

SAFRED

Saving Freshwater Biodiversity Research Data

Aaike De Wever (RBINS)¹ - Pieter Lemmens (KULeuven)² - Tanja Milotic (INBO)³ - Astrid Schmidt-Kloiber (BOKU)⁴ - Koen Martens (RBINS)¹

¹ Operational Directorate Natural Environment, Royal Belgian Institute of Natural Sciences, Brussels, Belgium

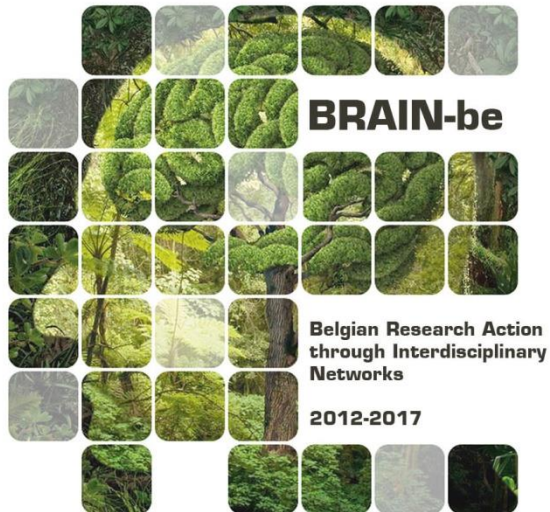
² Ecology Evolution Biodiversity Conservation, KULeuven, Leuven, Belgium

³ Research Institute for Nature and Forest, Brussels, Belgium

⁴ Institute of Hydrobiology & Aquatic Ecosystem Management, BOKU, University of Natural Resources & Life Sciences

Axis 6: Management of collections





NETWORK PROJECT

SAFRED

Saving Freshwater Biodiversity Research Data

Contract - BR/154/A6/SAFRED

FINAL REPORT

PROMOTORS: Koen Martens (RBINS)
Luc De Meester (KULeuven)
Elie Verleyen (UGent)
Annick Wilmotte (ULg)
Daniël Du Seuil (INBO)
Patrick Kestemont (UNamur)
André Heughebaert (BBPf)
Astrid Schmidt-Kloiber (BOKU)

AUTHORS: Aaike De Wever (RBINS)
Pieter Lemmers (KULeuven)
Tanja Milotic (INBO)
Astrid Schmidt-Kloiber (BOKU)
Koen Martens (RBINS)



Published in 2018 by the Belgian Science Policy Office
Avenue Louise 231
Louizalaan 231
B-1050 Brussels
Belgium
Tel: +32 (0)2 238 34 11 - Fax: +32 (0)2 230 59 12
<http://www.belspo.be>
<http://www.belspo.be/brain-be>

Contact person: Maaïke Vancauwenberghe
Tel: +32 (0)2 238 36 78

Neither the Belgian Science Policy Office nor any person acting on behalf of the Belgian Science Policy Office is responsible for the use which might be made of the following information. The authors are responsible for the content.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without indicating the reference :

Aaike De Wever (RBINS), Pieter Lemmers (KULeuven), Tanja Milotic (INBO) - Astrid Schmidt-Kloiber (BOKU) - Koen. ***Saving Freshwater Biodiversity Research Data***. Final Report. Brussels : Belgian Science Policy Office 2018 – 55 p. (BRAIN-be - (Belgian Research Action through Interdisciplinary Networks))

TABLE OF CONTENTS

ABSTRACT	5
CONTEXT	5
OBJECTIVES	5
RESULTS	5
CONCLUSIONS	5
KEYWORDS	5
1. INTRODUCTION	6
2. STATE OF THE ART AND OBJECTIVES	6
3. METHODOLOGY	9
4. SCIENTIFIC RESULTS AND RECOMMENDATIONS	13
NATURE OF THE KEY PROJECT RESULTS.	13
PUBLISHED (META-) DATA	14
<i>Background information on the key datasets.</i>	16
<i>Background information on additional datasets</i>	19
<i>Datasets for future publication</i>	22
TOOLS AND RECOMMENDATIONS FOR FUTURE DATA MANAGEMENT AND PUBLICATION	26
<i>Workflow & lessons learnt document (Task 5.2. Lessons learnt)</i>	26
<i>Data management guidelines, DMP tool and template (Task 5.3. Data management plans)</i>	26
<i>Other direct outputs</i>	26
<i>Knowledge exchange and training in data publication</i>	31
INTEGRATED ANALYSES	34
5. DISSEMINATION AND VALORISATION	40
6. PUBLICATIONS	41
7. ACKNOWLEDGEMENTS	45
ANNEX 1	47
1. MOLECULAR DATA	48
2. SEQUENCE DATABASES FOR DEPOSITION AND THE GENOMIC STANDARDS CONSORTIUM	49
3. DATA RESOURCES MANAGEMENT INITIATIVE	51
4. BELGIAN FRESHWATER MOLECULAR DATA DEPOSITION STRATEGY	53
REFERENCES	54

ABSTRACT

CONTEXT: Data produced by publicly financed research projects are often lost and seldom curated or published in open access repositories.

OBJECTIVES: (1) to facilitate exchange of expertise in data mobilisation, publishing, management, etc., (2) to promote data exchange and re-use of data building on existing standards and tools in order to create homogeneous data series, (3) to ensure the visibility and correct attribution of data contributors by publishing data online along with an accompanying (meta)data paper and/or synthetic scientific paper and (4) to develop a sustainable solution for future data management by implementing data policy and data management plans and by planning future activities in terms of data mobilisation, recovery and digitisation.

RESULTS: Key outcomes include: (1) publication of a wide range of (meta-) datasets on freshwater biodiversity stemming from research in Belgium (e.g. B-BLOOMS, BIOMAN, PONDSCAPE, MANSCAPE, etc.); (2) availability of tools and recommendations for future data management and publication and (3) manuscripts based on the integrated analysis of published databases as a showcase for the value of data mobilisation and integration.

CONCLUSIONS: SAFRED formulates recommendations regarding data management and data publication that can have a wide application for the management of future projects on aquatic biodiversity. Especially BELSPO, but also other funding agencies, can apply these best practices in data management. This will support the golden rule that reliable data resulting from publicly funded research projects should never become lost and should be made accessible in open access in standardised formats.

KEYWORDS:

FRESHWATER, BIODIVERSITY, DATA RECOVERY, DATA STANDARDISATION, DATA PUBLICATION

1. INTRODUCTION

Research projects typically have a short duration (2-5 years) and have very focussed research questions. Often, the researchers who manage the project and conduct the research are employed only for the duration of such projects and do not follow up on data management after the project ends, because they are then off to their next project or postdoc. In other cases, a researcher might be involved in long-term monitoring of aquatic biodiversity, but after his or her retirement, the database resulting from decades of monitoring remains on a private hard disk, is not publicly accessible and is eventually lost. There are many other possible scenarios that lead to the loss of valuable data that result from public funding.

Funding agencies have realised the value of such data, which might be useful for other purposes and research questions and for researchers outside of the research project or the monitoring programme that produced the data. Present day funding programs thus demand that applicants provide data management plans (DMP) in the project applications. However, there are few sources that provide recommendations to draft tailor-made DMPs for the different research fields, and the situation in Freshwater Biodiversity Research is not different. In addition, there are also large numbers of databases, resulting from older projects, that are at acute risk of being lost, unless they are ‘saved’ very soon. The SAFRED project has addressed the above problems, applying the most recent and most modern tools and standards.

2. STATE OF THE ART AND OBJECTIVES

Importance of freshwater biodiversity.

Freshwater environments are known to harbour a comparatively large fraction of the global biodiversity, while experiencing high rates in biodiversity decline (Dudgeon et al. 2006). At the same time, while information on freshwater biodiversity is relatively scarce compared with terrestrial biota (Heilpern 2015), freshwater species play a crucial role in ensuring a wide range of ecosystem services including water purification, nutrient cycling or food production. On the other hand, the proliferation of toxic cyanobacteria has impacted the quality and sustainable use of eutrophied waterbodies in terms of recreation, food production and water supply (Huisman et al. 2005; Hudnell and Steffensen 2008). Belgian researchers

have actively studied freshwater biodiversity over the last decades in the framework of a wide range of projects (including the BELSPO funded B-BLOOMS, MANSCAPE and PONDSCAPE projects). Although selected results from these projects have been the subject of multiple scientific papers, the underlying data have until now not been publicly released in a standardised and systematic manner.

The attention to the importance of freshwater biodiversity has not decreased in recent years, rather to the contrary. For example, GEOBON (the earths' biodiversity observation network) has recently (2017) endorsed the creation of a new thematic "BON", the freshwater biodiversity observation network (FWBON). FWBON currently has 108 members from 42 countries and is attracting considerable interest from researchers and practitioners involved in observing and evaluating freshwater biodiversity (<https://geobon.org/freshwater-biodiversity-observation-network-fwbon-endorsed-by-geobon-as-a-thematic-bon/>). Even more recently (august 2018), the Alliance for Freshwater Life (AFL) was launched. AFL is an interdisciplinary network of scientists, conservation professionals, educators, policy experts, creative professionals, and engaged citizens, working to improve the conservation and sustainable use of freshwater ecosystems and the biodiversity therein (<https://allianceforfreshwaterlife.org/>). Both new initiatives show that the relevance of Freshwater Biodiversity is considered higher than ever by the international research community.

Status of biodiversity (research) data sharing and publishing.

Data publication and sharing is a topic that is high on the international research agenda, e.g. through the activities of the Research Data Alliance (<https://rd-alliance.org/about.html>). In the biodiversity community, the Global Biodiversity Information Facility (GBIF) constructed an open data infrastructure and actively promotes the publication of data. Nevertheless, the online publication of biodiversity research data is not yet commonplace. This is particularly the case for freshwater biodiversity and motivated a consortium of 18 partners to construct a freshwater biodiversity data and information platform as part of the EU funded FP7 project BioFresh (Biodiversity of Freshwater Ecosystems: Status, Trends, Pressures, and Conservation Priorities).

Within BioFresh, RBINS was the main data partner in charge of the construction of a data portal and gained a rich experience in data mobilisation and processing. The present BRAIN project was meant to provide an impulse to the Freshwater Information Platform (FIP - <http://www.freshwaterplatform.eu/index.php/>) initiative, which emerged from the BioFresh project. In parallel, the Research Institute for Nature and Forest (INBO) has actively been publishing its datasets – including a large number of freshwater ones (e.g. [VIS – Fishes in inland waters in Flanders, Belgium](#), [TestWat](#), [Carabid beetles near the river Meuse in Flanders, Belgium](#)) – and has implemented open data policies, setting an example for similar research institutes (e.g. those connected in the ALTER-Net network of excellence).

Objectives.

The main SAFRED project objective was to achieve systematic recovery and publication of data generated in freshwater research by joining forces, standardising data, releasing data online and working out procedures to improve future data management.

To this end, the SAFRED network has:

1. facilitated the exchange of expertise, as well as support the assembly of datasets stemming from multi-partner projects and the documentation of information on data generation and storage (metadata).
2. built on existing standards, tools and expertise available within the network, in order to create homogeneous data series, promote data exchange and re-use of data.
3. ensured the visibility and correct attribution of data contributors by publishing data online along with an accompanying (meta-) data paper and/or synthetic scientific paper. While data processing mainly focussed on existing data, further attention has been given to exploring the possibilities for publishing (microbial) molecular data.
4. developed a sustainable solution for future data management by implementing data policy and management plans and planning future activities in terms of data mobilisation, recovery and digitisation.

