

The imperative for assessment and strategic improvement in Europe

To fully realise eDNA metabarcoding's potential for harmonised aquatic monitoring in Europe, a pressing need exists to assess the current status, content, and workflow gaps within European DNA barcode reference libraries. This evaluation must examine taxonomic and geographic completeness, data quality, metadata consistency, and adherence to FAIR principles, identifying biases and deficiencies, particularly for under-sampled regions, ecosystems or taxa. Assessing data generation and management workflows—from vouchering to dissemination—is crucial to pinpoint bottlenecks and areas needing standardisation. Only through such a comprehensive stock-take can evidence-based recommendations for future improvements be formulated. These recommendations should also aim to guide strategic investments in targeted barcoding initiatives, foster enhanced collaboration among researchers, taxonomists, data scientists, and monitoring agencies, as well as advocate for the development of sustainable infrastructures and policies that support the long-term curation, accessibility, and utility of these invaluable reference resources for European aquatic ecosystems.

Aims

Accordingly, the primary aims of this paper are to: (1) assess the current status and landscape of DNA barcode reference libraries for aquatic organisms, with a European focus, including their taxonomic scope, geographical coverage, data standards, and accessibility; (2) identify and analyse key gaps and deficiencies within these reference libraries, particularly concerning taxonomic and geographic completeness, data quality,

metadata consistency, and adherence to FAIR principles; (3) evaluate prevailing workflows and approaches for data generation, storage, and accessibility in DNA-based aquatic biodiversity monitoring, and finally (4) propose concrete, evidence-based recommendations for a federated system of taxonomically curated DNA barcode reference libraries, establishing community-driven guidelines for essential metadata, curation procedures, and strategic investments to support harmonised eDNA-based aquatic monitoring across Europe and beyond.

Box 1: Glossary of database-related terms. Importantly, the definitions provided in this glossary are not mutually exclusive; for example, some databases may also function as repositories, or repositories may serve as archives.

Archive	A long-term preservation service for primary data submissions. Archives assign stable accession identifiers and retain the submitted records, for example ENA, GenBank, and DDBJ.
Repository	A service or platform that stores, manages, and provides access to data. A repository may function as an archive, but can also host derived datasets, metadata, software, or other research outputs.
Database	A structured, queryable set of records. A database may contain primary submissions, data harvested from archives, or curated and reformatted records; it does not necessarily provide long-term archival preservation.
Collection	A defined set of related specimens, sequences, records, or datasets assembled for a particular taxonomic, geographic, or scientific purpose. A collection may be held within a repository or database, or maintained locally by an institution or research group.
Reference library	A digital database where DNA sequence data and associated metadata are conserved to be used as a reference to assign sequences of unknown identity to a taxon (e.g., species, genus).

Material and methods

Two complementary approaches were used to gather data on DNA reference libraries for European aquatic biodiversity: i) existing reference libraries were surveyed (n = 26); and ii) a questionnaire was distributed to key actors involved in developing reference sequence libraries. The questionnaire was intended to capture data on ongoing projects not publicly available. By employing these two approaches, a more comprehensive overview of the current state of DNA reference sequence libraries was achieved.

Reference library survey

We compiled a list of the most relevant DNA reference libraries, accounting for 26 databases. We selected them to represent a wide range of taxonomic groups and molecular markers, with reference sequence data representative of European aquatic biodiversity. Although MARES is a bioinformatic pipeline rather than a reference library (Arranz et al. 2020), the test database built for its original publication was included in our analysis. We evaluated the libraries using a comprehensive set of 40 pre-established attributes covering database accessibility, metadata availability, data curation practices, taxonomic information, technical features (e.g., presence of Application Programming Interfaces [APIs] or web interfaces), and information on sequencing and analysis protocols (Table S1). For each database, we considered only the primary target markers, i.e., those for which the database was designed and curated; accessory markers represented by a limited number of records were not included.

Next, we took into account the taxonomic, geographic and molecular scope and coverage of the selected reference libraries, and evaluated to what extent reference sequence data were verified and curated. Data curation, as we understand it here, is the selection, validation, and organisation of reference sequence data to ensure its accuracy, consistency, and utility. We defined four non-exclusive categories to describe the different types of curation (Table 1). Type A refers to taxonomic annotation, specifically whether the taxonomic information and ranks associated with sequences were curated, including the use of standardised taxonomic nomenclature and the removal of records with invalid or ambiguous annotations such as 'cf., aff., sp., or spp.'. Type B corresponds to sequence quality control, involving filtering steps based on criteria such as sequence length, ambiguous base calls, chimeras, indels, or potential pseudogenes. Type C refers to expert or phylogenetically-informed curation, encompassing cases where taxonomic domain expertise or tree-based methods were applied to identify and remove or correct ambiguous or potentially erroneous records. Because libraries may have different combinations of the above categories, and implemented to varying extents, we also consider a Type D category that is attributed to those libraries that integrate the full set of requirements with full traceability, which can be considered the optimal procedure for quality assurance/quality control (QA/QC) and curation. On top of the other requirements, this category includes long-term curation of reference specimens and information. As reference libraries get complemented over time, new sequences with deviating taxonomic information can generate conflicts with existing ones due to misidentification of some specimens, young or hybridising species and the use of synonyms. Dynamically addressing these issues (e.g., re-verifying vouchers, checking for synonyms, analysing additional genetic markers) and documenting the outcome should be the final goal of curation.

Table 1. Data curation types to categorise reference libraries based on sequence quality control, taxonomic curation, and phylogenetic analysis.

Curation type	Defining parameters
A	Taxonomic annotation (e.g., nomenclature validation; standardised taxonomic ranks; removal of records without taxonomy)
B	Sequence quality filtering (e.g., length; ambiguous base pairs; chimeras, indels; pseudogenes)
C	Phylogenetic congruence analysis and/or domain expert validation

D	Cumulative implementation of ABC curation procedures with complete curation of all records down to species level; dynamic long-term, transparent, traceable, and accessible curation actions

DNA reference libraries questionnaire

To support the assessment, a structured questionnaire was designed and distributed to gather information on critical aspects of DNA reference library creation. The survey included 45 questions covering project scope, targeted biodiversity (taxa and ecosystems), specimen sources and identification methods, storage and curation of biological materials, molecular markers, sequence data deposition, and metadata management. The questionnaire was designed to be concise and accessible to encourage broad participation. Questionnaire dissemination was carried out through established European barcoding and genomic networks (e.g., iBOL Europe, Biodiversity Genomics Europe). It was supplemented by targeted outreach to researchers (n = 290) and institutions known to be active in relevant initiatives, from consultation of >2500 papers compiled from Web of Science. Only responses from projects addressing European aquatic biodiversity were considered for further analysis. For details on the questionnaire structure, see Costa et al. (2024).

Developing the recommendations

The survey and questionnaire results formed the basis for developing recommendations for a future landscape of interoperable, reusable, and reliable reference libraries founded on common curation principles and essential metadata. These recommendations were developed as part of the wider vision of improved data reporting and sharing for eDNA-based aquatic biodiversity monitoring in Europe, within the EU-funded eDNAqua-Plan project (https://ednaquaplan.com/). The recommendations were shaped by internal workshops and external stakeholder consultations that we organised or attended over the course of the project. We held an in-person workshop with parallel group discussions covering 1) the storage of vouchers, DNA material and DNA barcodes, data and associated metadata, 2) taxonomic curation procedures, and 3) machine-readable workflows and data standards for interconnectable libraries, on 12–13 September 2024, at the University of Lodz field station in Spała, Poland, gathering 29 participants from 10 European countries and Australia. The following year, on 28 February 2025, we organised an internal online workshop to document and evaluate the current eDNA data ecosystem in Europe at different levels, including reference libraries, considering the different databases and repositories, data flows between them, metadata requirements, standards and taxonomies used. These discussions defined the eDNAqua-Plan vision for a federated network of reference libraries, which was further developed throughout the project and stakeholder consultations. A final workshop, held online on 8–9 October 2025, which included 65 external stakeholders and other attendees from 17 European countries, the USA, and the Philippines, allowed us to refine the recommendations and priorities as well as evaluate their feasibility within the wider landscape and community. A detailed account of the different discussions and outcomes can be found in the respective workshop proceedings and landscape reports (Babo et al. 2024; Boulanger et al. 2025a, 2025b; Exter et al. 2026).

Results

Reference library survey

We assessed 26 databases, spanning a diverse range of taxonomic groups, molecular markers, and data structures (Table 2; detailed results available in Table S2). They typically harvest nucleotide data from global repositories such as GenBank (https://www.ncbi.nlm.nih.gov/genbank/; Benson et al. 2013) or the European Nucleotide Archive (ENA; https://www.ebi.ac.uk/ena/; Yuan et al. 2026), to which different levels of quality control and data curation can be applied. In addition, the harvested data can be complemented with newly collected sequences. The genetic databases include global primary data repositories, data quality-controlled databases, and databases with some level of taxonomic curation (Fig. 1). Of the assessed databases, 16 (61.5%) were one-time snapshots of the available data, while 10 (38.5%) are updated periodically. Nineteen repositories cover all aquatic environments, while six focus on marine waters, and one contains only freshwater records.

