eDNAqua-Plan

NEXT GENERATION OF AQUATIC BIODIVERSITY MONITORING

The eDNAqua-Plan proposal towards a future of federated, curated and reliable reference libraries

Frédéric Rimet, Ian Probert, Filipe Costa, Peter Woollard, Saara Suominen, **Emilie Boulanger**

LIVING DATA 2025

21/10/2025

Universidade do Minho

OBIS OCEAN BIODIVERSITY INFORMATION SYSTEM

unesco Intergovernmental Oceanographic Commission

SORBONNE UNIVERSITÉ

EMBL-EBI

Funded by the European Union

Funded by the European Union under the Horizon Europe Programme, Grant Agreement No. 101112800 (eDNAqua-Plan). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

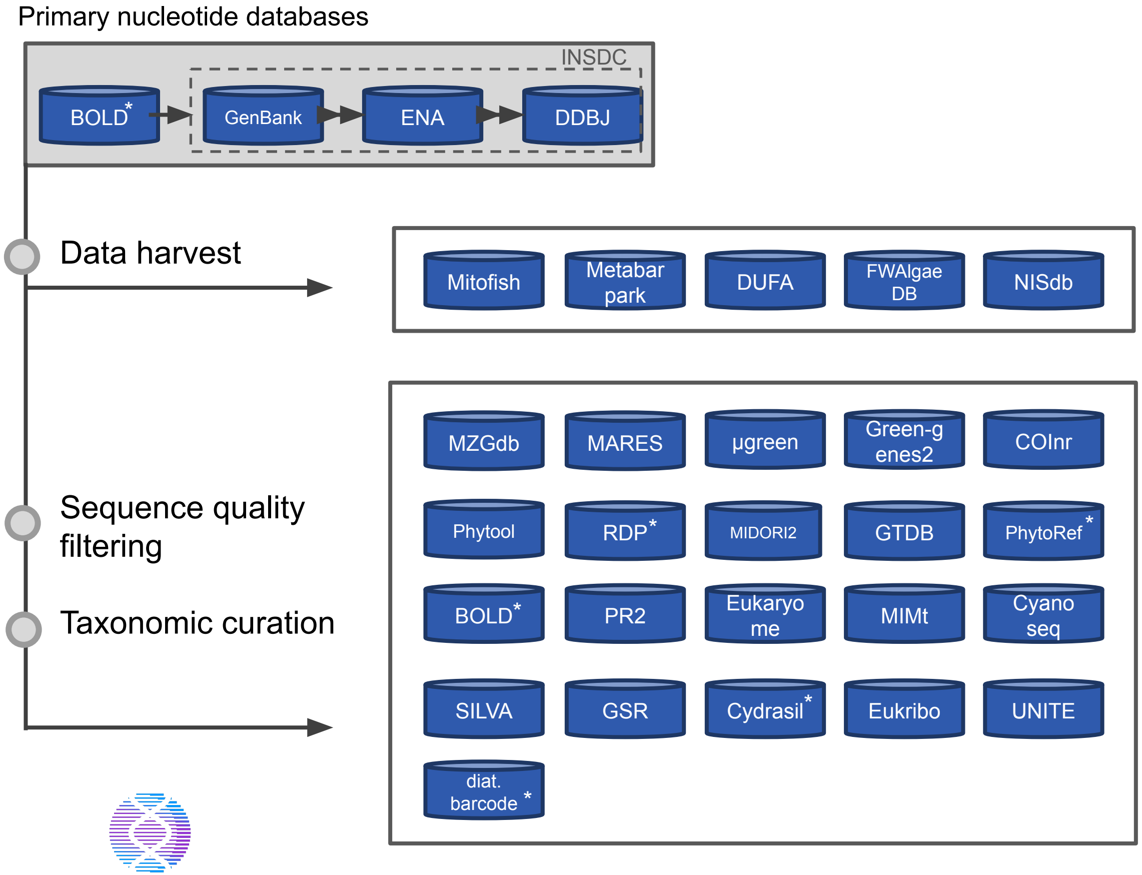
The importance of curated reference sequences

- DNA barcoding & metabarcoding from specimen or environmental samples increasingly applied to biomonitoring
- Reliable identifications need reliable (and comprehensive) reference sequences
- Inaccurate, incomplete, or inconsistently curated databases

→ misidentification, false positives, underestimated diversity

Reliable, taxonomically curated reference libraries provide the foundation for accurate species identification, ecological inference, and biogeographical comparison.