3. METHODOLOGY

SAFRED has focused on mobilising existing data. We engaged in active data mobilisation among all project partners and multiple external data holders to acquire a large number of datasets to be standardised, integrated and published.

Data management and publication is not only a technological issue, but also a sociological one. Edwards et al. (2011) indicate for example that there is a need for information on data for reducing “data friction” among scientists. Similarly, there is a need for building a common understanding of what is envisaged by data sharing/publishing practices and which metadata are required for a layperson to understand the data well enough to be able to re-use them.

A first step in the dataset mobilisation was the construction of an **inventory of datasets** available within the SAFRED network, through collaborators and additional interested data holders. In first instance, this inventory was an informal shared document in which we documented existing data files, involved partners and processing status. Eventually we aimed to document all available datasets in a more structured way in the Freshwater Metadatabase, which uses the Environmental Metadata Language (EML) standard (see http://en.wikipedia.org/wiki/Ecological_Metadata_Language).

The original datasets, which were mostly available in MS Excel format were **mapped to the Darwin Core standard** (<http://www.tdwg.org/standards/450/>). Darwin Core is a standard governed by the Biodiversity Information Standards organisation (TDWG - *from the original name “Taxonomic Databases Working Group”*) and is loosely defined as a “bag of terms”, standard field names and definitions and a standard organisation into “core” and “extension” file(s). As it is widely used within the broader biodiversity community and also contains terms less relevant in freshwater sciences, RBINS compiled recommendations for the use of specific fields in freshwater sciences as “freshwater core” (<https://code.google.com/p/freshwatercore/>). For standardising sample-based data, SAFRED built on recent developments to improve the Darwin Core standard for this type of data and provide recommendations for further refinements to the standard led by GBIF in the framework of the FP7 EU BON project. Recommendations for using this new format were included in an update of the “freshwater core” which we named the “SAFRED recipe”.

For standardising the publication and archiving of complex **microbial molecular datasets**, we studied the most relevant and practical approaches from existing initiatives. These included the marine focused Megx.net Micro B3 information system (<http://mb3is.megx.net>), the Genomics Standards Consortium (GSC), the MIMARKS and MIxS checklists and standards (Yilmaz et al. 2011) developed through the GSC consortium and the experience with regards to its implementation in the Microbial Antarctic Resource System (MARS; <http://mars.biodiversity.aq>) project and the activities of the TDWG Genomic Biodiversity Working Group. Several approaches were considered, and our experiences were reported and discussed during the workshop on microbial ecology data management at the projects' final event.

The data publication or online sharing of the mobilised data was achieved through the **use of the Integrated Publishing Toolkit** (IPT; <http://www.gbif.org/ipi>). This tool was developed by GBIF and facilitates the publication of data in the Darwin Core standard. It exposes the data to the GBIF network and allows direct harvesting from the server. Upon publication, the data publisher can assign a Digital Object Identifier (DOI – see <https://www.crossref.org/>) to the dataset, which can facilitate citation and citation tracking of the dataset.

In parallel to publishing the data as such online, we also released descriptions of the datasets in the form of **(meta-) data and (meta-) data papers**. A 'data paper', 'metadata paper' or 'database paper' focuses on the description of a dataset. This can either be a pure description of the dataset for publication in a specialised data journal (such as the Freshwater Metadata Journal or the Biodiversity Data Journal) or a more extensive scientific article giving a broader insight in the database including analysis and discussion of its content, which might be targeted at a regular scientific journal.

Through the involvement of a subcontractor, the University of Natural Resources and Life Sciences Vienna (BOKU), the existing Freshwater Metadatabase was updated to feature a **multilingual search interface and allow multilingual data entry**, while ensuring the accessibility of the metadata for the international scientific community. The metadatabase questionnaire makes ample use of pre-defined keywords and checkboxes which were translated without interference of the data/information provider. Further, additional free text fields (title, abstract) for the new languages (Dutch, French, German) were provided. Along with the translation of the metadatabase elements, we explored the possibilities for

developing multilingual thesauri specific for freshwater metadata and datasets. To do so, we studied existing thesauri and ontologies such as Enviro

(<http://www.environmentontology.org/Browse-EnvO>) and PATO

(http://wiki.obofoundry.org/wiki/index.php/PATO:Main_Page) and engaged with representatives of LTER-Europe involved in the development of EnvThes.

To demonstrate the value of the data compilation activities undertaken in the SAFRED project and to present a convincing case for prospective data holders to encourage the publication of their data, we worked out a **showcase** consisting of two types of integrated analyses based on databases that have been processed as part of the SAFRED project. We focused on zooplankton communities as they play a pivotal role on the functioning and dynamics of lakes and ponds and are among the best sampled organism groups. The general objective of the first analysis was to identify how environmental and spatial variation determine functional characteristics and phylogenetic structure of zooplankton communities. The analysis of functional characteristics (e.g. body size, grazing capacity) allows us to obtain an idea of the robustness and resilience of freshwater systems and how it depends on environmental and spatial drivers. The analysis of the phylogenetic structure might be important as it can capture non-measured trait variation. More specifically, we (1) explored the extent to which patterns in taxonomic, functional and phylogenetic alpha and beta diversity are determined by the same set of explanatory variables (environmental variables and space), and (2) assessed whether including functional and phylogenetic information increases explanatory power on the effects of drivers for variation in community composition. In the second type of analysis, we investigate the association of taxonomic and functional diversity with functional redundancy in relation to eutrophication. For this purpose, we complemented the integrated SAFRED database, with a zooplankton trait database based on information on nine relevant functional traits that we extracted from literature. In addition, we created a phylogenetic tree based on all cladoceran taxa in our database based on four genetic markers (COI, 16S rRNA, 18S rRNA, 28S rRNA). Genetic information was obtained from GenBank, as well as from sequencing own samples from lab cultures at KULeuven. We present the key results from these analyses below. The resulting manuscript will soon be distributed among all project partners to receive additional input and feed-back. We anticipate that both studies can be submitted for publication in a scientific journal by October 2019.

To ensure that the SAFRED project has a lasting impact on the data management at the participating institutes/research groups, we envisaged two main activities with regards to sustainability. The project featured a **“Sustainability and freshwater data management policy”** work package (WP5). As part of this, INBO took the lead in sharing its experience with implementing a research data policy and data management plans and actively supported the development of such policies and plans by the project partners. This involved a review of the current data lifecycle from the creation to the re-use of data in the partner institutes, in order to develop a research data policy and templates for data management plans that supports the functioning of the research groups. Current data management practices at the partners’ institutes were reviewed by conducting two different surveys: one at the researcher level in order to study individual based data management initiatives, and one at the institute or research group level in order to survey commonly used data management practices and the content of a research data policy if present. Based on the outcome of these surveys, a document was written containing guidelines for tackling commonly encountered data management bottlenecks and practical tips for writing data management plans. In addition to these written guidelines, lectures were offered to the partners’ institutes. In these lectures which were attended by a large group of (PhD) students, post docs and researchers, the different aspects of research data management were covered. Essential components of the data management cycle include documenting procedures for sampling and processing, the use of data and metadata standards, quality control steps, short and long-term storage of data and the planning of (public) archiving and online publication (e.g. Jones 2011). In addition to our efforts to work out data management plans, we ran a task for building an **inventory of relevant (legacy) freshwater datasets**, which project partners want to mobilise or digitise in future activities and explored funding sources to do so.

As part of the development of a data processing workflow, we also explored possibilities for creating a reproducible workflow with the “literate programming” approach by using the functionalities available in scripting languages such as R and Python (among others). This approach provided the ability to combine documentation about the transformation process and code that executes the transformations and allowed to trace and re-run the exact data manipulation that were executed. We acknowledged this approach as being very valuable, but it would have required more time for training project partners to implement this.

During the project, we organised several **outreach workshops** targeting a wide range of audiences. From the start of the project, we wanted to inform potential users and data providers of the project and invited them to actively contribute to the project. Towards the end of the project, we organised a **final conference for a broad audience**, including policy makers, water managers and scientists from the freshwater and biodiversity research domain and the interested public. At this event, we presented – among others – lessons learned, test case results and motivated participants to publish their data.

4. SCIENTIFIC RESULTS AND RECOMMENDATIONS

Nature of the key project results.

The key results of this project in the BRAIN Axis 6: “Management of collections” and the priority topic “New digital collections and data” cover 3 main categories: (1) published (meta-) data, (2) tools and recommendations for future data management and publication and (3) two scientific manuscripts based on the SAFRED database as a showcase for the value of data mobilisation and integration.

In terms of data mobilisation and publication, the SAFRED network represented a substantial volume of high-quality data on freshwater biodiversity research from Belgian scientists which was made publicly available during the project. Data were released among others through the Freshwater Biodiversity Data Portal managed by RBINS (accessible at data.freshwaterbiodiversity.eu and part of the Freshwater Information Platform) and exchanged with the GBIF (Global Biodiversity Information Facility) network. Owing to the high availability of freshwater data realised through SAFRED, this project also represented an example for a thematic data mobilisation initiative that could be deployed elsewhere and could be a test case to demonstrate what is possible if other regions would reach a similar level of coverage.

The tools and recommendations for future data management and publication in (freshwater) biodiversity research range from project outputs such as the workflow and lessons learnt document, data management guidelines and templates for data management plans to indirect outputs such as the increased exchange of expertise among the project partners and with external parties through a wide range of meetings and workshops.

Finally, the “showcases” consisting of two manuscripts based on integrated analysis of the compiled data clearly demonstrate the scientific value of these kinds of data compilations. Below, we discuss each of the 3 types of key results in further detail.

Published (meta-) data

As a first step in the data mobilisation work, we compiled an overview of available datasets and documented all relevant datasets in a central metadatabase, regardless of the fact whether these data could be published immediately, are under embargo or will not be made publicly available at all. As a central metadatabase we used the Freshwater Metadatabase at freshwatermetadata.eu constructed in the framework of the EU FP7 project BioFresh and currently integrated in the Freshwater Information Platform (www.freshwaterplatform.eu). During the project, we entered information on 15 datasets, 13 of which are now publicly available (Tables I and II) and metadata are also available in the Freshwater Metadata Journal.

The publication of data itself is the result of a wide range of project tasks starting with obtaining permission from the different data contributors (Task 5.1.1) and locating data files for joint publication (e.g. project data from different organism groups stored in separate files) (Task 4.1.1) over transfer of data to a standard format (Task 4.1.2) and quality control (Task 4.2) to their online release (Task 4.3) and description in a (meta) data paper (Tasks 1.3 and 4.4). Table I reports on the processing status of the datasets considered as “key” datasets during the proposal writing, including microbial ecology datasets (Task 3.2). Further details on these datasets are provided below. The processing status for additional datasets that were treated during SAFRED is further described in the text. Both additional datasets that are not fully processed, as well as datasets that were identified but not processed, are included in Table II which suggests further processing steps.

Table I: Overview of the publishing status for metadata, data and (meta)data papers for the key datasets of the SAFRED project. Completely finished tasks are indicated with an “X” or the resource URL, where data or information are available.