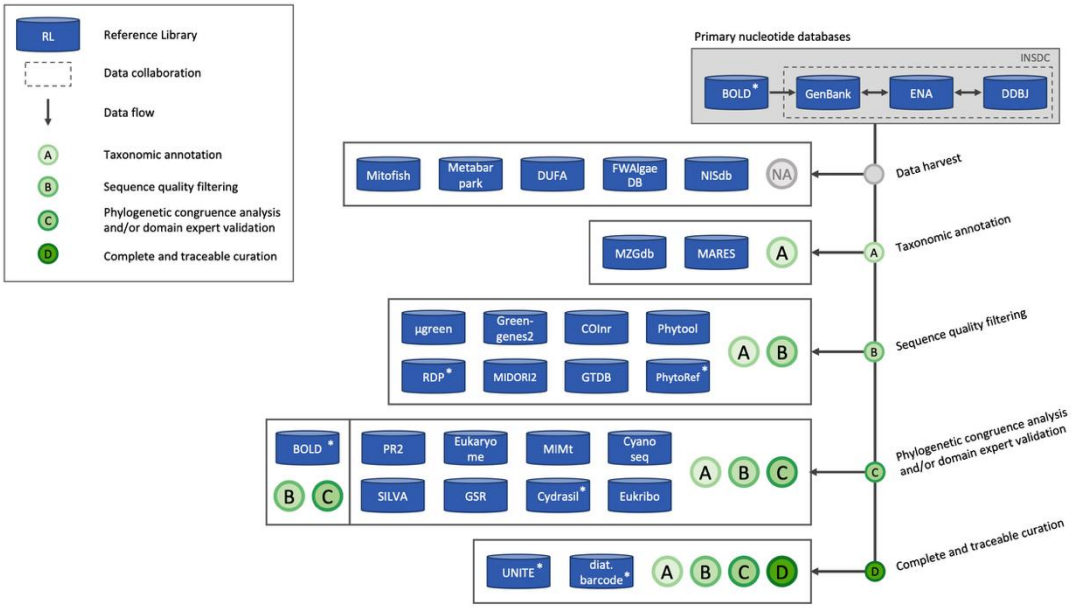

Figure 1. Overview of the current landscape of reference libraries for metabarcoding studies. Broadly, reference sequences are uploaded to primary nucleotide archives, including GenBank, ENA, and the DNA Data Bank of Japan (DDBJ), which continuously share data through the International Nucleotide Sequence Database Collaboration (INSDC; https://www.insdc.org/). From these, nucleotide data can be harvested to constitute (secondary) reference libraries. The asterisk* designates reference libraries that are both primary and secondary databases, providing original sequence data in addition to those harvested elsewhere. The reference libraries surveyed here were classified by the type of data curation that is applied upon data harvest, following the criteria and classification defined in Table 1.

The number of records ranged from ~200 for FWAlgaeDB (comprising complete genomes) to ~18 million (BOLD Systems; Table 2). Of the assessed libraries, six provide original primary-source reference sequence data (BOLD, UNITE, Diat.barcode, Cydrasil, PhytoRef, and RDP), although they incorporate records from additional sources. The remaining libraries consist exclusively of sequences retrieved from secondary sources. NCBI GenBank was the most common source of reference sequences, contributing data to 17 of the assessed databases (Table S2).

Nearly half of the databases (42%) store data on multiple molecular markers (n = 11). The most represented genetic markers across libraries were 16S rRNA, COX1, and 18S rRNA, present in eleven, nine, and nine libraries, respectively (Table 2; Table S2; list of acronyms of the markers available in Table S3). Notably, 16S sequences encompass both prokaryotic ribosomal genes and organelle-derived 16S rRNA (i.e., from mitochondria or chloroplasts) in eukaryotic contexts. Multi-marker databases included between two and five markers; however, two exceptions, MIDORI2 and BOLD Systems, included data from more molecular markers. The former contains 15 markers (mtDNA), while the latter integrates data from any molecular marker submitted by the user. In terms of taxonomic coverage, the databases spanned a wide range of organism groups, from specific clades such as diatoms (Diat.barcode) or algae (Phytool_v2) to broad collections targeting multiple domains of life (e.g., SILVA and BOLD).

Furthermore, most databases do not allow uploading information about steps taken in the study workflow, such as sample collection data, DNA extraction, sequence protocol (e.g., https://www.protocols.io), sequencing strategy, and sequence analysis (n = 20). In fact, COInr allows merely the upload of the former and latter, GSR only the latter, while Diat.barcode does not allow information regarding the latter two. Additionally, only two databases enabled all of the mentioned workflow steps (NISdb and DUFA), whereas EukRibo did not allow uploading information on DNA extraction.

Regarding data curation, Diat.barcode and UNITE implemented all four data curation types (A, B, C, D), eight databases covered the three types A, B, and C, BOLD covered types B and C, eight databases covered types A and B, and two included only type A curation (Table 2). Conversely, specific information on data curation was not available for five databases (MitoFish, Metabarpark, DUFA, NISdb, and FWAlgaeDB). The specific criteria and processing steps underlying each data curation type for each database are detailed in Table S4.

Fourteen (54%) databases lack a web-based interactive interface that would enable browsing available data and performing upload or download of data. Furthermore, there is no API support for 22 databases (85%). Regarding dedicated application or command-line support, 19 (73%) databases provide no such support. In fact, almost half of the databases (n = 12) were shown to have neither a web interface, API support, nor desktop/console support (COInr, NISdb, DUFA, GSR, EukRibo, MIMt, MARES, Metabarpark, µgreen, Diat.barcode, MZGdb, and Cyanoseq). PR2 and Phytool were the only databases that supported all of them. On the other hand, seven were found to display a single-access approach, GreenGenes2 and RDP supporting desktop/console access. Meanwhile, PhytoRef, MIDORI2, Cydrasil, Eukaryome, and MitoFish offer a web interface for browsing and downloading data. Three other databases were found to support both a web interface and desktop/console access (SILVA, GTDB, and FWAlgaeDB). The remaining databases, BOLD and UNITE, have a web interface and also allow data access through an API.

Table 2. List of databases containing DNA sequences searched in the scope of this study. Databases with confirmed current maintenance are underlined. Versions and number of records as of September 2025.

Database (version)	Target organisms	Markers	No. of records	Data curation type*	URL	Source
BOLD Systems (v5)	Animals, protists, fungi, and plants	COX1; rbcL; matK; ITS; plus other markers submitted by users	17.65 M	B and C	[1]	Ratnasingham et al. (2024)
SILVA (v138.2)	Bacteria, Archaea, and Eukaryota	16S; 18S; 23S; 28S	11.39 M	A, B, and C	[2]	Quast et al. (2013)
UNITE (v10.0)	Eukaryota (focus on fungi)	ITS	3.85 M	A, B, C, and D	[3]	Kõljalg et al. (2005); Abarenkov et al. (2024)
Eukaryome (v1.9.4)	Eukaryota	18S; 28S; ITS	1.3 M	A, B, and C	[4]	Tedersoo et al. (2024)
MIMt	Prokaryotes and Eukaryotes	16S and 23S (prokaryotic); 18S and 28S (eukaryotic); ITS (fungi)	740.4 k	A, B, and C	[5]	Cabezas et al. (2024)***
PR2 (v5.1.0)	Eukaryota (focus on protists)	18S	240.2 k	A, B, and C	[6]	Guillou et al. (2012)
GSR	Archaea and Bacteria	16S	90.4 k	A, B, and C	[7]	Molano et al. (2024)
EukRibo (v2.0)	Eukaryota (focus on protists)	18S	80 k	A, B, and C	[8]	Berney et al. (2022)
Diat.barcode (v15)	Diatoms	rbcL; 18S; 28S; COX1; ITS	10.5 k	A, B, C, and D	[9]	Rimet et al. (2019); Gelis et al. (2025)
Cyanoseq (v1.3)	Cyanobacteria	16S	5.5 k	A, B, and C	[10]	Lefler et al. (2023)
Cydrasil (v3)	Cyanobacteria	16S	1.3 k	A, B, and C	[11]	Roush et al. (2021)
GreenGenes2 (2024.09)	Archaea and Bacteria	16S	0.32 M (+ 16 k genomes)	A and B	[12] [13]	McDonald et al. (2024)

Database (version)	Target organisms	Markers	No. of records	Data curation type*	URL	Source
MIDORI2	Eukaryota	COX1; COX2; COX3; 12S; 16S; A6; A8; CytB; ND1; ND2; ND3; ND4; ND4L; ND5; ND6	3.74 M	A and B	[14]	Leray et al. (2022)
COInr (2025_05_11)	Metazoa	COX1	3.26 M	A and B	[15]	Meglécz (2023)
GTDB (R226)	Archaea and Bacteria	Draft genomes	732.5 k	A and B	[16]	Chaumeil et al. (2020)
Phytool_v2	Algae	16S (prokaryotic); 18S (eukaryotic); 23S (prokaryotic); rbcL (diatoms)	30.6 k	A and B**	[17]	Canino et al. (2021) Tapolczai et al. (2025)
PhytoRef	Photosynthetic eukaryotes	16S (chloroplastic)	6.5 k	A and B	[18]	Decelle et al. (2015)
µgreen	Eukaryotic algae, cyanobacteria, and non-vascular land plants	23S	2.3 k	A and B	[19]	Djemiel et al. (2020)
RDP (training set 19)	Archaea, Bacteria, and Fungi	16S; 28S	6.1 k	A and B	[20]	Maidak et al. (1996)
MZGdb (v2025-m03-01)	Zooplankton, benthic invertebrates, fish, marine mammals, and phytoplankton	COX1; 12S; 16S; 18S; 28S	91.8 k	A	[21]	Bucklin et al. (2021)
MARES	Marine eukaryotes	COX1	132.6 k	A	[22]	Arranz et al. (2020)
MitoFish	Fish	Full/partial mitochondrial genome	883.8 k	NA	[23]	Iwasaki et al. (2013) Zhu et al. (2023)
Metabarpark	Eukaryota	COX1; 18S	200 k	NA	[24]	Wangensteen and Turón (2016) Turón et al. (2020)
DUFA	Fish	COX1; 12S	175 k	NA	[25]	NA

Database (version)	Target organisms	Markers	No. of records	Data curation type*	URL	Source
NISdb	Marine NIS	COX1	8.9 k	NA	[26]	Zarcero et al. (2024)
FWAlgaeDB	Freshwater algae	Genome	0.2 k	NA	[27]	Lai et al. (2023)

***** Data curation type was categorised as follows: A - Taxonomic annotation (e.g., nomenclature validation; standardised taxonomic ranks; removal of records without valid taxonomy or with unaccepted expressions); B - Sequence quality filtering (e.g., length; ambiguous base pairs; chimeras, indels; pseudogenes); C - Phylogenetic congruence analysis and/or domain expert validation; D - cumulative implementation of ABC curation procedures with complete curation of all records down to species level; curation actions are transparent, traceable and accessible.