The current landscape

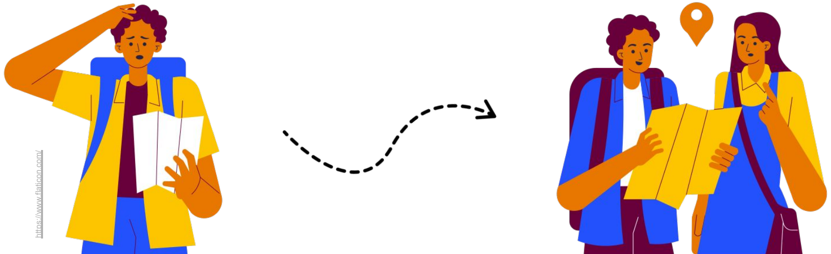


- Important effort in data harvest for specific markers and taxonomic groups, additional sequence quality filtering, verification and curation of taxonomic names
- But no global agreement on curation procedures
- Few carry out full curation procedures & are consistently maintained
- Because of their disparity, landscape can be difficult to navigate

Costa, F., et al. *in prep*

eDNAqua-Plan: A plan for a European DNA digital ecosystem for the next generation of aquatic biodiversity monitoring

- Focus on reference libraries and eDNA data
- Evaluating the existing data infrastucture
- Planning a digital ecosystem that will be fully transparent and interoperable