<i>Dataset name</i>	<i>Meta data</i>	<i>Data</i>	<i>(Meta)data paper</i>
BIOMAN Belgium & Netherlands	X	data.freshwaterbiodiversity.eu/ipt/resource?r=bioman_belgium	doi:10.15504/fmj.2018.29
Manscape, Belgium	X	data.freshwaterbiodiversity.eu/ipt/resource?r=manscape	doi:10.15504/fmj.2017.26
Midden-Limburg, Belgium	X	data.freshwaterbiodiversity.eu/ipt/resource?r=midden-limburg	doi:10.15504/fmj.2017.27
Pondscape, Belgium & Luxembourg	X	http://data.freshwaterbiodiversity.eu/ipt/archive.do?r=pondscape	doi:10.15504/fmj.2018.31
Tommelen, Belgium	X	http://data.freshwaterbiodiversity.eu/ipt/resource?r=tommelen	doi:10.15504/fmj.2018.40
De Maten (Genk, Belgium)	X	http://data.freshwaterbiodiversity.eu/ipt/resource?r=de_maten_belgium	doi:10.15504/fmj.2018.30
B-BLOOMS2	X	data.freshwaterbiodiversity.eu/ipt/resource?r=sf13-bblooms13	doi:10.15504/fmj.2017.28
MIDI-CHIP	X	http://hdl.handle.net/2268/228149	doi:10.15504/fmj.2018.37

Belgian River Meuse: fish data	X	data.freshwaterbiodiversity.eu/ipt/resource?r=sf6-meuse_fish	doi:10.15504/fmj.2018.33
Belgian River Meuse : environmental data	X	data.freshwaterbiodiversity.eu/ipt/resource?r=sf3-meuse_physicochemistry	doi:10.15504/fmj.2018.32
Belgian River Meuse: macroinverte- brate data	X	Under embargo	doi:10.15504/fmj.2018.34
Phytoplankton in rivers in Flanders Belgium	X	ipt.biodiversity.be/resource?r=flemish_rivers_phytoplankton	doi:10.15504/fmj.2018.39
ECOPOT1	X	ipt.biodiversity.be/resource?r=ecopot_phytoplankton	doi:10.15504/fmj.2018.38
ECOPOT2	X	ipt.biodiversity.be/resource?r=flemish_waterbodies_phytoplankton_ecopot_2	doi:10.15504/fmj.2018.38
Southern hemisphere diatoms	X	Under embargo	Under embargo

Background information on the key datasets.

The **lake and pond datasets** comprise data on the taxonomic diversity of multiple aquatic organism groups (including bacterioplankton, phytoplankton, phytobenthos, macrophytes, diatoms, zooplankton, zoobenthos, aquatic macro-invertebrates, fish and amphibians) and major local environmental characteristics.

The **BIOMAN** dataset comprises local environmental data and community data of different organism groups (bacterioplankton, zooplankton, ciliates, phytoplankton, macro-invertebrates, fish, protists and aquatic vegetation) from 98 shallow lakes covering three geographic regions in Europe sampled in 2000-2001. The database BIOMAN-Belgium is a subset of the overall BIOMAN dataset and includes data from 39 shallow lakes located in Belgium and The Netherlands. For reasons of data availability and quality, we were only able to process and publish the Benelux data during SAFRED.

The **MANSCAPE** dataset (Belspo: **Integrated management tools for water bodies in agricultural landscapes**) comprises species occurrence data of seven different organism groups (phytoplankton, diatoms, zooplankton, macro-invertebrates, macrophytes, amphibians and fish) and data on physical, chemical and morphometric variables of 126 small farmland ponds distributed over almost the entire Belgian territory. Ponds are located in different land-use intensity and were sampled in 2003.

The **PONDSCAPE** dataset (Belspo: **Towards a sustainable management of pond diversity at the landscape level**; <http://www.pondscape.be/>) comprises taxon occurrence data of eight different organism groups (bacteria, phytoplankton, diatoms, cladocerans, macro-invertebrates (molluscs, heteropterans and coleopterans), macrophytes, amphibians and fish) and data on physical, chemical and morphometric variables of 125 farmland ponds covering five biogeographic regions in Belgium/Luxembourg.

The **Tommelen** dataset contains occurrence data of macrophytes and macro-invertebrates, as well as information on major local environmental variables from bomb crater pools (n=23) at Tommelen nature reserve (122 pools in an area of 12 ha, Limburg, Belgium), sampled multiple times between 2007 until 2012.

The **De Maten** and **Vijvergebied Midden-Limburg** datasets consist of data from shallow interconnected fish ponds in Belgium. The De Maten pond dataset contains data on local pond conditions and taxonomic community composition of phytoplankton, zooplankton, macro-invertebrates and fish from 34 interconnected fish ponds in the “De Maten” nature reserve (Limburg, Belgium). The Midden-Limburg pond dataset contains data on local habitat conditions and taxonomic community composition of multiple aquatic organism groups (phytoplankton, zooplankton, aquatic vegetation, macro-invertebrates and fish) from 38 interconnected fish ponds in the fish pond complex Midden-Limburg (Limburg, Belgium). The selection of fish ponds represents five different pond management types.

The **B-BLOOMS2** (www.bblooms.be) dataset contains monitoring data on five Belgian reference lakes, which were regularly sampled throughout the study period (mostly in 2007-2008) to follow-up the evolution of cyanobacterial blooms. Sampled variables include information on the weather, nutrients, toxin concentrations, phytoplankton and zooplankton densities. In addition, molecular data from DGGE, clone library analysis and information on the detection of *mcy* genes (coding for microcystin synthetase) are available from this project and are released as a sequence set in addition to the occurrence dataset.

The **MIDI-CHIP** (<http://www.cip.ulg.ac.be/midichip/>) database contains data on the microbial biodiversity in 205 lakes sampled in 2001 in Belgium, Czech Republic, Finland, France, Luxembourg, Italy and Poland. The available data covers 563 strains, 69 different taxa, 2069 sequences from strains and environmental samples (obtained by DGGE and clone libraries), and physical and chemical data.

The **Meuse datasets** include physical, chemical and biological data of the Belgian Meuse river compiled from different sources (UNamur, SETHY, ISSeP and DEMNA) covering 1973 to 2012. The dataset was further extended during the project. Currently, 8 physical/chemical variables were monitored at 3 sites approx. 8-12 times a year between 1973 and 2012. The macroinvertebrate data cover 2 sites sampled each year between 1998 and 2011 with an IBGA method. The fish dataset consists of electrofishing data and fish-pass data from 1989 to 2012. Macroinvertebrate data were published by DEMNA, fish and physical/chemical data were processed during SAFRED.

The **Phytoplankton in rivers in Flanders Belgium** dataset contains phytoplankton monitoring data from Flemish rivers and channels.

The **ECOPOT** datasets contain monitoring data from the ECOPOT projects (Determining the maximal and good ecological potential, as well as the current state of Flemish regional water bodies). The data contains phytoplankton counts of Flemish lakes and pools. Publication was done in two sets of data because of significant differences in taxonomic granularity at which the identification of organisms was executed. The ECOPOT_1 dataset contains phytoplankton data identified at (mostly) genus level for 2 sampling locations monthly monitored during the growing seasons (May-October) of 2006 and 2007. ECOPOT_2 dataset contains phytoplankton data identified predominantly at a (morpho-) species level and covers 8 sampling locations monitored for 5 months of a single growing season (May-September) in 2008 (Blokkeerdijk, Schulsenmeer and Vinne) to 2014 (Hazewinkel). The different

ECOPOT datasets, and the dataset on **phytoplankton in Flemish rivers** were separate archived. Metadata on all three Darwin Core archives has been committed to the Freshwater Biodiversity Data Portal (not yet made public), and the files are ready for upload on the IPT after the finalization of a final quality control together with INBO.

The **Southern hemisphere diatoms** dataset contains relative abundances on oxidised diatom valve counts (200 valves per sample) from 604 individual samples taken from 439 freshwater lakes situated in fifteen ice-free regions and islands distributed over the three main biogeographical regions in the AR between 45°S and 82°S, namely Continental Antarctica, Maritime Antarctica and the main sub-Antarctic Islands. These samples were collected over the course of different international expeditions from 1992 up to 2013. The database was compiled by combining existing and published datasets with newly obtained inventories of species presence-absence data in a number of lake districts, including Marion Island and James Ross Island. For all lakes, air surface temperature data, as well as measurements of the specific conductance and pH of the lake water are available. For a subset of 213 lakes the concentrations of Na^+ , K^+ , Mg^{2+} , Ca^{2+} , Cl^- , NO_3^- , NH_4^+ and PO_4^{3-} are also available. The metadata have been submitted to the Freshwater Biodiversity Data Portal and are being revised after the quality check by BOKU. They will be made public afterwards. The data files are ready for upload on the IPT and will be made public after the resulting paper is published.

Background information on additional datasets

A wide range of additional datasets were identified and processed during SAFRED. These datasets vary widely in size, complexity and processing status. More details are here provided on some of the key additional datasets/projects considered.

The "**Inland water macro-invertebrate occurrences in Flanders, Belgium**" dataset is the major external dataset that was published during SAFRED (cfr. Task 6.3. Mobilising external datasets). This sampling event dataset holds data on macroinvertebrate occurrences sampled in the water quality monitoring network from the Flemish Environment Agency (VMM). The dataset contains more than 280.000 occurrences at over 4.100 monitoring sites across Flanders (Belgium). The data is published under a CC-BY license at ipt.inbo.be/resource?r=vmm-macroinvertebrates-events. A data paper, Vannevel et al. (2018), describing this dataset has been published in the international journal Zookeys.

Another important external dataset is the **Macro-invertebrate collection data from the Province of Antwerp - Provinciaal Instituut voor Hygiëne (PIH)**. Processing these data proved rather tedious as we aimed to integrate them into the RBINS collection database DaRWIn (Data Research Warehouse Information Network) before publishing to the GBIF network, because a large part of the physical collections linked to this database are curated at RBINS. At this stage, the original data has been re-organised into import templates for integration in DaRWIn and will be further followed up by RBINS.

We also envisaged to publish selected freshwater datasets from the **SPEEDY** project (Belspo IAP project P7/04 SPEEDY : SPatial and environmental determinants of Eco-Evolutionary DYnamics: anthropogenic environments as a model). SAFRED partners from RBINS, INBO and the BBPf have been approached by the SPEEDY consortium to provide advice on the projects' data archiving and participated in several meetings to discuss the approach. At this stage the SPEEDY consortium has assembled most raw data and will pick up the data processing in autumn 2018. To facilitate the process, we started preparing the data publication for selected freshwater datasets (e.g. zooplankton data from 81 pools in Belgium along a gradient of urbanization) and will follow-up on the data publication during the presently running BRAIN project ORCA (A comparative analysis of ORganic and Conventional Agriculture's impact on aquatic biodiversity).