** Although classified as type B curated, no active sequence data curation was performed within Phytool_v2 itself; instead, it relied on data mined, and previously curated, from other databases.

*** Cabezas et al. (2024) pertains to an earlier version of the MIMt database, previously compiling only prokaryotic 16S. It has since been updated to encompass reference data from a broader diversity of molecular markers.

DNA reference libraries questionnaire

The survey yielded replies from 45 contributors, representing 39 institutions. Only replies from projects generating reference data relevant to European aquatic biodiversity were considered for further analysis, including projects led by non-European institutions. After filtering, replies from 33 institutions (85%) were considered for further analysis (see Table S5). In general, the analysed dataset was mostly represented by European institutions (88%), of which approximately 79% and 10% corresponded to academic and governmental institutions, respectively. The remaining institutions included a non-governmental organisation, a research foundation, and a stakeholder. Non-European contributors represented three academic and a single governmental institution (Table S5).

Regional-scale projects contributed most to the survey (34% of the replies). These were either within country borders (18%) or across two or more countries (international; 16%). The former generally encompasses reasonably sized areas within a single country, while the latter comprises regions across several countries and shares similar characteristics (usually at a larger geographic scale than the former). International and national-sized projects, by contrast, included cases targeting full-scale countries—either multiple or single—representing around 28% and 23%, respectively. Localised projects accounted for a mere 13% of cases. A single submission did not provide information regarding the scale of the developed project.

The bulk of the projects accounted for marine (ca. 42%) and freshwater biodiversity (33%), while the remaining submissions (15%) addressed biodiversity in transitional waters. Only six projects were not categorised within the three main environment types: three groundwater taxa-specific, two terrestrial, and a single unspecified case. Two projects declared as terrestrial could not be unambiguously separated from aquatic data and were retained. Despite their inclusion, these cases are expected to have a negligible impact on the downstream analysis and were thus not considered a variable of interest. Further details on the main highlights of the questionnaire results can be found in Table S6.

The majority of the data input corresponded to eukaryotic taxa (83%), with metazoans representing the bulk, comprising approximately 67% of the reported libraries. Prokaryotic libraries accounted for around 17%, with Archaea and Bacteria similarly represented (Fig. 2). Among non-metazoan eukaryotes, chromists and plants were reported roughly in 16% and 3% of the reference libraries, respectively. Within metazoans, invertebrates dominated, contributing approximately 53% of the reported data, whereas vertebrates, primarily fish, accounted for around 15% (Fig. 3).

Several generalist terms were used to describe the target taxa. For instance, the term “plankton” was used to simultaneously refer to prokaryotes and eukaryotes. In such cases, these groups were treated separately in the present analysis. However, in instances where it was not possible to discriminate further (e.g., “Eukarya” and “microalgae”), finer classification was not possible.

Morphology-based, species-level taxonomic identifications were the predominant approach in metazoan-dominated reference libraries. In contrast, microbial libraries more commonly employed molecular identifications, typically resolved to genus or family level. Analysis of taxonomy cross-reference sources revealed frequent use of the World Register of Marine Species (WoRMS; 24%), including for freshwater taxa. Notably, freshwater libraries cited both WoRMS and the Global Biodiversity Information Facility (GBIF) equally (9%), although the NCBI Taxonomy database emerged as the most used source overall (10%). In addition to these online platforms, some contributors reported

relying on literature or expert consultation, though these cases were much less frequent (e.g., Diat.barcode).

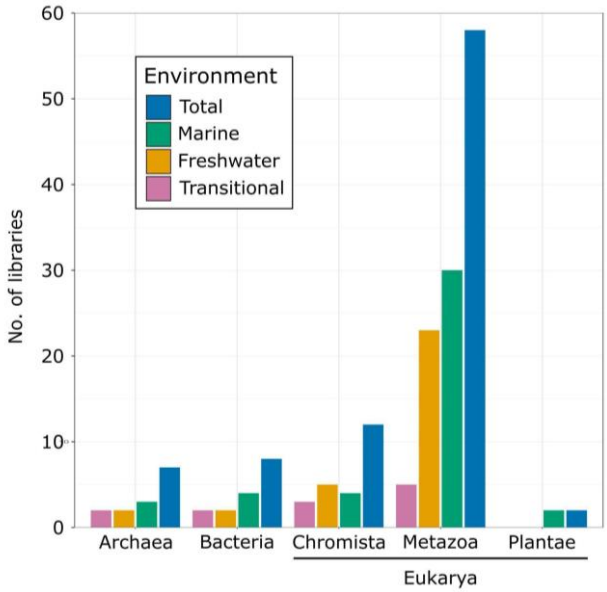

Figure 2. Number of reference libraries reported for major organismal groups across different aquatic environments.

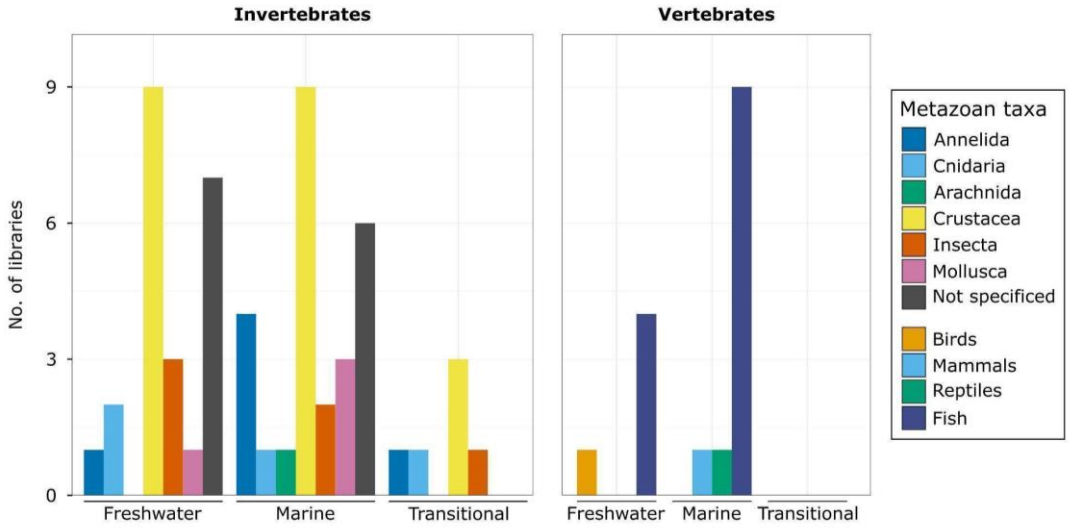

Figure 3. Number of reference libraries reported for metazoan taxa across different aquatic environments.

Among the available data, most reference libraries (~59%) contained fewer than 200 species, with species-rich libraries being relatively rare (12% contained more than 1000 species). Two projects stood out: one reported encompassing over 10,000 taxa across all environment types (MIMt, CIBIO), and another documented more than 50,000 taxa (Tara Oceans Foundation). Both targeted Bacteria and Archaea, with the latter further targeting exclusively marine plankton (prokaryotes, eukaryotes, and viruses). Furthermore, sequencing effort was reflected by the number of specimens, which fell into two main ranges: fewer than 500 specimens (38%) and 1,000 to 3,000 specimens (30%) per reference library.

A diverse array of molecular markers, 21 in total, were identified in the development of DNA reference libraries for European aquatic biodiversity. Consistent with findings from the genetic database survey, COX1, 16S, and 18S emerged as the most frequently targeted markers. However, in the present analysis, COX1 alone accounted for 30% of the reference libraries, followed by the mitochondrial 16S (15%), 18S (10%), and ITS (7%). This general pattern was consistent across all assessed aquatic environments. An exception was observed in transitional waters, where both COX1 and 16S had the same number of cases, but a poorer diversity of molecular markers.

Most projects (~62%) had deposited at least part of the sequences in primary public databases. They deposited their data either in BOLD (37%), in NCBI (GenBank) (13%), in a combination of both (33%), in a few other databases (11%; ENA, MIMt, Phytool, and an unnamed database under development from France) or a combination of NCBI with other databases (7%; ENA, Diat.barcode, SILVA, PR2, MetaZooGene, VAMPS). Although most sequences were said to be accompanied by metadata (91%), the majority of those (55%) did not declare following any metadata standards. The most frequently deposited metadata included were: project identification, GPS coordinates, voucher ID and location, person collecting and identifying, and depth/altitude. Collection date was mentioned only in two cases, and specimen images were reported in only about half of the libraries. Taxonomic annotation procedures were applied in nearly 65% of the libraries, while phylogenetic congruence checks were reported in about one third of them. An array of taxonomic references were reported for cross-referencing identifications, mainly including broad taxonomic references such as WoRMS (24%), NCBI Taxonomy (20%), GBIF (14%), Catalogue of Life (8%), BOLD (5%) and ENA (5%), as well as more taxon-specific references such as FishBase (2%), the literature (6%), AlgaeBase (3%) and Diat.barcode (2%).

Discussion

While evaluating the current status of European aquatic DNA reference libraries highlights significant fragmentation, quality gaps, and limited interoperability, our findings underscore the urgent need for a unified framework. The discussion translates these empirical insights into a strategic, actionable blueprint for the future. By detailing the three fundamental pillars of a federated network of reference libraries, i.e., essential metadata, standardised curation workflows, and seamless technical interoperability, we demonstrate how local taxonomic expertise can be connected with global repositories. We then detail the proposed data architectures, quality ranking criteria, and a sustainability roadmap designed to turn today's dispersed data into a federated network of curated reference libraries that will enhance aquatic (e)DNA biomonitoring in Europe.

Understanding primary data generation for aquatic reference libraries in Europe

The landscape of DNA reference libraries in Europe is considerably dispersed and uneven in content, structure, management and interoperability. Disproportions are evident across major taxonomic groups, markers and countries. Part of this imbalance reflects necessary adjustments in research practice, operational approaches and metadata requirements for handling distinct types of aquatic organisms (microorganisms, animals, plants; Rimet et al. 2021). Primary data from large-scale initiatives, such as iBOL, contribute to the cohesiveness of reference libraries by promoting standardised (meta)data and interoperable, publicly accessible systems. Yet, library completeness remains uneven

among taxonomic groups, regions and environments (Leite et al. 2020; Marques et al. 2021; Schultz and Hebert 2024). For many of them, the curation systems are still under development or implementation (e.g., Fontes et al. 2026; Kühbandner et al. 2026).

