Ref. Ares(2023)2125783 - 23/03/2023

Horizon Europe

Data Management Plan

Version 1.0
23 March 2023

HISTORY OF CHANGES		
Version	**Publication date**	**Changes**
1.0	23.03.2023	▪ Initial version

QUALITY CONTROL		
Activity	**Name**	**Date**
V1.0 created	Dr Miloš Ćirić, Dr Srđan Miletić, Dr Vladimir Petrović	24.02.2023
V1.0 reviewed	Dr Aleksandra Marković, Dr Susanne Schneider, Dr Andreas Ballot, Dr Agnès Bouchez, Dr Frédéric Rimet, Hervé Dumont	10.03.2023
V1.0 revised	Dr Miloš Ćirić	23.03.2023

Action Number: [101079234]

Action Acronym: [BIOLAWEB]
Action title: [Boosting Institute of Chemistry, Technology and Metallurgy in Water Biomonitoring]

Date: [23.03.2023]

DMP version: [DMP 1.0]

Table of Contents

1. Data Summary

1.1. Re-use of existing data

The BIOLAWEB project will strongly rely on DNA sequences available in international OA databases such as the reference barcoding library Diat.barcode and GenBank. These data are neccessary to compare the results generated in the BIOLAWEB project.

1.2. Generation of new data

The BIOLAWEB project will generate new DNA sequences of diatom, phytoplankton and macrophyte species. BIOLAWEB will collect data on water chemistry, macrophyte, benthic diatom and phytoplankton assemblages from study lakes. BIOLAWEB will also collect data from internal surveys (questionnaires and interviews).

1.3. Data types, purpose, formats, size, origin, and utility outside the project

BIOLAWEB datasets will include (i) quantitative and qualitative physical, chemical, hydrological and species data (species names and abundances of algae and macrophytes, including mapping data), (ii) DNA sequences from metabarcoding of the samples collected, and (iii) relevant third-party data from stakeholders. Table 1 summarizes data types linked with their purpose and format in which the data will be stored during the BIOLAWEB project.

2. FAIR data

2.1. Making data findable, including provisions for metadata

Persistent identifier

The BIOLAWEB consortium will store generated and collected data in an open online research data repository. The BIOLAWEB team decided to use a digital repository of the Institute of Chemistry, Technology and Metallurgy, University of Belgrade, CeR - Central Repository ICTM (https://cer.ihtm.bg.ac.rs/) as its data archive, based on the compliance of the repository structure and facilities with the FAIR data principles. This CeR repository allows researchers to deposit both publications and data, providing tools to linking them to these through **persistent identifiers** and data citations.
Quality assured sequence data for diatoms, phytoplankton, and macrophytes will be submitted to GenBank, an annotated collection of all publicly available DNA sequences.

Metadata types and standards

CeR repository provides rich metadata about researchers' publications, posters, and presentations (e.g., around 29 metadata for each publication including file name, size, format, date accessioned, date available, rights, license, for mapping results – lake name, organism group, method used for mapping/counting/measuring, coordinates, sampling date, etc.).

Quality assured macrophyte sequence data will be submitted to GenBank. All data will contain the necessary metadata required by GenBank, including species name, sampling location, collection date, publication reference and a persistent identifier for each entry. GenBank is an annotated collection of all publicly available DNA sequences, so that the quality-assured data of macrophytes generated in BIOLAWEB can be easily found and accessed.

Sequence data will be also integrated into the reference barcoding library "Diat.barcode". For each sequence, all necessary metadata are given and follow the CEN TR 17244 (CEN, 2018. Water quality - CEN/TR 17244 - Technical report for the management of diatom barcodes 1–11.). Diat.barcode is an open-access library available at https://doi.org/10.15454/TOMBYZ.