The **Meuse data**, "Macro-invertebrate data of the Belgian River Meuse from 1989 to 2012", and the **FAME dataset**, were also envisaged for publication through SAFRED. To do so, UNamur initially established contact with DEMNA in July 2016. At this point, UNamur received the authorisation to publish the "Meuse" data belonging to DEMNA under the "CC0" license waiver. During a follow-up meeting on 30/05/2017, the approach for publishing the DEMNA databases on the GBIF platform was discussed with representatives of the Belgian Biodiversity Platform (BBPf). Unfortunately the plans for releasing the DEMNA macro-invertebrate data were not clearly communicated, which meant that the publication of selected data used by UNamur was prepared in parallel, but could eventually not be released. Metadata were made available by DEMNA at <https://www.gbif.org/dataset/76a03a1c-fe0b-496e-adad-cd19af2ae043>.

The metadata and data availability from the **Boyekole Ebale Congo 2010** expedition were reviewed by RBINS with the aim to increase the available data to justify the public release of the Congo River portal developed by the Belgian Biodiversity Platform. Metadata were

completed and requests for data updates were sent out in collaboration with Erik Verheyen (RBINS) the expedition leader. At this stage, the datasets on vascular plants and myxomycetes are in an advanced stage of processing but are still pending, while other datasets are often still incomplete.

Part of the **SEXASEX** datasets (EU FP7: From Sex to Asex: a case study on the interactions between sexual and asexual reproduction) has been published as Supplementary Electronic Data in Schmit et al. (2013): <https://doi.org/10.1111/jbi.12174>.

The **Louette dataset** (zooplankton colonization dynamics in 25 newly created pools in Belgium; see also Louette & De Meester 2005, Louette et al. 2008), the **zooplankton species distribution database** (taxonomic data cladocerans in Belgian lakes and ponds (De Meester et al. 2002; Forró et al. 2003) and a database on **reservoirs in Northern Ethiopia** (taxonomic data on zooplankton, fish and phytoplankton from 30 newly created reservoirs in the highlands of Tigray, Dejenie et al. 2008; Dejenie et al. 2012; Teferi et al. 2014) were identified to be potentially further processed and made publicly available through GBIF in the future.

As mentioned under the “key datasets”, the **TIPPINGPOND** dataset represents a re-sampling of a subset of 61 MANScape ponds in 2013. The data from this project is currently still being used for drafting scientific manuscripts and the data contributors/owners therefore prefer to make the data publicly available at a later stage (likely in 2019).

As the molecular data of **B-BLOOMS(1)** is distributed over multiple files with complex annotations and would require a good deal of “data archaeology”, we decided to (i) upload the located data files of B-BLOOMS(1) as such on ORBI for archival purposes, and (ii) prepare a metadata paper to document the existence and background of these data.

INBO is gradually publishing its biodiversity data as Open Data on GBIF via its own IPT installation (ipt.inbo.be). Among these datasets are a lot of freshwater datasets. **INVEXO** (“Invasive species - American bullfrog (*Lithobates catesbeianus*) in Flanders, Belgium”) was mentioned in the project proposal and has now been published. Metadata for this dataset have been imported into the Freshwater Metadatabase using a tool for importing Ecological Metadata Language (EML) files developed by the University of Natural Resources and Life Sciences (BOKU Vienna). Other potential INBO datasets that were mentioned in the project proposal are **Macrof_I_REFCOND** and **Macrof_II, Monitoring data** including FytobII:

Fytobenthosgegevens Vlaamse Waterlichamen and “Historische collecties van diatomeeën in Vlaanderen”.

UGent has been registered as a GBIF data publisher through the Belgian Biodiversity Platform. In addition to the datasets on **phytoplankton in rivers** in Flanders Belgium, **ECOPOT** and **Southern hemisphere diatoms**, they published the **Global Lacustrine Diatoms dataset** (<http://doi.org/10.15468/q457rc>) on Diatom distribution data at morphospecies level for the Southern Hemisphere (40-72°S) in the course of SAFRED. The monitoring data from the **VLINA/Kraenepoel** projects remain to be retrieved from a former collaborator.

Datasets for future publication

Here, we briefly summarize the datasets mentioned above that require further follow-up and additional datasets that were identified during the project (cfr. Task 5.5). An overview of potential future freshwater data mobilisation activities is given in Table II.

Table II: Overview of datasets which are either being processed or could be considered for future publication and proposed actions for expediting the data publication.

dataset name	short description	processing status and actions required	contact(s)
BIOMAN Spain	See BIOMAN	To be mobilised from Spanish partner	KULeuven
BIOMAN Denmark	See BIOMAN	To be mobilised from Danish Partner	KULeuven
PIH data	Macro-invertebrate collection data from pools and ponds in Flanders, spanning several decades	Data in DaRWIn import template, Integration in DaRWIn to be completed	RBINS

SPEEDY-Ostracods	Ostracod identifications from 81 SPEEDY pools	In Excel sheet, to be formatted	RBINS
SPEEDY-Zooplankton	Zooplankton identifications from 81 SPEEDY pools	In Excel sheet, to be formatted	KULeuven
Congo River-Vascular Plants	Occurrence data of vascular plants from the Boyekole Ebale Congo 2010 expedition	In Excel sheet, to be formatted	RBINS
Congo River-Myxomycetes	Occurrence data of myxomycetes from the Boyekole Ebale Congo 2010 expedition	In Excel sheet, to be formatted	RBINS
COBAFISH	Biodiversity and environmental data from parts of the Congo River	Data not yet compiled or processed	RBINS
SEXASEX	Ostracod/ macro-invertebrates/ physical/chemical data from 147 SEXASEX ponds throughout Europe.	Part of the raw data are published as “supplementary electronic material” attached to Schmit et al. (2013): https://doi.org/10.1111/jbi.12174	RBINS

Louette	zooplankton colonization dynamics in 25 newly created pools in Belgium; see also Louette and De Meester 2005, Louette et al. 2008)	In Excel sheet, to be formatted	KULeuven
TIPPINGPOND	Re-sampling of subset MANScape ponds	In Excel sheet, to be formatted	KULeuven
Reservoirs in Northern Ethiopia	Zooplankton data from Ethiopian reservoirs	In Excel sheet, to be formatted	KULeuven
Zooplankton species in Belgian waterbodies	Zooplankton species in Belgian waterbodies	In Excel sheet, to be formatted	KULeuven
B-BLOOMS(1)	Molecular data on cyanobacteria in Belgian lakes	Uploaded as a zip-file in ORBI (orbi.uliege.be)	ULg
MICROMAT	Environmental data from the project “microbial mats in Antarctica” Additional molecular sequence data	Metadata for a dataset with the environmental data produced during this project were already published at the Australian Antarctic Data Centre (https://data.aad.gov.au/metadata/records/ASAC_2112)	ULg

BCCM-Cyanobacteria	Belgian Co-ordinated Collections of Micro-Organisms ULC Cyanobacteria Collection catalogue	In Excel sheet, to be formatted	ULg
VLINA	Monitoring data	Data to be retrieved from former collaborator.	Ugent
Kraenepoel	Monitoring data	Data to be retrieved from former collaborator.	Ugent
Van Meel reports	Numerous physical and chemical data on Lake Tanganyika, saline lakes in north-western Belgium, etc.	Presently data are in printed tables; data need to be scanned and digitised.	RBINS
Belgian Land and Freshwater Mollusca (BLF)	Mollusca collection data registered from 1974 onward (also BLZ).	Data in Excel and Access files, sampling stations encoded in DaRWIN, specimen details to be imported Data to be transferred to DaRWIN import templates	RBINS
Selected “waarnemingen.be” data	Citizen science data from the “waarnemingen.be” platform	Selected data were published as thematic datasets (e.g. Odonata)	Natuurpunt

Tools and recommendations for future data management and publication

The direct outputs under this theme were realised as part of Work package 5. “*Sustainability and freshwater data management policy*” and included the “workflow and lessons learnt” document, “data management guidelines and templates” for Data Management Plans. In addition, the project had considerable ‘indirect’ impact through the increased exchange of expertise among the project partners and with external parties during a wide range of meetings and workshops.

Workflow & lessons learnt document (Task 5.2. Lessons learnt)

We documented the data processing and quality control in a workflow description, together with the lessons learnt during this project in a report that will be published as a paper in an international open access journal (De Wever et al., in prep.). With this document, we want to inspire and assist prospective data holders to publish their data and provide suggestions to funders and policy makers how they could contribute to the more systematic and rigorous publication of data from publicly funded projects.

Data management guidelines, DMP tool and template (Task 5.3. Data management plans)

Complementary to the workflow and lessons learnt document, we compiled a document with general data management guidelines to improve the day-to-day management of data, based on the outcomes of the surveys run at the partner institutes. As part of these guidelines, we recommend initiating data management plans (DMP) for selected projects and activities and consider data publication practices at the level of a wider data policy for the partner research laboratories and institutes. We also propose DMP tools and a generic template for (fresh-water) biodiversity data. The guidelines document is published online at the Zenodo repository (<https://zenodo.org/record/1421739#.W64xoHtMGpo>; doi: 10.5281/zenodo.1421739) and later also as a webpage.

Other direct outputs

Other tools that contribute to future data management and publication include (1) recommendations on microbial ecology publication, (2) the development of the improved metadatabase with multi-lingual capabilities, (3) the inventory of future freshwater data mobilisation activities and (4) support for IPT set-up and use.

ULg performed an extensive literature analysis (Task 3.1. Horizon scanning of developments in standardisation of data in molecular studies on environmental samples) in collaboration with UGent and consulted with the Microbial Antarctic Resource System team (mARS; <http://mars.biodiversity.aq/>). This work is reported in Annex 1. Based on this analysis, we proposed, depending on the project, to store sequence templates and MiMarks/MIMS/MIGS/MIxS in open source data repositories, such as GenBank or GBIF, which are linked via the mARS portal. These records will contain the metadata information and link to both occurrences and molecular datasets with standard compliant contextual data.

For the deposition of molecular data generated from High Throughput Sequencing, minimum information is now required for the deposition of data in the Sequence Read Archive (SRA, NCBI) database, which was specifically developed for High Throughput Sequencing datasets. After the publication, it generates Minimum Information about any Sequence (MIxS) files that are referenced with a BioSample ID. We recommend to compress MIxS files with a mARS sequence set template (available at: <http://mars.biodiversity.aq/>; Figure 1) as a .zip/.gz/.tar folder, which will then be uploaded on an open repository (e.g. SRA). The mARS sequence set template gathers specific molecular metadata (e.g. primers, sequencing technology, ...) and locations of molecular data sets (accession number, database, or web address). Then metadata can be filled, mentioning the web address of the compressed file. In the case of Antarctic data (e.g. MICROBIAN, CCAMBIO, etc.) we recommend to directly use the mARS platform for the GBIF entry through the mARS IPT (http://mars.biodiversity.aq/site_pages/datasets).

Regarding molecular data generated with Sanger technology, we also recommend using the approach described above when data are published on a public database (e.g. GenBank). Unpublished sequence related data, which was not included in scientific papers (e.g. from MIDI-CHIP, B-BLOOMS2 and MICROMAT), can also be packaged together with a sequence set and uploaded on an open repository as was done for B-BLOOMS2.

Occurrence data associated with microbial ecology data can be prepared for publication to the GBIF network. The location where the molecular data is stored must be reported into a mARS sequence template. In order to realise the link between the two files, we implemented a sequence list with two columns with the sequence names and the eventID. During the course of this project we successfully tested this methodology with the B-BLOOMS2 datasets

Figure 1: Example of a completed mARS sequence set template. For each set of primers used (a “sequence set”), details on the methodology, sequence repositories and links and/or accession numbers are provided in this file.