European cross-country networks and governmental institutions (such as natural history museums and environment-related agencies) often contribute primary data and have their own standards, database structures and policies, which do not always align easily with large-scale initiatives and repositories. Relevant primary data structures also include numerous other research and academic institutions dispersed across Europe, often focused on particular taxa or geographic regions, thereby filling unique gaps in libraries. However, such contributions may be sporadic, depend on short-term funding, not follow common standards, or lack the capacity to ensure proper data management and FAIRness.

Implementing dependable European-wide reference libraries to sustain eDNA-based monitoring is particularly challenging given the diversity of primary contributors. The current landscape, combining multiple autonomous agents, country-level institutions, and European-level networks, has long been established and is expected to persist in the future. From this perspective, the challenge is to build on the existing dispersed structure, leverage the disparate contributions and sources of the various agents and find sustainable solutions to harmonise them into a dependable product.

Rationale and architecture of a federated system of taxonomically curated reference libraries

While decentralised, locally curated collections often offer superior local taxonomic coverage and flexibility, their lack of common standardised curation procedures and uniform metadata protocols exacerbates fragmentation and limits cross-study comparability. Overcoming these bottlenecks requires establishing a federated network grounded in shared standards for data quality, curation, and connectivity. Rather than forcing all locally gathered data into a single monolithic archive, such a system should rest on three essential pillars: comprehensive essential metadata using metadata terminology standards, standardised and transparent curation workflows, as well as seamless technical interoperability (Fig. 4). Broad consensus on such principles would allow independently managed local and regional databases to preserve their autonomy while uniting into a cohesive, machine-readable ecosystem. Keeping the diversity of these reference libraries would also let them meet different needs; it may improve taxonomic experts' commitment to maintaining these libraries, while their unison allows for easier data access and reuse. Ultimately, such a federated system would ensure that distributed taxonomic expertise could directly power reproducible, high-throughput biodiversity monitoring across regional and national scales as well as across diverse governance frameworks.

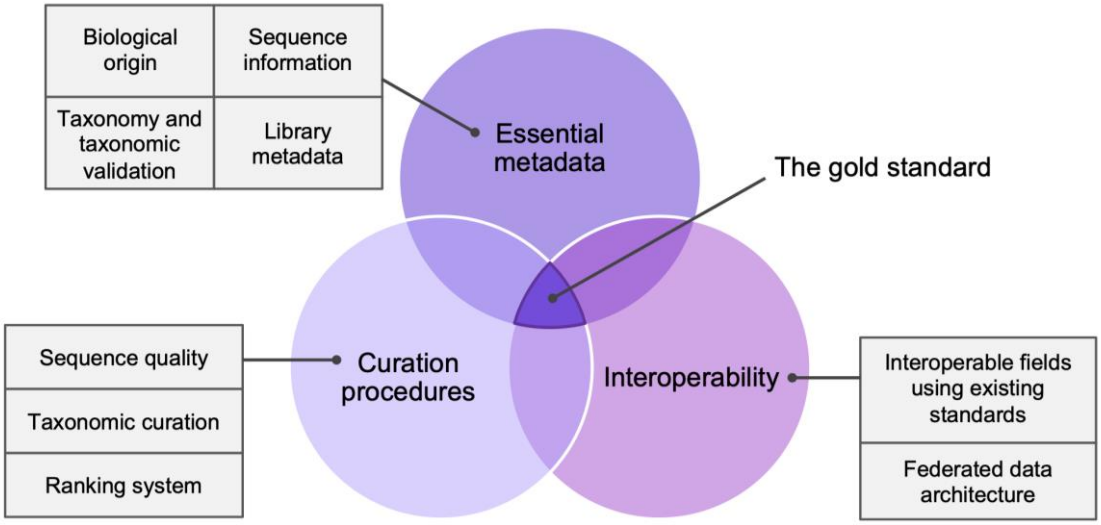

Figure 4. Summary of the three pillars needed to improve curated reference libraries for Europe.

The architecture of the proposed federated system relies on a common data portal providing access to FAIR data held in the participating reference libraries, which abide by a set of prerequisites on essential metadata, curation procedures and interoperable formats (Fig. 5). Rather than creating a new repository replicating the information already available in other databases, the goal is to create a portal that aggregates minimal and compliant data, while maintaining interoperable links to the primary repositories containing the full (meta)data elements. Data sources can include newly generated sequences, data mined from primary nucleotide databases, such as BOLD and INSDC, and other repositories and databases from multiple institutions and initiatives that characterise the reference libraries' landscape. Hence, by keeping the bulk of data in their original repositories, this federated system would require comparatively low logistics and lower the burden of sustainability, while ensuring easily accessible, taxonomically curated and reliable reference DNA sequences for eDNA-based monitoring in Europe.

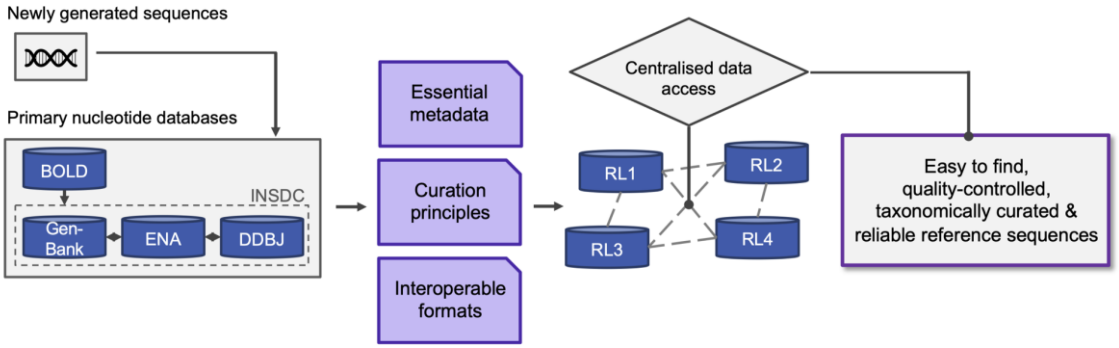

Figure 5. Schematic representation of the eDNAqua-Plan vision for a federated system of curated reference libraries. RL1, RL2,… represent independent reference libraries following common principles that enable them to be searchable through a common data portal.

Several data architecture and interoperability options can be considered for the federated system. Two possible solutions are depicted in Fig. 6. In the first solution, reference libraries are individually managed on separate servers, with each source database possessing its own API, compliant with the federated, distributed system. This will enable queries across the distributed databases' APIs from a single federated portal. Although this model would provide a straightforward solution, establishing the inter-connections may be challenging, servers may go down episodically, and many databases lack IT personnel to maintain such a structure permanently. In such cases, an alternative second solution would be to periodically (e.g., monthly, quarterly, yearly) export data to a central server in an agreed-upon format, while keeping the reference libraries individually managed (Fig. 6, solution 2). In both cases, the common portal provides a single point of access, while the participating reference libraries remain the authoritative sources and retain management of their records.

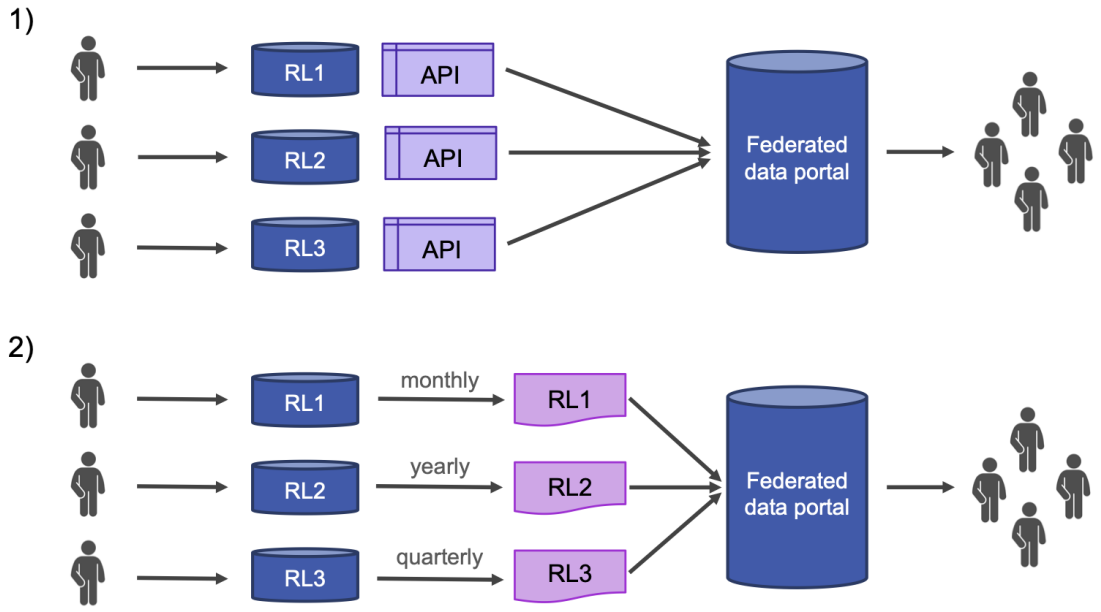

Figure 6. Several data architectures are possible for federation, where all reference libraries either 1) have standardised APIs or 2) periodically exchange their latest data in a standardised format. Both solutions allow for individual teams or researchers to generate and maintain databases that are then made centrally accessible in one federated portal that the community can easily find and access.

Sets of essential metadata

The "essential metadata" is one of the three pillars of the proposed federated system of curated libraries (Fig. 4). Such metadata ensure data traceability and enable the assessment of data quality. The rationale for the essential metadata is twofold: a) ensure that records imported to the federated library comply with minimal quality (meta)data standards, and b) retain only essential information to enable autonomous and reliable taxonomic assignments of (e)DNA data through the federated system of libraries, without

over-replicating massive amounts of data from full records, which will be accessible anyway through agreed linkages when surplus information is needed.