https://www.flaticon.com/

The eDNAqua-Plan approach

WP2

Landscape analysis

Sep 2024

Reference Libraries and eDNA data

WP4

Use cases

WP3

Discussions

Oct 2024

Workshop: discussion on standards requirements

Jan 2025

Workshop: conceptualise the eDNA landscape

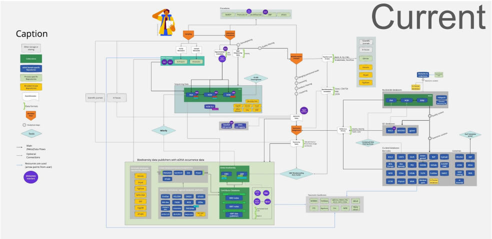
Landscape proposal

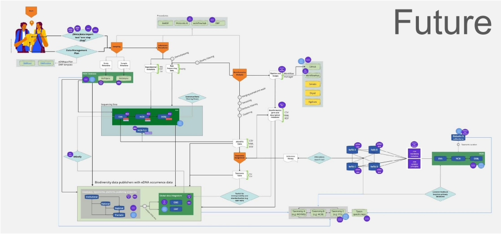
Apr 2025

Detailed Doc

Jul 2025

Publish v1





Recommendations

Feb 2026

Good Practice reports

- Reference libraries
- eDNA data standards
- Connectivity

WP5

Oct 2025

Workshop: Stakeholder input

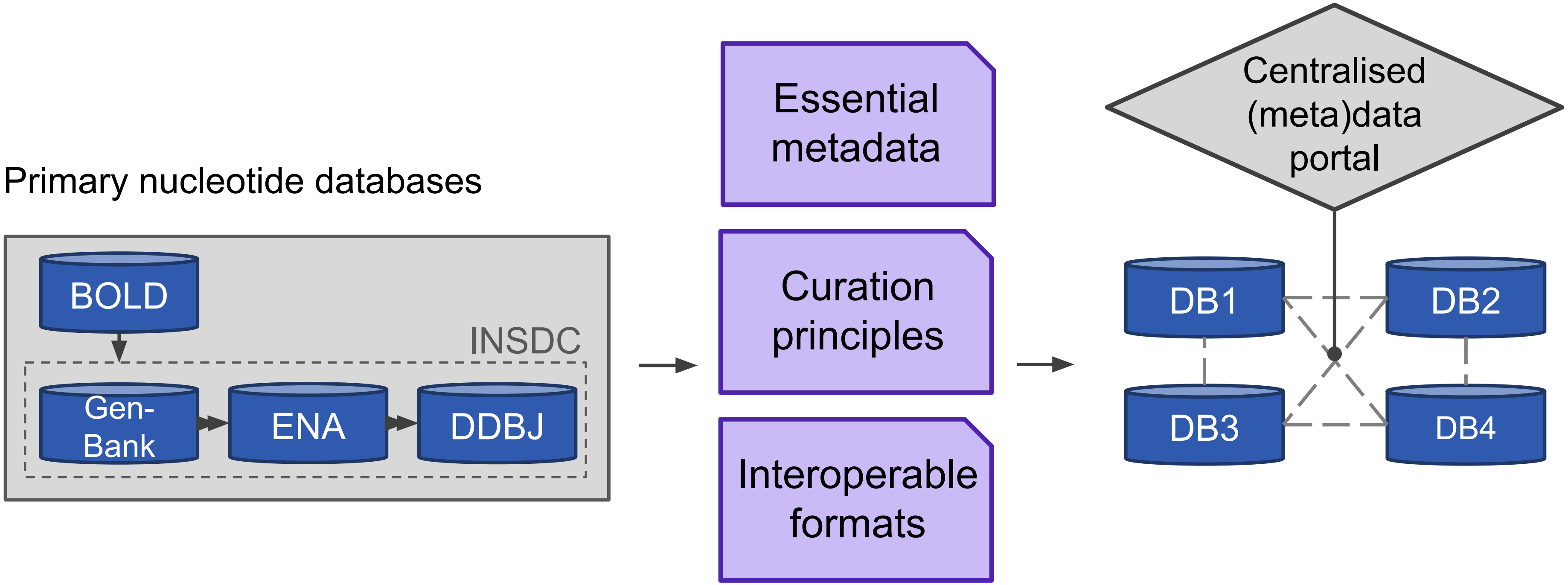
The vision for the future:
a federated system of curated reference libraries

- All taxonomic assignments of eDNA data are done using quality-controlled, taxonomically curated and thus reliable reference databases.
- These are easy to find and connected to each other (mapped to each other as well).
- It’s easy for users to know whether any reference sequence follows a curation and quality standard.
- Not one large taxonomically-curated DNA reference database of aquatic organisms of Europe (not feasible, necessary or desirable) but a diverse portfolio of independent databases.

Each of the current curated databases are created and maintained by different groups or institutions and each of these databases have their own focus, specialisation, taxonomic specialists and management. This diversity in existing databases should be maintained.

- A network of open linked data resources, QC & curated reference libraries.
- A federated (meta)data search engine to find the right (standard-approved) library and sequence.

The vision for the future: a federated system of curated reference libraries



A set of essential metadata

Essential metadata

- Reference libraries should provide an assessment of the confidence level in taxonomic assignment of barcodes
- Allow for relative restriced set of fields for end-users to:
  - Easily identify whether the taxonomic assignment linked to a barcode is trustworthy
  - Filter database entries by criteria. Geographic location, habitat type, environmental parameters, etc.

→ Each barcode record should have standardized, traceable, and verifiable metadata describing the **biological origin, sequence quality,** and **taxonomic validation** of the reference.

A set of essential metadata (1-2/4)

Essential metadata

We propose the following set of metadata that is considered **obligatory** or recommended for reference libraries:

BIOLOGICAL ORIGIN

- **Specimen identifier**
- Repository
- **Sample type**
- **Date of collection**
- **Geographic origin**
- **Habitat / environment type**
- **Depth / altitude**
- **Temperature** and/or **salinity**, when relevant
- Sample permits? (e.g. ABS)
- Specimen identification
- Original identification method
- Type material

SEQUENCE QUALITY

- **Sequence accession number**
- **Source database name**
- **Source database version**
- **Marker / gene region**
- **Sequence length**
- Sequencing platform
- **Quality metrics**
- Quality control
- BLAST nearest match
- Link to associated publication or project

A set of essential metadata (3-4/4)

Essential metadata

We propose the following set of metadata that is considered **obligatory** or recommended for reference libraries:

TAXONOMIC VALIDATION

- **Curation status**
- **Curator / expert validator name** (+ institute / contact details?)
- **Date of expert validation**
- **Curation method** (phylogeny, cluster, etc. - need defined list?)
- **Confidence level method** (defined list?)
- **Confidence level score/rank**
- **Scientific name** (current valid name according to an accepted taxonomy)
- **Taxonomic rank(s)**
- **Taxonomic source**
- Synonyms / previous names

LIBRARY METADATA

- **Library name**
- **Library version number**
- **Library version data**

Common guiding curation principles

Curation principles

Step 1. Pre-curation, Sequence quality check

Step 2. Taxonomic curation

To homogenise taxonomic names for a given phylogenetic clade (or for identical / near-identical sequences), the following methods:

1. **Expert curation based on phylogenies.** If neighbouring sequences have non-homogeneous names, check for
   a. synonyms in scientific publications
   b. taxonomic databases
   c. check metadata
2. **Expert curation based on clustering genetically similar sequences.** Check non-homogeneous names following a, b, c.
3. **Expert curation based on self assignation** + a, b, c.
4. **Automated curation procedures based on bioinformatics or AI.** Semi-automated to fully automated as developments progress, with human intervention for clarifying intricacies at low taxonomic levels. e.g. Diat.barcode, UNITE, BAGS, BGE library curation tool

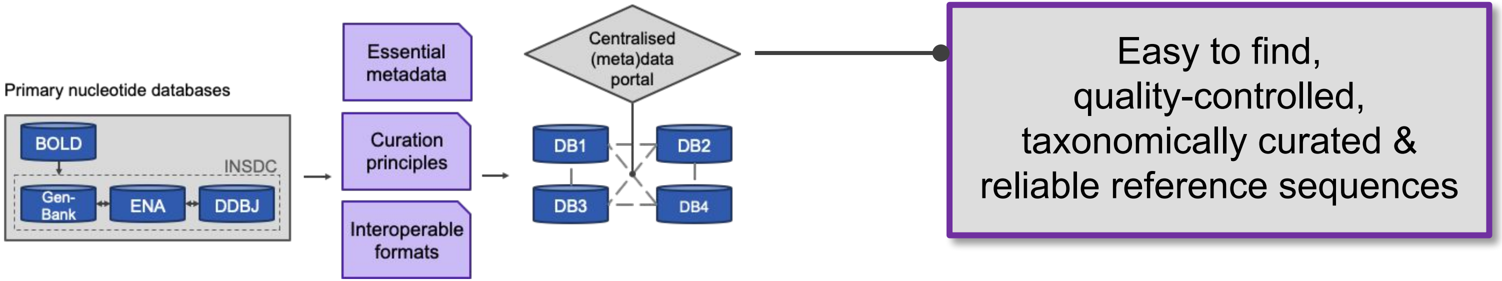
Common guiding curation principles

Interoperable formats

At the sequence/library level

- Following **interoperable metadata standards**, consisting of or aligned with existing standards (MIxS, environmental checklists,)
- This includes what **metadata fields** we believe are **essential for interoperability** (e.g. use of taxonID)
- Using m**achine-readable fields and formats** (JSON, XML)

The achieve the vision for the future



Easy to find, quality-controlled, taxonomically curated & reliable reference sequences

eDNAqua-Plan

NEXT GENERATION OF AQUATIC BIODIVERSITY MONITORING

Emilie Boulanger
e.boulanger@unesco.org

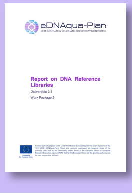
Funded by the European Union

More on

zenodo

eDNAqua-Plan

http://ednaquaplan.com/



Report on DNA Reference Libraries

Deliverable 2.1

Work Package 2

The eDNAqua-Plan conceptual landscape proposal v1

Milestone 3.3

Work Package 3

More at Living Data 2025:

Improving data management strategies, sharing and FAIRness of DNA-derived Biodiversity Data (Part 1) 6796330-1 Ballroom B2

21/10 3:15 PM Camila Babo - Improving interoperability towards FAIR eDNA by aligning data standards

Posters EVT-POSTERS

22/10 5:00 - 6:00 PM Our vision for a future eDNA digital ecosystem