Référence : Molecular dataset from BELSPO project B-BLOOMS 2

Type de document : Rapports : Autre
Discipline(s) : Sciences du vivant : Sciences de l'environnement & écologie
Pour citer cette référence : <http://hdl.handle.net/2268/213145>

Titre : Molecular dataset from BELSPO project B-BLOOMS 2
Langue du document : anglais
Auteur, co-auteur : [Lara, Yannick](#) [Université de Liège > Département des sciences de la vie > Centre d'ingénierie des protéines >]
[Wilmotte, Annick](#) [Université de Liège > Département des sciences de la vie > Physiologie et génétique bactériennes >]
Date de publication : Non daté
Mots-clés : [en] cyanobacteria ; molecular diversity ; freshwater
Résumé : [en] This entry is dedicated to answer the need of a link between molecular data, environmental parameters, and counting depositions in the field of microbial ecology. Indeed, it is necessary to implement a new approach to link data that are deposited in GBIF with molecular data set that are deposited on GenBank or that stay in lab computers. In the frame of the BRAIN BELSPO project SAFRED, we attempt to achieve this goal by supplying a sequence template set as designed by the Microbial Antarctic Resource System (mARS). For the B-BLOOMS 2 molecular data set, we also supply a fasta file which contains curated sequences obtained by the DGGE method. These sequences are not published elsewhere. In addition, we added a sequence list which contains Event ID information for each sequences.
Organisme(s) subsédant(s) : BELSPO
Intitulé du projet de recherche : SAFRED
Public cible : Chercheurs ; Professionnels du domaine ; Grand public
URL permanente : <http://hdl.handle.net/2268/213145>


Document(s) associé(s) à cette référence :

Document(s) en texte intégral :

	Fichier	Commentaire	Version	Taille	Accès
 Accès ouvert	MolecularDataSet_BB2.zip	this zip file contains 3 files	Postprint éditeur	4.13 kB	Voir/Ouvrir

Accord(s) & licence(s)

Licence de diffusion :

	Fichier	Commentaire	Taille	Accès
 Accès privé	license.txt	Autorisation accordée par les auteurs pour la diffusion en accès ouvert	20.38 kB	Voir/Ouvrir


Statistiques


Figure 2: Deposition of the B-BLOOMS2 molecular dataset on ORBI.

Since January 2018, the Freshwater Metadatabase features a 4-language interface (English, Dutch, French and German). The language selection is visible after logging in the metadatabase by clicking on the respective flag symbol (Figure 3). Data can then be filled in any of these four languages. To guarantee maximum comprehensibility, title and abstract always need to be entered in English, supplementary to the main language. The four languages can also be selected for the full-text-search as well as the query tool and database entries can be viewed in all languages.

The figure displays four screenshots of the Freshwater Metadatabase interface, arranged in a 2x2 grid. Each screenshot shows a different language version of the same form, used for entering metadata for the AQEM/STAR invertebrate database. The languages are Dutch (top-left), English (top-right), German (bottom-left), and French (bottom-right). Each form is divided into sections: 'Algemene informatie' (General information), 'Technische en administratieve specificaties' (Technical and administrative specifications), and 'Publicatie' (Publication). The forms include various input fields for data format, access level, operating system, language, and contact information, along with checkboxes for publication status and update frequency. The interface is designed to be user-friendly and accessible to researchers from different countries.

Figure 3: Screenshots of the Freshwater Metadatabase interface in four different languages.

Towards the end of the project, we compiled an inventory of future freshwater data mobilisation activities (Task 5.5). The result of this exercise is included in Table II.

Within the SAFRED project we offered support for the Integrated Publishing Toolkit (IPT) set-up (Task 5.4). Since the SAFRED proposal was accepted, GBIF has implemented the option to manage multiple organisations through a single IPT installation. Due to this change it is often unnecessary to install the tool on partners' servers, unless they have an explicit demand for this. At this stage, datasets that are connected to multiple partners are published on the Freshwater Data Portal/FIP installation (through the platform "BioFresh"; data.freshwaterbiodiversity.eu/ipt), external parties such as VMM and Natuurpunt received an organisational account on the INBO installation (ipt.inbo.be), and UGent, Spw-DEMNA and RBINS are able to publish through the Belgian Biodiversity Platform (ipt.biodiversity.be).

Knowledge exchange and training in data publication

Knowledge exchange and training was an essential component of the SAFRED project. This happened through ad hoc informal as well as planned meetings including project meetings (Task 6.2.2) – on which we covered among others licenses and rights waivers for publishing data (Task 5.1.2) – information sessions for external data holders and project partners (Task 6.2), the workshop on microbial ecology data management (Task 3.3), the final workshop on data management in freshwater sciences (Task 6.4) and the consultations with the Follow-up Committee and its members.

The first partner meeting was held on February 25th 2016, and was followed by a series of hands-on data meetings with selected partners during the months May to July 2016. A second round was started in December 2016 and January 2017. Project work meetings were held in October 2016 and February 2017 to share experiences in data processing, metadata entry and plan further work.

As most scientific partners were not very familiar with the Darwin Core (DwC) exchange standard prior to the SAFRED project, we organised an introduction at the kick-off meeting. This introduction was given by the BBPf and INBO. Several hands-on data meetings with selected data holders were organised by RBINS and supported by BBPf and INBO to work on specific datasets. At these occasions we also initiated discussions on the use of the “event core” organisation for sample-based data, minimally required fields and the inclusion of supporting environmental data in the export for data publication.

The SAFRED project and the importance to make (biodiversity) monitoring data systematically available was discussed during the presentation entitled “*Beheer van data rond zoetwaterbiodiversiteit: naar een systematische aanpak en het online beschikbaar stellen van gegevens*” by Aaike De Wever at the Water Forum in Ghent on the 14th of October 2016. This event reached a large audience of Dutch-speaking freshwater managers and researchers. A similar presentation entitled “*Sauver et préserver les données sur la biodiversité des eaux douces: gestion et publication des données, un exemple concret sur les inventaires biologiques en Meuse*” was delivered at the Aquapôle meeting on 4th of May 2017 with presentations by Aaike De Wever and Adrien Latli.

To share our experience from Tasks 3.1 and 3.2 and exchange views with experts and scientists working with microbial ecology data, we originally planned a workshop around this

topic during the Scientific Committee on Antarctic Research (SCAR) biology symposium in Leuven (10-14 July 2017). However, to broaden the scope of the workshop and not only focus on sequence data from Antarctica, this workshop was moved to the final event on February 28th, 2018.

UGent prepared an interactive presentation in collaboration with ULiège. This presentation started with a brief overview of the molecular sequencing approaches used during the past decades and their specific data formats. The most widely used platform to date for Next Generation Sequencing (Illumina) was discussed in more detail. Next, an overview was given on the sampling preparation, the output file, genes generally sequenced and the pitfalls related to these methods. Good practices when publishing these data and the need to provide (standardised) metadata were subsequently discussed. A step-by-step tutorial was given regarding how to submit amplicon sequencing and metagenome data to NCBI's Sequence Read Archive (SRA). This included the selection of the appropriate minimum information environmental data package, adding all necessary metadata (cf. mARS), and the subsequent publication of the raw sequencing data itself.

In a second tutorial, the different possibilities and procedures for downloading sequence data from data repositories were summarised, to demonstrate the benefits of open access data, using the SRA toolkit to download the data. In a third demonstration, the different bioinformatics pipelines to process sequencing data were presented. Some basic Linux commands and the different steps in a bioinformatics pipeline were presented. These included quality control of the sequencing, paired-end merging, quality filtering, dereplication, and chimera detection. The final part of this tutorial was dedicated to different algorithms used for clustering the sequences into Operational Taxonomic Units (OTUs) and how to assign these to known sequences available in reference databases (e.g. Greengenes, PR2). The aim of this last tutorial was to show how the downloaded sequences could be processed into a usable dataset.

A first successful follow-up committee meeting was held on the 9th September 2016 at INBO, Brussels. During this meeting, we focused on introducing the project and planned work and getting advice from external experts early in the project. We had a fruitful discussion on building an inventory of (Belgian) freshwater biodiversity datasets and mobilising external data. Main recommendations included: documenting important (non-digital) works in the metadatabase; follow-up developments in terms of citizen science data

publications; engage with provincial citizen scientist bodies “Koepels”; evaluate options for “low threshold” data repository to which researchers can upload any data.

Following the meeting we initiated discussions with representatives of the Flanders Environment Agency (VMM) and the Province of Antwerp - Provinciaal Instituut voor Hygiëne (PIH) to explore possibilities for the publication of biological data associated with monitoring campaigns and linked to collections stored at RBINS (Task 6.3). These discussions were followed up by Dimitri Brosens (BBPf) and Aaike De Wever (RBINS) and led to the publication of the macroinvertebrate data and import of collection data in the RBINS DaRWIN database.

The second follow-up committee meeting was held on 27th February 2018 at the occasion of the projects’ final event. Although most of the follow-up committee members (8) indicated that they were available and registered for this event, only 4 persons made it on this day (among others because of a train strike): Sami Domisch (IGB, as replacement for Sonja Jähnig), Pieter Boets, Rudy Vannevel and Anton Van de Putte. In addition, Thierry Vercauteren was present during the workshop on the 28th.

Members of the follow-up committee were contacted in the run-up to the final event and during this meeting. At this meeting we received the following feedback:

- The project is on track to meet its goals.
- As a type of project SAFRED is “ahead of time” and clearly sets an example and provides a template for other countries or regions.
- As such it will be very important to work out a “lessons learnt” document that lists obstacles encountered during the project and provides guidance for other interested persons to set up similar activities/projects.

The final event “*Safeguarding Biodiversity Data for the Future – Conference & workshop of the "Saving Freshwater Biodiversity Research Data" project*” on February 27th and 28th 2018 at Royal Belgian Institute of Natural Sciences, Brussels, consisted of a two full days program with a conference on the first day and a hands-on workshop on the second day.

With the conference, we aimed to attract a wide audience of scientists and data holders interested in the topic of biodiversity data publication and analysis. The morning session with three international keynote speakers (Jonathan Chase [iDiv, Leipzig, Germany], Jörg Freyhof

[Leibniz-Institute of Freshwater Ecology and Inland Fisheries (IGB)], Julia Steward Lowndes [National Centre for Ecological Analysis and Synthesis (NCEAS)]) focussed on different aspects and motivations for making biodiversity data publicly available. During the afternoon, we provided an overview of the SAFRED project and the obtained results, interacted with the members of the follow-up committee and held a final discussion on the impact and future work of the project. The full program can be found at odnature.naturalsciences.be/safred/workshops.

During the hands-on workshop targeting research institutes, individual scientists and other data holders, we discussed how to effectively work on standardisation and publication of datasets, and shared experiences and knowledge about data management and biodiversity data publication. During the morning sessions, simple guidelines were given to improve day-to-day data management in research projects, the concept of tidy data was introduced, and practiced during an exercise on structuring data in spread sheets. In the afternoon, participants were introduced to the Darwin Core data and trained in transforming biodiversity data to Darwin Core using the R package ‘dplyr’. Workshop presentations and exercises are available online at <https://inbo.github.io/dwc-in-R/>. In parallel, we ran a workshop on microbial ecology data management as outlined under Task 3.3.