Four sets of (meta)data elements are considered, namely “Biological origin”, “Taxonomy and taxonomic validation”, “Sequence information”, and “Library metadata”. Within each set, eDNAqua-Plan proposes a list of recommended essential elements (Rimet et al. 2026; Table S7).

Curation procedures

The proposed curation workflow involves three sequential stages, starting with sequence quality curation, then taxonomic curation and, finally, quality ranking (Fig. 7).

Sequence quality

Sequence quality curation is essential for detecting sequencing artefacts and structural inconsistencies prior to any taxonomic assignment. Basic sequence curation steps are relatively well established and followed by most repositories; sequence validation workflows have been described, for example for DNA barcode data (see Pentinsaari et al. 2020).

Acceptable levels of sequence quality must be defined using transparent, reproducible and marker-specific criteria. Among others, recommended thresholds to be set include (Rimet et al. 2026): ambiguous bases (“N”), sequence length, reading frame integrity (coding genes), indels (coding regions), alignment coherence, marker confirmation (rRNA genes) and contamination screening.

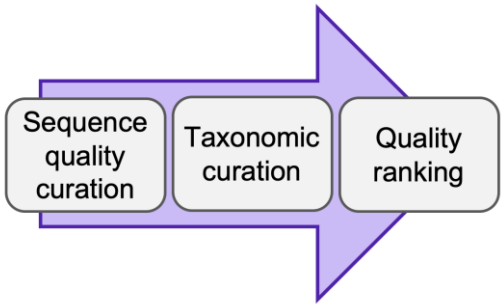

Figure 7. Curation procedures workflow. There are three steps necessary to make library taxonomic assignments reliable.

Taxonomic curation

As observed in the landscape assessment, most repositories and libraries lack a complete curation system of the taxonomic identity of DNA sequence data. Low levels of completeness of libraries, together with data FAIRness and the taxonomic reliability of records, are among the main concerns in (e)DNA monitoring and arguably one of the most challenging issues to tackle so far. However, much effort has been made lately to

develop taxonomic curation systems that provide pathways to overcome major difficulties.

Taxonomic curation involves more than harmonising multiple taxonomic reference systems or rectifying errors such as misspelt taxonomic names. It also pertains to the validation and reliability of the association between a DNA sequence and the respective taxa (and taxonomic name) from which it was obtained. For the latter, several approaches may be used, as for example: expert curation based on phylogenies, expert curation based on clustering genetically similar sequences, and expert curation based on self-assignment [see Rimet et al. (2026) for further details of these approaches]. However, one must recognise a fundamental discrepancy that persists between the strict principles of Linnaean taxonomy and the practical reality of generating reference DNA barcodes in the metazoan world. While the ideal standard dictates that reference sequences should derive directly from morphological holotypes or topotypical populations, achieving this across expanding databases remains largely unrealistic. Acknowledging this limitation would be a desirable step in modern curation, requiring frameworks to pragmatically bridge the gap between historical type-specimen descriptions and contemporary sequence data without hindering the biomonitoring practice. For example, reference libraries could implement a confidence-ranking system that scores sequences based on their morphological, genetic and geographical proximity to the original type material. By categorising records into distinct traceability levels, ranging from sequences derived directly from type specimens down to those from non-type strains lacking morphological documentation, researchers could fully utilise available data while maintaining a transparent and reliable measure of taxonomic authenticity.

Therefore, as part of the curation process, researchers should put particular emphasis on taxonomic traceability and morphological verification. To promote harmonisation across reference libraries, we recommend the following criteria, grounded in these two features, to assist database managers and experts in scoring and annotating records (Rimet et al. 2026):

Level 1 – Sequence derived from a type specimen or type strain.

Level 2 – Sequence from a non-type specimen/strain collected near the type locality, with published morphological information sufficient to unambiguously assign the specimen/strain to the species or its equivalent (OTU).

Level 3 – Sequence from a non-type specimen/strain collected distant from the type locality, with published morphological information sufficient to clearly assign the specimen to the species or its equivalent (OTU).

Level 4 – Sequence from a non-type specimen/strain with published morphological information, but insufficient to unambiguously confirm species or equivalent identity.

Level 5 – Sequence from a non-type specimen lacking published morphological documentation.

These levels describe the strength of the taxonomic evidence supporting each record and are intended as a scoring aid, not as admission thresholds for the federated system. Applying these expert-driven taxonomic curation and scoring approaches to DNA sequence databases requires considerable individual expert input, making them impractical for handling massive, ever-growing databases (e.g., Radulovici et al. 2021).

Semi-automated systems have been developed to assist taxonomic curation and lower the effort required by experts, as for example the Barcode Audit and Grade System (BAGS; Fontes et al. 2021), the Biodiversity Genomics Europe curation tool (Kühbandner et al. 2026) and the dbDNA system (available at https://github.com/TillMacher/dbDNA).

BAGS provides an auditing system for BOLD COI barcode data based on the concordance between the species names assigned to the records and their clustering into BINs, followed by the annotation of each species and its records with one of five grades (A–E), according to the level of name–BIN concordance and the number of available records. Although it provides an efficient tool to quickly screen large datasets and flag incongruent records, it does not include any metadata scoring system and, at least so far, can only be applied to COI barcode data deposited in BOLD.

The BGE curation tool is similarly focused on BOLD's COI barcodes, but in addition to BAGS grades, it also comprises a ranking system for metadata elements associated with each record, and labels for representative records for each country. Hence, it provides a tool for automated screening and grading of taxonomic congruency together with metadata quality, delivering a structured output that greatly facilitates downstream inspection and annotation by experts.

dbDNA stands out for being a marker- and database-independent automated rating system for reliable taxonomic annotation of metabarcoding data, requiring minimal downstream expert input. It uses taxonomic concordance criteria, together with a metadata scoring system to annotate reference library records with quality scores (gold, silver, bronze or unreliable). Among several unique features, dbDNA includes lists of taxonomic experts to assist in scoring library records with different levels of reliability. Ultimately, rated sequences can be used to build curated reference libraries of high quality.

Although these semi- or fully automated systems may greatly contribute to improving and facilitating large-scale curation across aquatic reference libraries, they are not exempt from shortcomings. More relaxed or superficial curation criteria may deem erroneous records as reliable, whereas more conservative criteria may leave out many trustworthy records that are signalled as unreliable. Indeed, detailed downstream "manual" analyses of records graded discordant by BAGS have shown that some may be in fact pseudo-discordant, among other reasons, due to morphology-based misidentifications between closely related species (Fontes et al. 2021; Lavrador et al. 2023). Hence, some detailed level of "manual" curation may still be required in these automated systems, especially to inspect the most intricate cases.

In this regard, artificial intelligence (AI) systems may offer tools with great potential to refine reference library curation, particularly by reviewing existing curation systems and attempting to solve manual curation-dependent cases. For example, a pilot study by Fontes et al. (2026) showed the considerable potential of machine learning tools to mimic human-based curation decisions in a test case using marine fish reference libraries.

Quality ranking

The landscape analysis revealed highly heterogeneous approaches and extent of curation of DNA sequence records across databases. The term "curation" itself appears to be

interpreted differently among researchers, and the curation procedures employed by repositories are not always available in sufficient detail or clarity.

The workflow proposed here systematises the steps that must be considered in a complete curation procedure of reference libraries: sequence quality curation, taxonomic curation and, finally, quality ranking. This final step is intended to give researchers, stakeholders and, in particular, non-specialist end-users a simple and reliable indication of the strength of evidence supporting each record's taxonomic assignment, without requiring familiarity with the detailed curation criteria applied upstream.

Several databases have already implemented ranking systems (e.g., UNITE, BOLD, PR2, SILVA), tailored to the needs of the communities they serve and following different approaches. Because every record admitted into the federated system proposed here has, by definition, already passed the essential-metadata and curation workflow described above, end-users can consider all such records reliable. Here, 'reliable' denotes compliance with the three pillars of the federated system rather than a uniform threshold of taxonomic evidence. The mandatory requirements are common to all libraries. Still, the specific curation criteria, including which records are retained, the taxonomic rank at which each record is validated, and how conflicting identifications are resolved or flagged, are defined and documented by each contributing database. This reflects the diversity of organisms, markers and taxonomic practices across libraries, which is best handled by the expert communities that maintain them.

Still, a further distinction helps highlight records with especially extensive evidence and solid agreement within the research community regarding their taxonomic authenticity. Such evidence may come from factors such as a sequence derived from a type specimen, multiple independent specimen records for the same taxon, or cross-validation between independent research sources. We propose flagging these records with a single "very high confidence" stamp. This is not intended as an additional multi-level scale, but as one optional, positive marker layered on top of the baseline "reliable" status shared by all records in the federated system.

The precise criteria for attribution of the "very high confidence" stamp are left within the scope of each contributing repository or database, since these are best placed to combine their specific taxonomic expertise and established curation procedures into a grounded decision. Nevertheless, to ensure the stamp conveys a comparable meaning across libraries, one should reserve it for records whose species-level identification is particularly robust and supported by multiple, convergent lines of evidence beyond the baseline requirements that apply to all records in the system. Examples of such evidence include: (i) sequences derived from type specimens or type strains; (ii) concordant sequences from multiple specimens of different geographic origins, collected and identified independently by different researchers; (iii) records of frequently occurring, well-studied species that are taxonomically stable, have extensive historical ecological and occurrence data, and whose morphology-based identification is well established and unambiguous; and (iv) records of species belonging to genera that are well represented in the library, for which morphology-based identification among congeners carries little ambiguity and which are clearly discriminated genetically from their closest relatives. These examples are indicative rather than prescriptive and need not all apply simultaneously; each database should document the criteria it applies, so that end-users can trace why a given record was awarded the stamp.