Table 1. BIOLAWEB project data summary

Data type	Objective	Description	Data utility*
Research datasets	**Objective 4.** To scale up and develop new approaches in the field of biomonitoring through joint research with leading EU partners	Purpose: To collect information on diatoms and phytoplankton composition in the sampled lakes using morphological and genetic approach; to develop a novel method to detect the occurrence of macrophytes from a water sample. Data origin: Sampling reports, list of species, HTS sequencing data for diatoms, phytoplankton, macrophytes and macrophyte mapping data from the same lakes; peer-reviewed scientific articles and conference papers. Format: .xls, .pdf, .docx, .jpg, raw sequence reads (fastq, fasta) Expected size: 40MB to GB (depending on number of sequences)	Data are primarily restricted to BIOLAWEB consortium. After consortium made data openly available (OA), they can be used by the scientific community (e.g. within freshwater ecology), including quality assured sequence published in GenBank etc. The data from "traditional" and genetic macrophyte mapping will be published in scientific publications and in the digital repository of UB-ICTM.
Personal datasets	**Objective 2**. To significantly raise and scale up UB-ICTM's staff competence and knowledge on how indices for the EU WFD are developed, intercalibrated and applied **Objective 3.** To raise the competence and skills of ICTM researchers in DNA-based biomonitoring methods **Objective 5.** To set up a fully operational IRMO** at UB-ICTM **Objective 6.** To considerably enhance strategic networking activities between UB-ICTM and internationally leading research institutions through sustainable partnerships	Purpose: To collect information for preparing consortium and internal meetings; to collect data of participants for workshops, trainings and roundtables; to collect information on potential candidates for the new office; to collect quantitative data on visitors of BIOLAWEB website, social media, Festival of Science and Researchers' night. Data origin: List of participants, pictures, promo videos; interview's reports, application material; Format: docx, .pdf,.xlsx, .ppt, .png, .jpg, .mp4, Web site, Google Analytics, .html Expected size: 10 GB	All files from internal UB-ICTM meetings are restricted to BIOLAWEB consortium. Pictures and promo videos from kick-off/consortium meetings will be made openly available. Data related to potential candidates for Research Management Office will be strictly and confidentially used by Human Resources Department of UB-ICTM. In case personal data is needed they will be anonymized.
Internal strategy datasets	**Objective 1.** To substantially raise scientific excellence and innovation capacity by developing a tailor-made scientific strategy for UB-ICTM	Purpose: To collect data on UB-ICTM's internal structure, research performance, and knowledge gaps and needs in the fields of aquatic ecology/ biomonitoring; Data origin: Attendance list, pictures, presentations, expert visit reports, questionnaires, interview's reports, Science Strategy and Action Plan. Format: .pdf, .docx, .ppt, .png, .jpg Expected size: 100 MB	Restricted to BIOLAWEB consortium.
Training materials datasets	**Objective 2**. To significantly raise and scale up UB-ICTM's staff competence and knowledge on how indices for the EU WFD are developed, intercalibrated and applied **Objective 3.** To raise the competence and skills of UB-ICTM researchers in DNA-based biomonitoring methods **Objective 5.** To set up a fully operational IRMO** at UB-ICTM	Purpose: to collect data for training in the bioindication (index development) and DNA-based biomonitoring methods thanks to expert visits, workshops, on-site trainings (STSM's); to collect data for training for administrative and management staff. Data origin: Presentations, teaching material Format: .ppt, .pdf, .docx, .xlsx, .jpg, .mp4 Expected size: 10 GB	Training materials from workshops, expert visits etc. will be made openly available (including participants of the courses). A somewhat reduced version (to respect copyrights) will be made publicly available.

*outside the project

**International Cooperation and Project Office (ICPO) is a new and officially accepted name (with acronym) for the established office. IRMO was an initial acronym used in the project proposal.

2.2. Making data accessible

Repository

BIOLAWEB consortium identified the CeR as a trusted repository suitable for storing datasets generated during the project lifetime. It is based on the platform DSpace (https://duraspace.org/dspace/) which is one of the most popular platforms for open science. Within the CeR, BIOLAWEB project data will be deposited and organized in a specific part of the repository (catalog) intended for international projects to facilitate data re-use. CeR repository ensures that the data is assigned an identifier and resolves the identifier to a digital object.

GenBank is an annotated collection of all publicly available DNA sequences, a well-established international database, so that the quality-assured data generated in BIOLAWEB can be easily found and accessed.

Embargo and restricted access

Following the HORIZON EUROPE guidelines scientific data will be made available for re-use (the sequence data, the mapping data, and the water chemistry) as well as workshops' materials. On the other hand, in case personal data is needed they will be anonymized, and restricted to the BIOLAWEB consortium (Table 2).