Integrated analyses

The different pond databases were integrated into one central database that provides unique material for integrated testing of specific hypotheses (Workpackage 6: Test cases and outreach). The central SAFRED pond database currently comprises data of major abiotic variables and data on occurrences of multiple taxonomic organism groups in >400 lentic Belgian waterbodies (Figure 4). These data were collected following during different projects but with a similar protocol. We first conducted multiple explorative analyses to screen the consistency of all datasets and to detect major environmental gradients within each dataset. We specifically aimed (1) to identify taxonomic-functional-phylogenetic generalities that structure zooplankton metacommunities at different spatial scales, and (2) to investigate the relation between taxonomic and functional diversity in relation to eutrophication. We used cladoceran crustaceans as model organism group as they play a key role in the functioning of pond ecosystems and they are also the best sampled organism group. In addition, we restricted our analyses to relatively small farmland ponds for reasons of consistency in ecosystem type and sampling effort. In the analyses presented below, data from lakes and larger ponds were thus excluded.

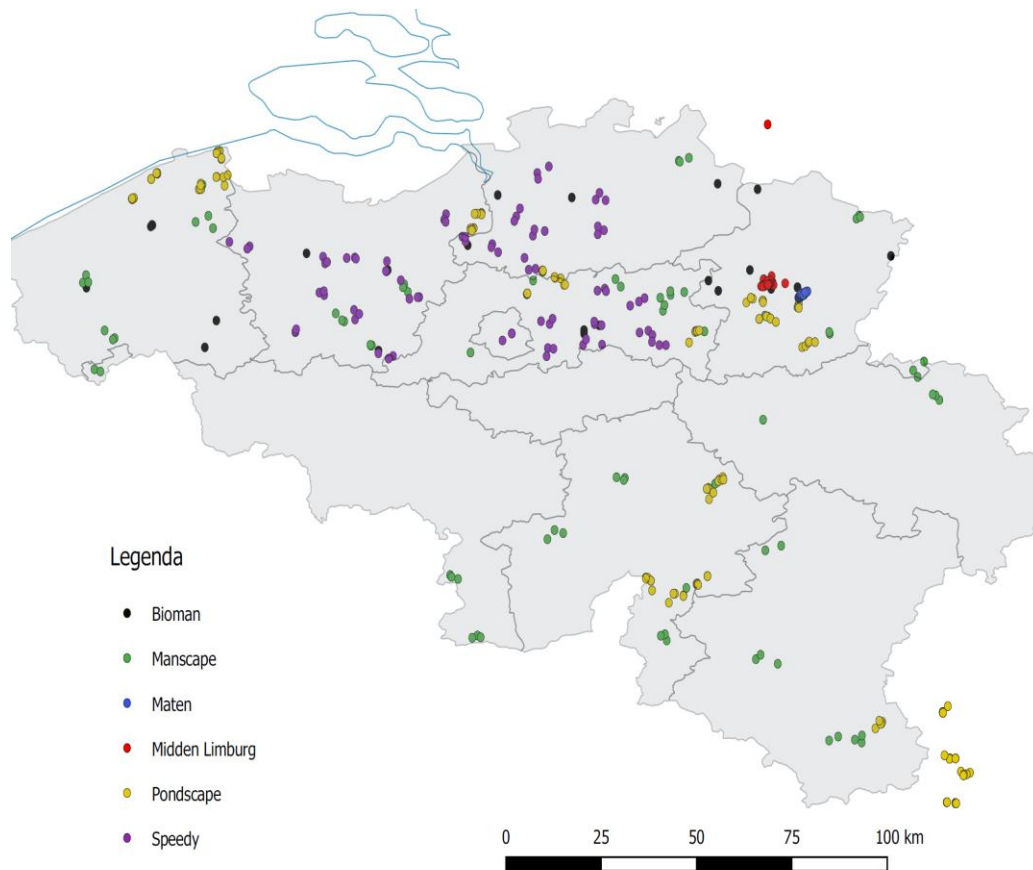


Figure 4. A map of Belgium with the location of all ponds sampled as part of different projects (indicated by different colours) and which are currently part of the central SAFRED pond database.

For objective 1, we complemented the taxonomic zooplankton database with a trait database that comprises information on nine ecological relevant traits of all cladoceran species in the central SAFRED database ($n = 52$). This information was largely extracted from published literature. In addition, we created an accompanying phylogenetic database comprising sequences of 4 markers (COI, 16S, 18S and 28S) by sequencing specimens from lab cultures (done for 17 species) and extracting additional information from GenBank. Based on this molecular information, we created a phylogenetic tree that can be used in further statistical analyses as a measure of phylogenetic diversity (Figure 5). This phylogenetic database will also be used in future research projects (currently for example in the Belspo funded project ORCA - <https://bio.kuleuven.be/eeb/ldm/ORCA>) as it includes almost all cladoceran species occurring in Belgium.

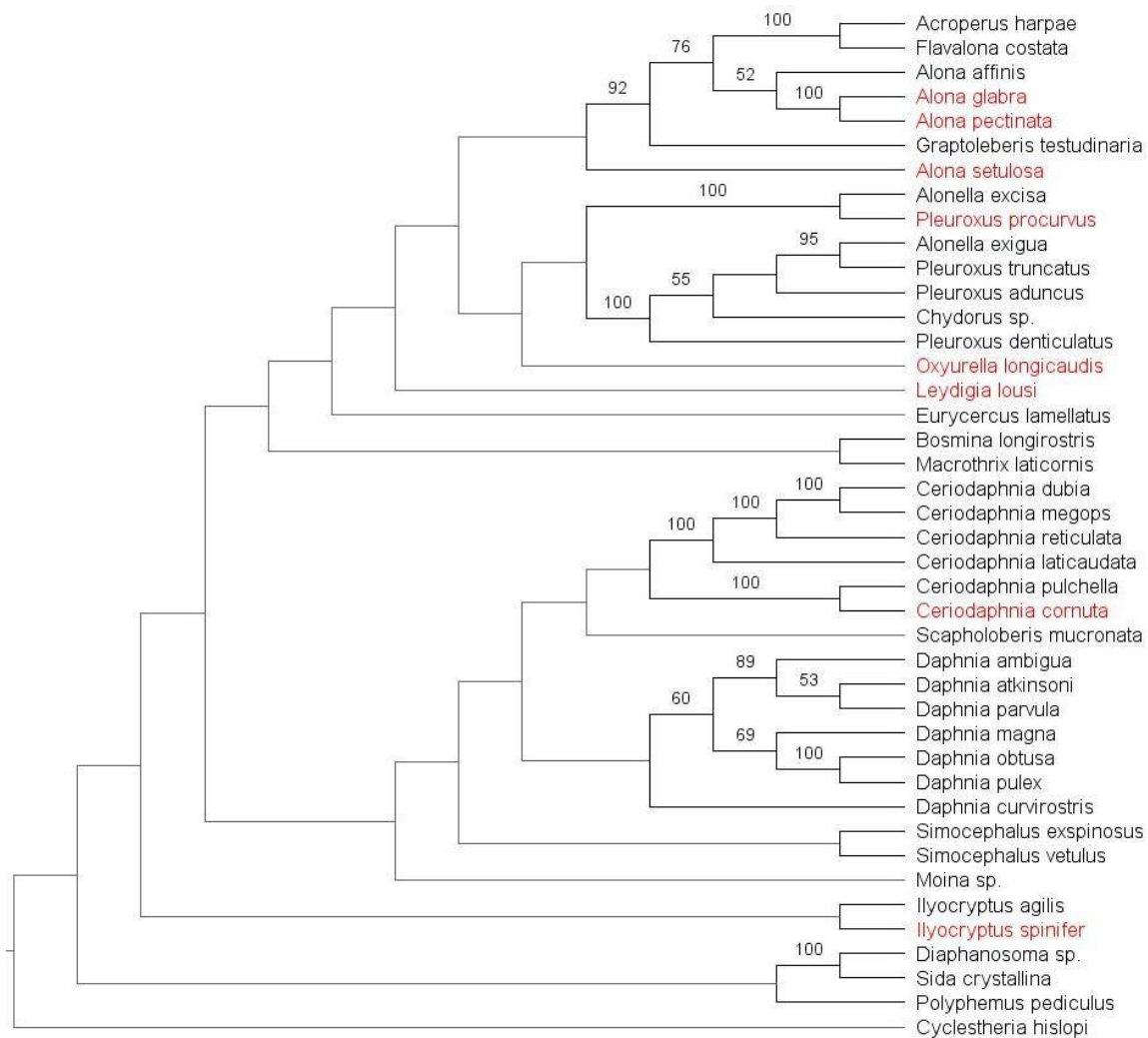


Figure 5. A phylogenetic neighbour joining tree (created in PAUP4, based on 4 markers: COI, 16S, 18S and 28S) and comprising 41 cladoceran species that occur in the integrated SAFRED pond database. The grey branches indicate relationships that were constrained based on previously recognized taxonomic grouping. Red species names indicate that these species were used as a substitute for species occurring in our dataset. Bootstrap values are given for unconstrained branches as a measure of the robustness of the tree. The method to integrate traits and phylogenetic distances to assess scale-dependent community assembly processes is described in Gianuca et al. (2017).

The overall variation in environmental conditions between ponds was visualized using an ordination plot of a principal component analysis. The first and second axes of the PCA ordination plot jointly comprise almost 37% of the variation in local environmental conditions between ponds (Figure 6.). The first axis (eigenvalue = 0.239) is positively

associated with eutrophication related variables, whereas the second axis differentiates ponds mainly based on pH and size. TP and TN were positively associated with chlorophyll a concentration and showed a negative association with water transparency, water depth, as well as with the coverage with submerged and emergent vegetation.