Interoperability

Accessibility and interoperability are essential for maximising the value of eDNA reference libraries. Reference data should be openly accessible wherever possible, supported by persistent identifiers, clear licensing, and sufficient metadata to enable reliable reuse. Interoperability requires that reference sequences have provenance, i.e., can be consistently linked to voucher specimens, taxonomic backbone identifiers, collection records, and sequence archives. It is highly recommended to collaborate with the GSC (Genomic Standards Consortium; https://www.gensc.org/) and TDWG (Biodiversity Information Standards; https://www.tdwg.org/) standards organisations to aid sustainable alignment and interoperability of the minimum metadata identified in this paper. Interoperability requires shared identifiers, controlled vocabularies, and machine-readable metadata. Improving these aspects allows reference records to be assessed, exchanged, and integrated across distributed databases. This supports a federated system in which individual reference libraries retain their specialist roles, whilst contributing to a more coherent and reliable eDNA data infrastructure. We note that this will all contribute to AI-ready data (Harrison et al. 2026).

Sustainability plan and roadmap for implementation

The long-term sustainability of eDNA reference libraries depends on stable funding, good governance, continued curation, and reliable technical infrastructure. Regular interaction with European and other policy leaders is needed to demonstrate how comparatively small amounts of eDNA infrastructure funding enable better decisions, from nature conservation and fisheries management to sustainable farming and to emerging pathogen monitoring, all powered by good-value, high-quality, data-driven biodiversity and other biological monitoring that complements other monitoring techniques. A federated model offers a sustainable foundation by allowing established repositories and specialist libraries to retain their roles while following common interoperability principles. This approach can minimise duplication, reduce dependence on any single resource, and distribute responsibility and funding for long-term maintenance across participating institutions (Huertas Olivares et al. 2026).

Initial implementation should establish a shared governance framework, minimum metadata and quality requirements, persistent identifiers, and versioning procedures. In the short to medium term, focus should be on European-based resources, while remaining inclusive and using international standards. Participating resources should then adopt interoperable exchange formats and documented APIs, enabling records and corrections to move between systems while preserving both provenance and local ownership. In the medium term, regular quality audits and procedures for updating taxonomies and metadata standards will be needed. We anticipate that reference libraries will significantly increase their taxonomic coverage of species and their infraspecific variants, adding scientifically important complexity, but that the data volumes will be low compared to sequence archives. In the longer term, advances in software and technology standards warrant periodic evaluation and updating, allowing the reference library system to evolve (Collart et al. 2026).

Conclusions

Our findings show that, while eDNA analysis is a transformative tool for monitoring aquatic biodiversity in Europe, the reference libraries underpinning its full functionality and reliability are highly fragmented and inconsistent. To address this, we propose the following practical, community-driven recommendations across four thematic areas to build a federated and reliable system of taxonomically curated reference libraries.

System Architecture and Interoperability

Build a Federated Network, Not a Centralised One: Develop a federated network that connects independently managed reference libraries into a cohesive whole, avoiding data fragmentation while preserving each library's independence and specialist expertise, rather than consolidating them into a single, centralised database.

Enable Technical Interoperability: Connect participating libraries through standards-compliant APIs; where this is not yet feasible, adopt a near-term alternative based on periodic, standardised data exports to a central aggregator.

Adopt Common Data Standards: Ground technical interoperability in well-established data frameworks, such as GSC MIxS and Darwin Core.

Data Quality and Metadata

Require Essential Metadata: All imported records should include a minimum set of essential metadata, covering biological origin, taxonomic validation, sequence information, and library-level descriptors.

Apply Transparent Sequence Quality Checks**:** Before any taxonomic assignment, apply clear, reproducible quality checks on sequences, with defined limits for ambiguous bases, sequence length, reading-frame integrity, and contamination.

Introduce a Two-Tier Quality Assessment: Within curation procedures, apply a detailed, traceability-based scoring system (Levels 1–5) to assist researchers and database managers in curating records. At the point of use, complement this with a simple, optional "very high confidence" flag that signals to non-specialist end-users which records are supported by especially strong evidence, without requiring familiarity with the underlying criteria.

Account for Biological Complexity in Curation Criteria: Curation procedures should explicitly accommodate biological complexities such as cryptic species, incomplete

lineage sorting, and intraspecific variation, rather than treating every taxonomic conflict as a simple error.

Advanced Curation Tools

Use Semi-Automated Curation Tools: Semi-automated taxonomic curation tools should be used to reduce the manual labour involved in screening large datasets.

Incorporate Artificial Intelligence: AI methods can be incorporated to progressively improve automated curation, given their potential to resolve complex taxonomic issues that currently require expert review.

Governance and Sustainability

Ensure Long-Term Sustainability: Secure stable funding and establish shared governance to support this technical infrastructure over the long term; anchor it wherever possible in established European research infrastructures, and maintain ongoing communication with policymakers to convey how essential well-curated, federated reference libraries are for robust eDNA-based aquatic biodiversity monitoring.

Acknowledgements

We thank the researchers and institutions that replied to the questionnaire, providing useful information for analysis, and the stakeholders participating in our discussion sessions, which helped shape the recommendations. We fully acknowledge the rest of the eDNAqua-Plan consortium (https://doi.org/10.5281/zenodo.17143114), who all directly or indirectly contributed to the material and ideas behind this paper. We thank Florian Leese (University of Duisburg-Essen, Germany), Ian Probert (Sorbonne University, France) and Pilar Cabezas (University of Minho, Portugal) for their contribution to earlier discussions on the topics of this article.

Conflict of interest

The authors declared that no competing interests exist.

Ethical statement

No ethical statement was required.

Use of AI

The authors used generative AI tools (ChatGPT 5.6 Sol Light; Claude Sonnet 5; Claude Opus 5.5) for language editing, improving clarity and readability, and refining the structure and presentation of selected parts of the text. All scientific content, interpretation, and conclusions were developed and verified by the authors.

Funding

This work was supported by the European Union under the Horizon Europe Programme (eDNAqua-Plan, Grant Agreement No. 101112800).

Author contributions

Conceptualization: FOC, FR, MG, EB. Investigation: FOC, IK, JM. Data curation: FOC, IK, JM. Formal analysis: FOC, JM. Visualization: FOC, JM, PW, EB. Writing – original draft: FOC, FR, AJB, PW, MG, EB. Writing – review & editing: FOC, FR, AJB, IK, JM, SP, PW, MG, EB. All authors read and approved the final manuscript.

References

Abarenkov K, Nilsson RH, Larsson K-H, Taylor AFS, May TW, Frøslev TG, Pawlowska J, Lindahl B, Põldmaa K, Truong C, Vu D, Hosoya T, Niskanen T, Piirmann T, Ivanov F, Zirk A, Peterson M, Cheeke TE, Ishigami Y, Jansson AT, Jeppesen TS, Kristiansson E, Mikryukov V, Miller JT, Oono R, Ossandon FJ, Paupério J, Saar I, Schigel D, Suija A, Tedersoo L, Kõljalg U (2024) The UNITE database for molecular identification and taxonomic communication of fungi and other eukaryotes: sequences, taxa and classifications reconsidered. Nucleic Acids Research 52: D791–D797. https://doi.org/10.1093/nar/gkad1039

Altermatt F, Couton M, Carraro L, Keck F, Lawson-Handley L, Leese F, Zhang X, Zhang Y, Blackman RC (2025) Utilizing aquatic environmental DNA to address global biodiversity targets. Nature Reviews Biodiversity 1: 332–346. https://doi.org/10.1038/s44358-025-00044-x

Appleyard SA, Ward RD, Pogonoski JJ, Graham A, Last PR, Deagle BE, Holmes B, Gomon MF, Bray DJ, Johnson JW, Hay AC, Moore GI, Hammer MP, Russell B, Graham KJ (2025) Australia's marine fishes DNA barcode reference library for integrated taxonomy, metabarcoding and eDNA research. Scientific Data 12(1): 21. https://doi.org/10.1038/s41597-025-04375-4

Arranz V, Pearman WS, Aguirre JD, Liggins L (2020) MARES, a replicable pipeline and curated reference database for marine eukaryote metabarcoding. Scientific Data 7(209): 1–8. https://doi.org/10.1038/s41597-020-0549-9

Babo C, Rimet F, Martins F, Camacho A, Probert I, Woollard P, Boulanger E (2024) eDNAqua-Plan Milestone 5: Proceedings of the WP3 Workshop on standards requirements. Zenodo. https://doi.org/10.5281/zenodo.14967949

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers ER (2013) GenBank. Nucleic Acids Research 41(D1): D36–D42. https://doi.org/10.1093/nar/gks1195

Berney C, Henry N, Mahé F, Richter DJ, de Vargas C (2022) EukRibo: a manually curated eukaryotic 18S rDNA reference database to facilitate identification of new diversity. bioRxiv 1–20. https://doi.org/10.1101/2022.11.03.515105

Boulanger E, Babo C, Veríssimo J, Rimet F, Exter K, Woollard P (2025a) eDNAqua-Plan Milestone 6: Proceedings of the WP3 workshop for conceptualising and writing on the eDNA data ecosystem. Zenodo. https://doi.org/10.5281/zenodo.14967994

Boulanger E, Exter K, Paupério J, Suominen S, Provoost P, Rimet F, Picazo Mozo A, Camacho A, Babo C, Veríssimo J, Norros V, Hablützel P, Pavloudi C, Meissner K, Woollard P (2025b) eDNAqua-Plan Milestone 7: the eDNAqua-Plan conceptual landscape proposal v1. Zenodo. https://doi.org/10.5281/zenodo.16367374

948 Bucklin A, Peijnenburg KTCA, Kosobokova KN, O'Brien TD, Blanco-Bercial L, Cornils A, Falkenhaug
949 T, Hopcroft RR, Hosia A, Laakmann S, Li C, Martell L, Questel JM, Wall-Palmer D, Wang M,
950 Wiebe PH, Weydmann-Zwolicka A (2021) Toward a global reference database of COI barcodes for
951 marine zooplankton. Marine Biology 168(78): 1–26. https://doi.org/10.1007/s00227-021-03887-y

952 Cabezas MP, Fonseca NA, Muñoz-Mérida A (2024) MIMt: a curated 16S rRNA reference database with
953 less redundancy and higher accuracy at species-level identification. Environmental Microbiome
954 19(88): 1–13. https://doi.org/10.1186/s40793-024-00634-w