Table 2. Description of internal data created during the BIOLAWEB project

Data type	Utility	Description and purpose
Internal strategy datasets	Datasets will be used for developing Science and Innovation Strategy and Action Plan (Objective 1)	All generated data including expert interviews and filled questionnaires will not be published due to the privacy reasons.
Personal datasets	To make a list of potential workshops, trainings and roundtables attendees and their expectations (Objectives 2, 3 and 5).	Attendance lists and filled questionnaires (workshops, trainings, roundtables, etc.) will be restricted to the BIOLAWEB consortium

BIOLAWEB researchers will publish their work in scientific journals with large audiences and aim to publish Gold Open Access (OA) wherever possible (research data immediate availability). According to the GA, the option for using repository-based or Green OA for some project publications is also forseen. In addition, the list of project publications with links to the publications and the selected supplementary data will be uploaded to BIOLAWEB website. According to the established access protocol within CeR repository, the BIOLAWEB research data will be accessible either free (biological, chemical and sequence data) or restricted only for registered users (available to the research staff of the institute only). The identity of the person accessing the data is linked with her/his institutional e-mail. At the moment, there is no need for a data access committee.

Metadata

In CeR repository, metadata associated with scientific publications, project presentations and reports will be openly available and licenced under a public domain (CC0) during the project lifetime and at least five years after the project ends.

GenBank is an annotated collection of all publicly available DNA sequences, and metadata are provided in a standardized way. This ensures that all macrophyte sequence data will be provided with the necessary metadata including species name, sampling location, collection date, publication reference and a persistent identifier for each entry. The BIOLAWEB team will ensure that all data will be made publicly available in GenBank and can be found immediately after publication.

2.3. Making data interoperable

The BIOLAWEB team will deposit data and related metadata in the open repository CeR which fulfils the criterion of interoperability. This repository uses general standard Dublin Core which will make BIOLAWEB data interoperable with other services that provide or offer metadata about scientific publications. The use of uncommon or creation of specific vocabularies is not foreseen for the project data, but this point could be discussed or reassessed in the consortium meetings.

All data submitted to Genbank follow the community-endorsed standard and therefore are in an interoperable format.

The sequencing data will be analysed using reference barcoding databases developed in former research programs (Diat.barcode and Phytool). These reference libraries are used by a worldwide audience of scientists working on algal metabarcoding.

2.4. Increase data re-use

Bioinformatic pipelines and statistical analyses will be published on Open Access platforms (GitHub, GenBank, entrepot.recherche.data.gouv.fr). For example, GenBank accession numbers will be provided in all publications using these entries, thereby ensuring maximum re-use of the data.

The BIOLAWEB will make research publications available through deposition of a copy of the published, or final, peer-reviewed version, in identified suitable institutional repositories. Datasets uploaded in the CER repository will be freely accessible, immediately in case of Gold Open Access (OA), or after an embargo period - Green Open Access. Potential users are expected to adhere with the CeR Terms of Use. Also, BIOLAWEB will publish the project results (list of scientific publications, reports and other results) on the project website as soon as possible, taking into account the EC deadline.

Datasets generated in BIOLAWEB project and deposited in the CeR repository under an Open Access license can be used without restrictions by third parties up to 5 years after the end of the project. Data deposited in GenBank are freely available.

The provenance of research data generated in BIOLAWEB project will be tracked and those responsible for its generation will be adequately credited for their work using appropriate Creative Commons copyright licenses. Data submitted to GenBank follow a community endorsed standard and will therefore be thoroughly documented.

Data quality assurance

The members of the Project Executive Board approve data collection and deposition in the open repository by checking the following critical data quality dimensions: 1. **Integrity** (data stability during the lifecycle); 2. **Validity** (data items can be traced); 3. **Accuracy** (data fits reality); 4. **Consistency** (values in former and latter datasets are consistent); 5. **Timeliness** (the time between expected and actual data availability); 6. **Currency** (data is up-to-date) and 7. **Completeness** (no missing values in table/box fields). After the end of the project, this role will be transferred to the repository administrator at UB-ICTM.

In general, only quality-assured data will be published and submitted to GenBank, thereby ensuring that only useful data will be made publicly available. Metabarcoding and sequence data will be screened by using the suitable software.