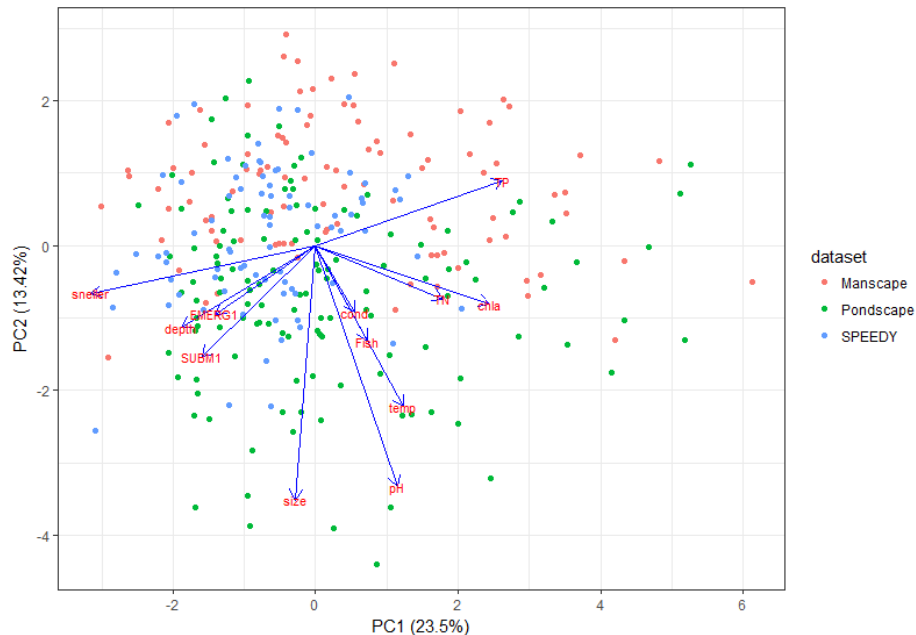
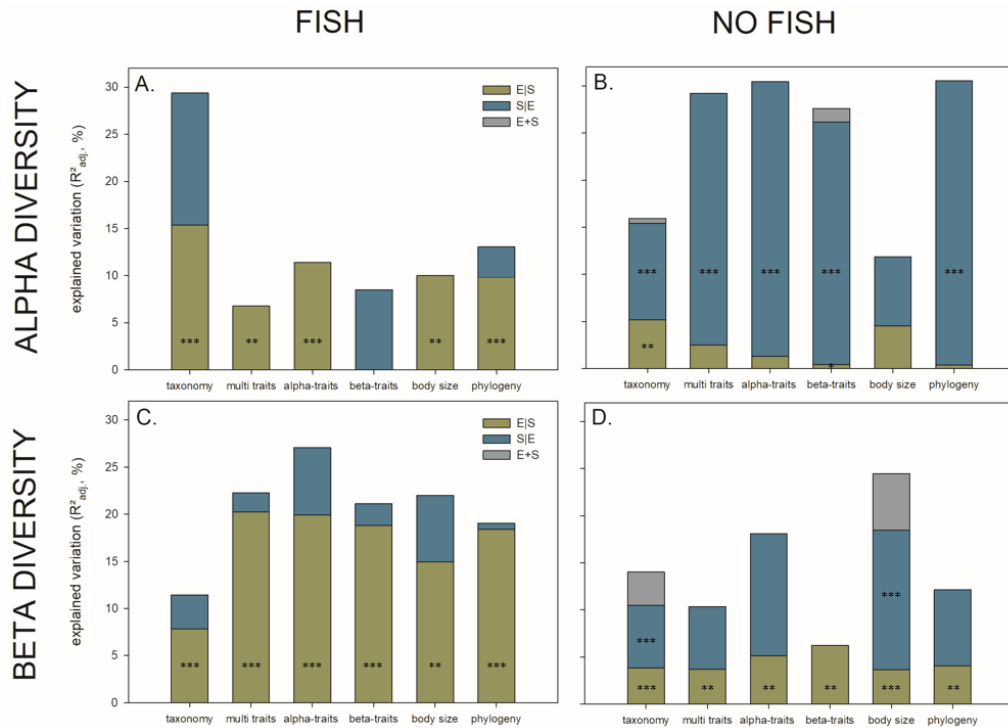


Figure 6. An ordination plot of a Principal Component Analysis (PCA) based on 12 major environmental pond variables. Coloured dots represent individual ponds. Arrows show the association between different environmental variables. Used abbreviations: pond area (size), water depth (depth), temperature (temp), conductivity (cond), water transparency (sneller), chlorophyll a concentration (chla), total phosphorus concentration (TP), total nitrogen concentration (TN), emergent macrophyte coverage (EMERG1) and submerged macrophyte coverage (SUBM1).

A first key finding of our study is that trait and phylogenetic information did not increase the overall explanatory power of space and environment in explaining variation of community composition in the set of investigated ponds (Figure 7). Based on our results we cannot support the idea that including functional and phylogenetic approaches consistently increase informative power in studies on community structure. Secondly, we observe profound differences in the mechanisms that shape community structure in ponds with and without fish. Local environmental conditions explained an important fraction of variation in community composition in ponds with fish, whereas this was not the case in ponds without fish. In contrast, spatial factors tended to be relatively more important in ponds with fish,

especially at the alpha scale. These findings suggest that environmental filtering is relatively more important in ponds with fish, whereas the dispersal limitation seems to be more relevant in ponds without fish.



*Figure 7. Result of multiple variation partitioning analyses testing for the relative importance of local environmental variables and space on variation on each dimension of diversity (taxonomic, functional and phylogenetic at the alpha and beta spatial scale (panel A and B, and panel C and D respectively) in the presence and absence of fish (left and right hand panels respectively). Stack bars show the amount of variation in community composition explained by taxonomy, multiple traits, alpha traits, beta traits, body size and phylogeny. The significance level of each fraction is indicated with ***: $p < 0.001$, **: $p < 0.01$ and *: $p < 0.05$. Bars without '*' are not significant. Note that the significance of the shared fractions cannot be tested.*

For the second objective, we selected ponds from the overall SAFRED pond database along the eutrophication gradient (represented by the first PCA axis, Figure 6) by omitting ponds with relatively large and relatively small PCA2 scores from further analysis. Subsequently, we investigated the relation between taxonomic and trait diversity in relation to eutrophication. The key findings of this analysis are the negative relation between taxonomic

diversity and eutrophication, while such negative relation is not significant between trait diversity and eutrophication (Figure 8). Consequently, we observe an important negative relation between functional redundancy and eutrophication. This finding suggest that pond ecosystems become more vulnerable to ecosystem perturbation with increasing eutrophication.

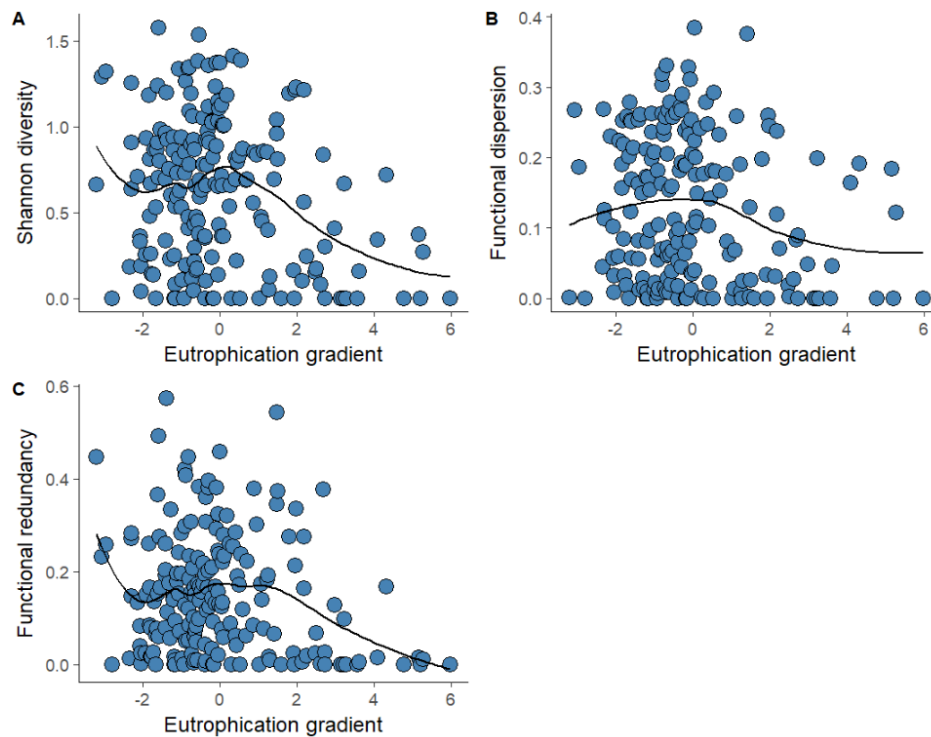


Figure 8. Taxonomic (Shannon), functional (functional dispersion) and functional redundancy along a gradient of eutrophication. The black trend line was created based on LOESS smoothing.

In addition to the above-mentioned analyses, the KU Leuven team is currently an active member of a working group with multiple international researchers focussing on the development of novel statistical methods integrating space and time into meta community analyses (sTURN working group, coordinated by Zsafia Horvath, Robert Ptacnik and Jon Chase). Since the “De Maten” database comprises data of phytoplankton, zooplankton and macro-invertebrates of 32 fish ponds from three subsequent years, it is well suited to test statistical techniques for simultaneously investigating the importance of space and time in

community assembly. The next working group meeting will be held in November 2018 at iDiv (Leipzig, Germany).

Finally, the KU Leuven team also plans an additional study to explore the patterns of biodiversity accumulation in space based on a subset of the SAFRED database. This study will be done simultaneously on multiple organism groups and will be developed in close collaboration with the team of Jon Chase (iDiv, Leipzig, Germany).

5. DISSEMINATION AND VALORISATION

The SAFRED project was presented at the following occasions:

- GEO BON Open Science Conference & All Hands Meeting - July 4–8th, 2016, Leipzig, Germany.
 - o Aaike De Wever coordinated a session on “Freshwater and wetland biodiversity monitoring and data mobilisation”
 - o Oral presentation: De Wever, A., Schmidt-Kloiber, A., Bremerich, V. Lessons learned from the freshwater biodiversity data mobilisation activities for the Freshwater Information Platform.
- Waterforum “Biologische monitoring in water, versie 2.0” – October 14, 2016, Gent. Oral presentation: De Wever, A. Beheer van data rond zoetwaterbiodiversiteit: naar een systematische aanpak en het online beschikbaar stellen van gegevens.
- Aaike De Wever was invited to talk at the “Werkgroep Ecologisch Waterbeheer” symposium in Amersfoort, The Netherlands (April 2017)

6. PUBLICATIONS

(Meta)data papers

- Lemmens, P., De Bie, T., De Roeck, E., Ercken, D., Vanhecke, L., Martens, K. & De Meester, L., 2018. Database of bomb crater pools in Tommelen nature reserve, Belgium. *Freshwater Metadata Journal* 40: 1-6.
(<https://doi.org/10.15504/fmj.2018.40>).
- Sweetlove, M., Van Wichelen, J., Verleyen, E. & Vyverman, W., 2018. Flemish rivers phytoplankton. *Freshwater Metadata Journal* 39: 1-4
(<https://doi.org/10.15504/fmj.2018.39>).
- Sweetlove, M., Van Wichelen, J., Verleyen, E. & Vyverman, W., 2018. Ecological potential datasets (Ecopot). *Freshwater Metadata Journal* 38: 1-4
(<https://doi.org/10.15504/fmj.2018.38>).
- Lara, Y., Wilmotte, A., Sivonen, K., De Bellis, G., Kuuppo, P., Ventura, S., Montarani, B., Henderson, P., Hoffmann, L., Zalewski, M. & Komárek, J., 2018. Metadata compilation for the MIDI-CHIP dataset. *Freshwater Metadata Journal* 37: 1-8 (<https://doi.org/10.15504/fmj.2018.37>).
- Latli, A., Chérot, F. & Kestemont, P., 2018. Macroinvertebrate data of the Belgian River Meuse from 1998 to 2011. *Freshwater Metadata Journal* 34: 1-5
(<https://doi.org/10.15504/fmj.2018.34>).
- Latli, A., Ovidio, M. & Kestemont, P., 2018. Fish data of the Belgian River Meuse from 1989 to 2012. *Freshwater Metadata Journal* 33: 1-5
(<https://doi.org/10.15504/fmj.2018.33>).
- Latli, A., Service Public de Wallonie, Kestemont, P., RIWA & CIM-Meuse, 2018. Physicochemical data of the Belgian River Meuse from 1972 to 2010. *Freshwater Metadata Journal* 32: 1-5 (<https://doi.org/10.15504/fmj.2018.32>).
- Lemmens, P. A. De Wever, N. Bonjean, A. Castiaux, L. Colson, T. De Bie, E. Decoster, C. Denis, E. De Roeck, D. Ercken, B. Goddeeris, R. Goyvaerts, S. N.M. Mandiki, K. Morelle, E. Praca, I. Schön, J. Van Wichelen, K. Van Der Gucht, J. Vandekerckhoven, P. Vanormelingen, M. Villena Alvarez, D. Bauwens, L. Denys, M. Herremans, L. Vanhecke, B. Losson, Y. Caron, H.-M. Cauchie, P. Kestemont, W. Vyverman, L. De Meester, S. A.J. Declerck & K. Martens 2018. Database of the PONDSCAPE project (Towards a sustainable management of pond diversity at the

landscape level). *Freshwater Metadata Journal* 31: 1-12

(<https://doi.org/10.15504/fmj.2018.31>)