955 Canino A, Bouchez A, Laplace-Treyture C, Domaizon I, Rimet F (2021) Phytool, a ShinyApp to
956 homogenise taxonomy of freshwater microalgae from DNA barcodes and microscopic observations.
957 Metabarcoding and Metagenomics 5(e74096): 199–205. https://doi.org/10.3897/mbmg.5.74096

958 Chaumeil P-A, Mussing AJ, Hugenholtz P, Parks DH (2020) GTDB-Tk: a toolkit to classify genomes with
959 the Genome Taxonomy Database. Bioinformatics 36(6): 1925–1927.
960 https://doi.org/10.1093/bioinformatics/btz848

961 Claver C, Canals O, de Amézaga LG, Mendibil I, Rodriguez-Ezpeleta N (2023) An automated workflow
962 to assess completeness and curate GenBank for environmental DNA metabarcoding: the marine fish
963 assemblage as case study. Environmental DNA 5(4): 634–647. https://doi.org/10.1002/edn3.433

964 Collart T, Koole CA, McMeel O, Exter K, Boulanger E, Woollard P, Hablützel P, Leese F, Derycke S,
965 Norros V, Laamanen T, Meissner K, Huertas Olivares CI, Arvanitidis C, Paupério J, Pade N (2026)
966 eDNAqua-Plan Deliverable 16: Roadmap for implementation of the optimal system(s)/workflow(s).
967 Zenodo. https://doi.org/10.5281/zenodo.22144786

968 Costa F, Moutinho J, Woollard P, Krawczyk D, Kavakiotis I, Seweryński M, Rimet F, Cabezas MP,
969 Beermann A, Leese F, Papakostas S, Grabowski M (2024) eDNAqua-Plan Deliverable 6: Report on
970 DNA Reference Libraries. Zenodo. https://doi.org/10.5281/zenodo.15167446

971 Decelle J, Romac S, Sern RF, Bendif EM, Zingone A, Audic S, Guiry MD, Guillow L, Tessier D, Le Gall
972 F, Gourvil P, dos Santos AL, Probert I, Vaulot D, de Vargas C, Christen R (2015) PhytoREF: a
973 reference database of the plastidial 16S rRNA gene of photosynthetic eukaryotes with curated
974 taxonomy. Molecular Ecology Resources 15: 1435–1445. https://doi.org/10.1111/1755-0998.12401

975 Djemiel C, Plassard D, Terrat S, Crouzet O, Sauze J, Mondy S, Nowak V, Wingate L, Ogée J, Maron P-A
976 (2020) µgreen-db: a reference database for the 23S rRNA gene of eukaryotic plastids and
977 cyanobacteria. Scientific Reports 10(5915): 1–11. https://doi.org/10.1038/s41598-020-62555-1

978 Exter K, Boulanger E, Camacho A, Collart T, Hablützel P, Koole CA, McMeel O, Pavloudi C, Picazo Mozo
979 A, Suominen S, Woollard P (2026) eDNAqua-Plan Milestone 14: Blueprint workshop report.
980 Zenodo. https://doi.org/10.5281/zenodo.18833437

981 Fontes JT, Vieira PE, Ekrem T, Soares P, Costa FO (2021) BAGS: an automated Barcode, Audit & Grade
982 System for DNA barcode reference libraries. Molecular Ecology Resources 21(2): 573–583.
983 https://doi.org/10.1111/1755-0998.13262

984 Fontes JT, Ferreira AO, Lavrador AS, Soares P, Costa FO, Canchaya CA (2026) Exploring machine
985 learning for the curation of DNA barcode reference libraries: a pilot study with marine fish species
986 of the Eastern Atlantic and adjacent waters. Molecular Ecology Resources (in press).

987 Gauvin P, Eme D, Domaizon I, Rimet F (2025) An annotated reference library for supporting DNA
988 metabarcoding analysis of aquatic macroinvertebrates in French freshwater environments.
989 Metabarcoding and Metagenomics 9(e137772): 149–173. https://doi.org/10.3897/mbmg.9.137772

990 Gelis MMN, Barry-Martinet R, Briand J-F, Chonova T, Cochero J, Gassiole G, Kahlert M, Karthick B,
991 Keck F, Kelly M, Kochoska H, Mann D, Nayak P, Pfannkuchen M, Servulo T, Trobajo R, Vasselon
992 V, Vidakovic D, Viollaz L, Wetzel C, Zimmermann J, Rimet F (2025) Assigning taxonomy and

traits to DNA sequences of river diatoms in a region with limited taxonomic knowledge using the updated and annotated reference library Diat.barcode. International Journal of Limnology 61: 10. https://doi.org/10.1051/limn/2025009

Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L, Boutter C, Burgaud G, de Vargas C, Decelle J, del Campo J, Dolan JR, Dunthorn M, Edvardsen B, Holzmann M, Kooistra WHCF, Lara E, Le Bescot N, Lugares R, Mahé F, Massana R, Montresor M, Morard R, Not F, Pawlowski J, Probert I, Sauvadet A-L, Siano R, Stoeck T, Vaulot D, Zimmermann P, Christen R (2012) The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. Nucleic Acids Research 41(D1): D597–D604. https://doi.org/10.1093/nar/gks1160

Harrison C, Clark-Casey J, Farrell G, Gibbs AJ, Occhipinti A, Overton I, Rusholme-Pilcher R, Spick M, Suchak T, Vollmar M, Zwiggelaar R (2026) AIRBDS: assessing AI readiness in bioscience datasets. Zenodo. https://doi.org/10.5281/zenodo.21803891

Hollingsworth PM, Fantoni K, Koureas D, Arakelyan M, Aravanopoulos FA, Bacela-Spychalska K, Ballesteros-Mejia L, Beja P, Bonchev GN, Bou Dagher Kharrat M, Bourlat SJ, Chkhartishvili T, Čiampor F, Costa FO, Csabai Z, Dankova G, Ekrem T, Emerson B, Ferreira S, Gadawski P, Gkagkavouzis K, Hebert PDN, Ichim MC, Jelić M, Kaitetzidou E, Kalamujić Stroil B, Kamenova S, Keskin E, Laini A, Lawniczak M, Madesis P, Montagna M, Mutanen M, Peters RS, Price B, Rewicz T, Rougerie R, Rulik B, Szucsich N, Vos R, Goodall-Copestake WP, Triantafyllidis A, Grabowski M (2026) Species identification, discovery, and biomonitoring: strategic priorities for DNA barcoding in Europe, set in a global context. BioScience biag052. https://doi.org/10.1093/biosci/biag052

Huertas Olivares CI, Arvanitidis C, McMeel O, Exter K, Hablützel P, Collart T, Leese F, Woollard P, Pade N (2026) eDNAqua-Plan Deliverable 17: Sustainability Plan. Zenodo. https://doi.org/10.5281/zenodo.22145308

Iwasaki W, Fukunaga T, Isagozawa R, Yamada K, Maeda Y, Satoh T, Sado T, Mabuchi K, Takeshima H, Miya M, Nishida M (2013) MitoFish and MitoAnnotator: a mitochondrial genome database of fish with an accurate and automatic annotation pipeline. Molecular Biology and Evolution 30(11): 2531–2540. https://doi.org/10.1093/molbev/mst141

Kaitetzidou E, Gadawski P, Goodall-Copestake WP, Dankova G, Gkagkavouzis K, Holak S, Rewicz T, Bącela-Spychalska K, Mamos T, Fantoni K, Jabłońska A, Tończyk G, Trębicki Ł, Aravanopoulos FA, Bruschini C, Bonchev G, Dagher Kharrat MB, Čiampor F, Costa FO, Dapporto L, Ekrem T, Ferreira S, Geiger M, Hausmann A, Hebert PDN, Kalamujić Stroil B, Kamenova S, Kautmanova I, Keskin E, Kučinić M, Lipinskaya T, Mutanen M, Papakostas S, Price B, Ramírez R, Rougerie R, Rulik B, Szucsich N, Van Der Bank M, Triantafyllidis A, Hollingsworth PM, Grabowski M (2026) Empowering national capacity for DNA-based approach to species identification and biodiversity monitoring. Metabarcoding and Metagenomics 10: e183268. https://doi.org/10.3897/mbmg.10.183268

Kavakiotis I, Woollard P, Babo C, Egge E, Krawczyk D, Laamanen T, Paupério J, Bacela-Spychalska K, Pade N, Costa FO, Grabowski M, Norros V, Meissner K, Papakostas S (2026) Navigating the European aquatic eDNA landscape: opportunities for metadata standardisation and data mobilisation. Metabarcoding and Metagenomics 10: e173612. https://doi.org/10.3897/mbmg.10.173612

Keck F, Couton M, Altermatt F (2023) Navigating the seven challenges of taxonomic reference databases in metabarcoding analyses. Molecular Ecology Resources 23: 742–755. https://doi.org/10.1111/1755-0998.13746

Kõljalg U, Larsson K-H, Abarenkov K, Nilsson RH, Alexander IJ, Eberhardt U, Erland S, Høiland K, Kjøller R, Larsson E, Pennanen T, Sen R, Taylor AFS, Tedersoo L, Vrålstad T (2005) UNITE: a

database providing web-based methods for the molecular identification of ectomycorrhizal fungi. New Phytologist 166(3): 1063–1068. https://doi.org/10.1111/j.1469-8137.2005.01376.x

Kühbandner S, Deister F, Ekrem T, Price B, Stur E, Emerson B, Hollingsworth PM, Vos R, Raupach MJ, Dapporto L, Bordoni A, Bruschini C, Ferreira S, Hausmann A (2026) A browser-based curation tool for expert review of DNA barcode records from BOLD Systems. Research Ideas and Outcomes 12: e191986. https://doi.org/10.3897/rio.12.e191986

Lai J, Liang Q, Zhang X, Liu Y, Wang M, Yang W, Sun T, Li Y, Jin H, Liu Y, Li W, Wu S, Xie Z, Zhou L, Luo M, Zeng L, Yan Q, Feng J, Sun L (2023) FWAlgaeDB, an integrated genome database of freshwater algae. Frontiers in Environmental Science 11(1178097): 1–8. https://doi.org/10.3389/fenvs.2023.1178097