3. Other research outputs

Diatom samples collected in BIOLAWEB project will be deposited in the Diatom Collection of Serbia (DCSR) at the University of Belgrade, Institute of Chemistry, Technology and Metallurgy, National Institute of the Republic of Serbia.

Teaching material generated in WP3 will be made available on our website, and as project deliverable. In some cases, we will make a slightly reduced version available, in order to avoid copyright issues (generally related to illustrations which may be used for teaching purposes only).

In the framework of BIOLAWEB project, protocols that will be used for diatom and phytoplankton metabarcoding will be protocols already developped in other projects of INRAE. BIOLAWEB members will refer in deliverables to these protocols which are open-accessible on platforms of INRAE. If changes are made to these protocols, BIOLAWEB members will notify it and will make these modified protocols open-accessible. The protocols for macrophytes generated in the project will be made openly available via our website, and as project deliverable.

4. Allocation of resources

No costs for curation of data related to DCSR are foreseen. Submission of data to GenBank is part of the publication process and is free of charge. No costs are incurred for further archiving.

Data protection and long-term storage

NIVA is responsible for data management of the macrophyte sequence data. The Data Protection Officer (DPO) will be responsible for overall data management in the BIOLAWEB project.

The Data Protection Officer (DPO) of the BIOLAWEB project is:

Dr Vladimir Petrović,
University of Belgrade
National Institute of the Republic of Serbia
Institute of Chemistry, Technology and Metallurgy
Department of Ecology and TechnoEconomics
Njegoševa 12, 11000 Belgrade, Serbia
Phone : +381 60 5866885
E-mail : petrovic@ihtm.bg.ac.rs

The Data Protection Officer (DPO) of the BIOLAWEB project is responsible for:

- the highest standards of the security measure to prevent unauthorized access to personal data. DPO should follow data minimization principle, where all data that are not relevant will be immediately deleted from the record;
- the storage of the data, collection, recording, organization, structuring restriction, erasure, availability, and destruction of the data;
- ensuring that essential documents are not destroyed before the end of the required retention of 5 years and that after that they will be destroyed in their entirety making sure that they cannot be recovered. Data retained for auditing processes will be stored securely and further processed for those purposes only;

The coordinating institution of the BIOLAWEB project (UB-ICTM) will use CeR for long-term storage research data and metadata produced in BIOLAWEB. Data will be machine-readable, citable, published in a registered long-term open access repositories and interlinked with other project outcomes. Each dataset will be accompanied by a full metadata description, explaining their contents, origin, methods, licensing, citation, etc.

Data submitted to GenBank are foreseen to be archived “for eternity,” and necessary resources are provided by the international scientific community.

5. Data security

During the implementation of the BIOLAWEB project, datasets, either generated or re-used, will be archived on the storage system of the beneficiaries. Each partner involved in collecting or analysing data is responsible for their secure storage and transfer, and follows national and institutional guidelines with respect to data security. Datasets that must be kept confidential are BIOLAWEB internal survey data and documents related to the Science Strategy and Action Plan (D2.2). Sensitive data will not be made publicly available.

GenBank is part of the International Nucleotide Sequence Database Collaboration. It is a long-standing foundational initiative that operates between DDBJ (DNA Databank of Japan), EMBL-EBI (EMBL’s European Bioinformatics Institute) and NCBI (National Center for Biotechnology Information, USA). Long-term storage and data security therefore are ensured in the best possible way.

After the end of the project, the responsibility for secure long-term preservation and curation will be transferred to the repositories storing the dataset.

6. Ethics

It is planned to collect local resources in the Republic of Serbia that might include samples of endangered flora. The project will comply with Serbian law with respect to sample collection. This activity is related to WP4 (Research and partnership) and WP3 (Training and networking). Samples containing diatoms, phytoplankton, and macrophytes including *Chara* species will be transferred from Serbia to France and from Serbia to Norway. They will be used only for the purpose of the project (material for training, material used for contributing new sequences to the reference barcoding library, material used for scientific publications). Each Partner is responsible for compliance with the Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity.

Questionnaires and other surveys dealing with personal data used in the project will include a disclaimer informing respondents on the purpose of data collection and the way the information is going to be further used within the scope of the project (e.g., confidentiality).

7. Other issues

No other procedures for data management are foreseen.