Lavrador AS, Fontes JT, Vieira PE, Costa FO, Duarte S (2023) Compilation, revision, and annotation of DNA barcodes of marine invertebrate non-indigenous species (NIS) occurring in European coastal regions. Diversity 15(2): 1–16. https://doi.org/10.3390/d15020174

Lefler FW, Berthold DE, Laughinghouse IV HD (2023) Cyanoseq: a database of cyanobacterial 16S rRNA gene sequences with curated taxonomy. Journal of Phycology 59: 470–480. https://doi.org/10.1111/jpy.13335

Leite BR, Vieira PE, Teixeira MAL, Lobo-Arteaga J, Hollatz C, Borges LMS, Duarte S, Troncoso JS, Costa FO (2020) Gap-analysis and annotated reference library for supporting macroinvertebrate metabarcoding in Atlantic Iberia. Regional Studies in Marine Science 36: 101307. https://doi.org/10.1016/j.rsma.2020.101307

Leray M, Knowlton N, Machida RJ (2022) MIDORI2: a collection of quality controlled, preformatted, and regularly updated reference databases for taxonomic assignment of eukaryotic mitochondrial sequences. Environmental DNA 4: 897–907. https://doi.org/10.1002/edn3.303

Maidak BL, Olsen GJ, Larsen N, Overbeek R, McCaughey MJ, Woese CR (1996) The Ribosomal Database Project (RDP). Nucleic Acids Research 24(1): 82–85. https://doi.org/10.1093/nar/24.1.82

Marques V, Milhau T, Albouy C, Dejean T, Manel S, Mouillot D, Juhel J (2021) GAPeDNA: assessing and mapping global species gaps in genetic databases for eDNA metabarcoding. Diversity and Distributions 27: 1880–1892. https://doi.org/10.1111/ddi.13142

McDonald D, Jiang Y, Balaban M, Cantrell K, Zhu Q, Gonzalez A, Morton JT, Nicolaou G, Parks DH, Karst SM, Albertsen M, Hugenholtz P, DeSantis T, Song SJ, Bartko A, Havulinna AS, Jousilahti P, Cheng S, Inouye M, Niiranen T, Jain M, Salomaa V, Lahti L, Mirarab S, Knight R (2024) Greengenes2 unifies microbial data in a single reference tree. Nature Biotechnology 42: 715–718. https://doi.org/10.1038/s41587-023-01845-1

Meglécz E (2023) COInr and mkCOInr: building and customizing a nonredundant barcoding reference database from BOLD and NCBI using a semi-automated pipeline. Molecular Ecology Resources 23: 933–945. https://doi.org/10.1111/1755-0998.13756

Molano L-AG, Vega-Abellaneda S, Manichanh C (2024) GSR-DB: a manually curated and optimized taxonomical database for 16S rRNA amplicon analysis. mSystems 9(2): 1–16. https://doi.org/10.1128/msystems.00950-23

Pentinsaari M, Ratnasingham S, Miller SE, Hebert PDN (2020) BOLD and GenBank revisited – do identification errors arise in the lab or in the sequence libraries? PLoS ONE 15(4): e0231814. https://doi.org/10.1371/journal.pone.0231814

Quast C, Pruesse E, Yilmaz P, Gerbe J, Schweer T, Yarza P, Peplies J, Glöckner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research 41(D1): D590–D596. https://doi.org/10.1093/nar/gks1219

Radulovici AE, Vieira PE, Duarte S, Teixeira MAL, Borges LMS, Deagle BE, Majavena S, Redmond N, Schultz JA, Costa FO (2021) Revision and annotation of DNA barcode records for marine invertebrates: report of the 8th iBOL conference hackathon. Metabarcoding and Metagenomics 5(e67862): 207–217. https://doi.org/10.3897/mbmg.5.67862

Ratnasingham S, Hebert PDN (2013) A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. PLoS ONE 8(7): e66213. https://doi.org/10.1371/journal.pone.0066213

Ratnasingham S, Wei C, Chan D, Agda J, Agda J, Ballesteros-Mejia L, Botou HA, El Bastami ZM, Ma E, Manjunath R, Rea D, Ho C, Telfer A, McKeowan J, Rahulan M, Steinke C, Dorsheimer J, Milton M, Hebert PDN (2024) BOLD v4: a centralized bioinformatics platform for DNA-based biodiversity data. In: DeSalle R (Ed.) DNA Barcoding. Methods in Molecular Biology, vol. 2744. Humana, New York, 403–441. https://doi.org/10.1007/978-1-0716-3581-0_26

Rimet F, Gusev E, Kahlert M, Kelly MG, Kulikovskiy M, Maltsev Y, Mann DG, Pfannkuchen M, Trobajo R, Vasselon V, Zimmermann J, Bouchez A (2019) Diat.barcode, an open-access curated barcode library for diatoms. Scientific Reports 9(15116): 1–12. https://doi.org/10.1038/s41598-019-51500-6

Rimet F, Aylagas E, Borja A, Bouchez A, Canino A, Chauvin C, Chonova T, Ciampor Jr F, Costa FO, Ferrari BJD, Gastineau R, Goulon C, Gugger M, Holzmann M, Jahn R, Kahlert M, Kusber W-H, Laplace-Treyture C, Leese F, Leliaert F, Mann DG, Marchand F, Méléder V, Pawlowski J, Rasconi S, Rivera S, Rougerie R, Schweizer M, Trobajo R, Vasselon V, Vivien R, Weigand A, Witkowski A, Zimmermann J, Ekrem T (2021) Metadata standards and practical guidelines for specimen and DNA curation when building barcode reference libraries for aquatic life. Metabarcoding and Metagenomics 5: 17–33. https://doi.org/10.3897/mbmg.5.58056

Rimet F, Costa F, Woollard P, Probert I, Boulanger E (2026) eDNAqua-Plan Deliverable 9: recommendations for a federated system of taxonomically curated reference libraries. Zenodo. https://doi.org/10.5281/zenodo.18849393

Roush D, Giraldo-Silva A, Garcia-Pichel F (2021) Cydrasil 3, a curated 16S rRNA gene reference package and web app for cyanobacterial phylogenetic placement. Scientific Data 8(230): 1–6. https://doi.org/10.1038/s41597-021-01015-5

Schultz JA, Hebert PDN (2024) Marine ecozones represent a major gap in Canadian DNA barcoding efforts. FACETS 9: 1–11. https://doi.org/10.1139/facets-2024-0015

Tapolczai K, Rimet F, Ćirić M, Ballot A, Laplace-Treyture C, Alric B (2025) A novel framework for phytoplankton biomonitoring: trait assignment of 23S rRNA sequences. Ecological Indicators 173(113361): 1–13. https://doi.org/10.1016/j.ecolind.2025.113361

Tedersoo L, Moghaddam MSH, Mikryukov V, Hakimzadeh A, Bahram M, Nilsson RH, Yatsiuk I, Geisen S, Schwelm A, Piwosz K, Prous M, Sildever S, Chmolowska D, Rueckert S, Skaloud P, Laas P, Tines M, Jung J-H, Choi JH, Alkahtani S, Anslan S (2024) EUKARYOME: the rRNA gene reference database for identification of all eukaryotes. Database 2024: baae043. https://doi.org/10.1093/database/baae043

Turón X, Wangensteen OS, Palacín C, Cebrían E (2020) Caracterización genética de la biodiversidad y sus alteraciones en comunidades bentónicas marinas de los parques nacionales. In: Proyectos de Investigación en Parques Nacionales 2013–2017. Organismo Autónomo Parques Nacionales, 349–368.

Wangensteen OS, Turón X (2016) Metabarcoding techniques for assessing biodiversity of marine animal forests. In: Rossi S, Bramanti L, Gori A, Orejas Saco del Valle C (Eds) Marine Animal Forests. Springer, Cham. https://doi.org/10.1007/978-3-319-17001-5_53-1

Weigand H, Beermann AJ, Čiampor F, Costa FO, Csabai Z, Duarte S, Geiger MF, Grabowski M, Rimet F, Rulik B, Strand M, Szucsich N, Weigand A, Willassen E, Wyler S, Bouchez A, Borja A, Čiamporová-Zaťovičová Z, Ferreira S, Dijkstra KD, Eisendle U, Freyhof J, Gadawski P, Graf W, Haegerbaeumer A, van der Hoorn BB, Japoshvili B, Keresztes L, Keskin E, Leese F, Macher JN, Mamos T, Paz G, Pešić V, Pfannkuchen DM, Pfannkuchen MA, Price BW, Rinkevich B, Teixeira MAL, Várbíró G, Ekrem T (2019) DNA barcode reference libraries for the monitoring of aquatic biota in Europe: gap-analysis and recommendations for future work. Science of the Total Environment 678: 499–524. https://doi.org/10.1016/j.scitotenv.2019.04.247

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3: 160018. https://doi.org/10.1038/sdata.2016.18

Yuan D, Ahamed A, Athar A, Devaraj R, Gupta D, Haseeb M, Ihsan M, Ivanov E, Kadhirvelu V, Khen A, Kumar M, Lathi A, Liyanage I, Meszaros L, O'Cathail C, Pauperio J, Paz R, Pesant S, Rahman N, Rajan J, Tutis I, Ventouratou M, Vijayaraja S, Waheed Z, Woollard P, Burdett T, Cochrane G, Sarkans U (2026) The European Nucleotide Archive in 2025. Nucleic Acids Research 54(D1): D120–D127. https://doi.org/10.1093/nar/gkaf1295

Zarcero J, Antich A, Rius M, Wangensteen OS, Turón X (2024) A new sampling device for metabarcoding surveillance of port communities and detection of non-indigenous species. iScience 27(1): 1–14. https://doi.org/10.1016/j.isci.2023.108588

Zhu T, Sato Y, Sado T, Miya M, Iwasaki W (2023) MitoFish, MitoAnnotator, and MiFish Pipeline: updates in 10 years. Molecular Biology and Evolution 40(3): 1–5. https://doi.org/10.1093/molbev/msad035