- Lemmens, P., K. Cottenie, F. Van de Meutter, P. Vanormelingen & L. De Meester 2018. Database of interconnected fish ponds in De Maten Nature Reserve, Belgium. *Freshwater Metadata Journal* 30:1-8. (<https://doi.org/10.15504/fmj.2018.30>).
- Vannevel R, Brosens D, De Cooman W, Gabriels W, Lavens F, Mertens J, Vervaeke B 2018 The inland water macro-invertebrate occurrences in Flanders, Belgium. *ZooKeys* 759: 117-136 (<https://doi.org/10.3897/zookeys.759.24810>)
- Lemmens, P., De Wever, A., Vandekerckhove, J., Muylaert, K., Van der Gucht, K., Zwart, G., Rommens, W., Van Wichelen, J., Geenens, V., Vyverman, W., Brendonck, L., Martens, K., Declerck, S.A.J. & De Meester, L., 2018. Database on environmental conditions and biodiversity in shallow lakes in Belgium and the Netherlands. *Freshwater Metadata Journal* 29: 1-9(<https://doi.org/10.15504/fmj.2018.29>)
- Lara, Y., De Wever, A., Verniers, G., Pirlot, S., Viroux, L., Leporcq, B., Vanormelingen, P., Van Wichelen, J., Van der Gucht, K., Peretyatko, A., Tessier, S., Lambion, A., Reilly, M., Menzel, D., Wojnicz, A., Codd, G.A., Triest, L., Vyverman, W., Descy, J.-P. & Wilmotte, A., 2017. Metadata compilation for the B-BLOOMS2 dataset: Cyanobacterial bloom monitoring. *Freshwater Metadata Journal* 28: 1-8 (<https://doi.org/10.15504/fmj.2017.28>)
- Lemmens, P., Mergeay, J., Ercken, D., De Bie, T., Van Wichelen, J., Declerck, S.A.J. & De Meester, L., 2017b. Database on local environmental conditions and biodiversity in fish ponds in Midden-Limburg, Belgium. *Freshwater Metadata Journal* 27: 1-8 (<https://doi.org/10.15504/fmj.2017.27>)
- Lemmens, P., De Wever, A., Hampel, H., De Bie, T., Ercken, D., Van Wichelen, J., Denys, L., Goddeeris, B., Mandiki, S.N.M, Vanhecke, L., van der Gucht, K., Bauwens, D., Denayer, S., Durinck, R., Dasseville, R., Lionard, M., van De Meutter, F., Louette, G., Hulsmans, A., De Gelas, K., Schön, I., Vrijders, H., Maes, A., Losson, B., Lasri, S., Kestemont, P., Vyverman, W., Vanormelingen, P., Brendonck, L., De Meester, L., Declerck, S.A.J. & Martens, K., 2017a. Database of the MANSCAPE project (Management tools for water bodies in agricultural landscapes). *Freshwater Metadata Journal* 26: 1-11 (<https://doi.org/10.15504/fmj.2017.26>)

Datasets published (see also Table I for complete list)

- B-BLOOMS2: Verniers, G., Pirlot, S., Viroux, L., Leporcq, B., Vanormelingen, P., Van Wichelen, J., Van der Gucht, K., Peretyatko, A., Teissier, S., Lara, Y., Lambion, A., Reilly, M., Menzel, D., Wojnicz, A., Everbecq, E., Codd, G.A., Triest, L., Vyverman, W., Wilmotte, A., Descy, J.P. (2018). Cyanobacterial blooms: toxicity, diversity, modelling and management “B-BLOOMS2”.
<http://data.freshwaterbiodiversity.eu/ipt/resource?r=sf13-bblooms13>
- River MEUSE : Latli Adrien, Service Public de Wallonie - DG03, RIWA-Maas, CIM-Meuse (2017). Physicochemical evolution of the Belgian River Meuse from 1972 to 2010. http://data.freshwaterbiodiversity.eu/ipt/resource?r=sf3-meuse_physicochemistry
- RIVER MEUSE : Latli Adrien, Ovidio Michael, Kestemont Patrick (2017). Fish abundance evolution in the Belgian River Meuse from 1989 to 2012.
http://data.freshwaterbiodiversity.eu/ipt/resource?r=sf6-meuse_fish
- BIOMAN: Pieter Lemmens, Aaike De Wever, Jochen Vandekerckhove, Koenraad Muylaert, Katleen Van der Gucht, Gabriel Zwart, Wouter Rommens, Jeroen Van Wichelen, Vanessa Geenens, Wim Vyverman, Luc Brendonck, Koen Martens, Steven A.J. Declerck and Luc De Meester. (2017). Biodiversity in shallow lakes along a gradient of eutrophication.
http://data.freshwaterbiodiversity.eu/ipt/resource?r=bioman_belgium
- MIDDEN LIMBURG: Pieter Lemmens, Aaike De Wever, Joachim Mergeay, Tom De Bie, Dirk Ercken, Jeroen Van Wichelen, Steven A.J. Declerck, Luc De Meester (2017). The importance of management for biodiversity conservation in interconnected man-made ponds.
<http://data.freshwaterbiodiversity.eu/ipt/resource?r=midden-limburg>
- MANSCAPE : Lemmens P., De Wever A., Hampel H., De Bie T., Ercken D., Van Wichelen J., Denys L., Goddeeris B., Mandiki S.N.M., van Hecke L., van der Gucht K., Bauwens D., Durinck R., Denayer S., Dasseville R., Lionard M., van De Meutter F., Louette G., Hulsmans A., De Gelas K., Schön I., Vrijders H., Maes A., Losson B., Lasri S., Kestemont P., Vyverman W., Vanormelingen P., Brendonck L., De Meester L., Declerck S.A., Martens K. (2017). Management tools for waterbodies in agricultural landscapes. <http://data.freshwaterbiodiversity.eu/ipt/resource?r=manscape>
- Vannevel R, De Cooman W, Gabriels W, Lavens F, Mertens J, Vervaeke B, Brosens

D (2017): Inland water macroinvertebrate occurrences in Flanders, Belgium. v1.8.

Flanders Environment Agency. Dataset/Samplingevent.

<https://doi.org/10.15468/4cvbka>. <https://ipt.inbo.be/resource?r=vmm-macroinvertebrates-events>

- Boets P, Brosens D, Lock K, Adriaens T, Aelterman B, Mertens J, Goethals P L (2014): Alien macro-invertebrates in Flanders, Belgium. v1.8. Research Institute for Nature and Forest (INBO). Dataset/Occurrence. <https://doi.org/10.15468/xjtfoo>. <http://data.inbo.be/ipt/resource?r=alien-macroinvertebrate-occurrences>
- WATER BIRDS : Devos K, T'Jollyn F, Brosens D, Desmet P (2014): Watervogels - Wintering water birds in Flanders, Belgium. v3.5. Research Institute for Nature and Forest (INBO). Dataset/Occurrence. <https://doi.org/10.15468/lj0udq>. <https://ipt.inbo.be/resource?r=watervogels-occurrences>
- Vanderhaeghe F, De Blust G, Brosens D (2018). InboVeg - Amphibious vegetation of shallow lakes at Turnhout. Version 1.8. Research Institute for Nature and Forest (INBO). Sampling event Dataset <https://doi.org/10.15468/xkuvwe>. <https://ipt.inbo.be/resource?r=inboveg-turnhout-ponds-events>

Dataset related publications

- Publication related to the Meuse data: Latli A., Descy J.P., Mondy C. P., Floury M., Viroux L., Otjacques W., Marescaux J., Depiereux E., Ovidio M., Usseglio-P. P. and Kestemont. P. (2017). Long-term trends in trait structure of riverine communities facing predation risk increase and trophic resource decline. *Ecological Applications*, 27(8), 2458 – 2474.
- Publication related to the pond data: Verbeek, L., Vanhamel, M., van den Berg, E., Hanashiro, F. T. T., Gianuca, A. T., Striebel, M., Lemmens, P., Declerck, S. A. J., Hillebrand, H. and De Meester, L. (2018). Compositional and functional consequences of environmental change in Belgian farmland ponds. *Freshwater Biology*. <https://doi.org/10.1111/fwb.13095>

7. ACKNOWLEDGEMENTS

List of the members of the follow-up committee

- Rudy Vannevel, Vlaamse Milieumaatschappij (VMM), Afdeling Rapportering Water
- Pieter Boets, Provinciaal Centrum voor Milieuonderzoek
- Steven Declerck, Nederlands Instituut voor Ecologie (NIOO-KNAW)
- Patrick Meire, Universiteit Antwerpen (UAntwerpen)
- Erwin De Meyer, Agentschap Natuur & Bos (ANB), Entiteit Strategie &
- Daniel Hering, Universität Duisburg-Essen (UDE, Germany), Faculty of Biology, Aquatic Ecology
- Sonja Jähnig, Leibniz-Institute of Freshwater Ecology and Inland Fisheries (IGB, Germany)
- Sami Domisch, Leibniz-Institute of Freshwater Ecology and Inland Fisheries (IGB, Germany)
- Thierry Vercauteren, APB Provinciaal Instituut voor Hygiëne (APB PIH), Provincie Antwerpen
- Pieter Vanormelingen, Natuurpunt Studie
- Anton Van de Putte, RBINS, Biodiversity.aq
- François Darchambeau, Département de l'Etude du milieu naturel et agricole – DEMNA

REFERENCES

Dejenie, T., T. Asmelash, et al. (2008) Limnological and ecological characteristics of tropical highland reservoirs in Tigray, Northern Ethiopia. *Hydrobiologia* 610: 193-209.

Dejenie, T., S. A. J. Declerck, et al. (2012) Cladoceran community composition in tropical semi-arid highland reservoirs in Tigray (Northern Ethiopia): A metacommunity perspective applied to young reservoirs. *Limnologia - Ecology and Management of Inland Waters* 42: 137-143.

De Meester, L., L. Forro, et al. (2002) The status of some exotic cladoceran (Crustacea: Branchiopoda) species in the Belgian fauna. *Bulletin de l' Institut Royal des Sciences naturelles – Biologie* 72: 87-88.

Dudgeon D, Arthington AH, Gessner MO, Kawabata Z-I, Knowler DJ, Lévêque C, Naiman RJ, Prieur-Richard AH, Soto D, Stiassny MLJ, Sullivan CA (2006) Freshwater biodiversity: importance, threats, status and conservation challenges. *Biological reviews* 81:163–182.

Edwards PN, Mayernik MS, Batcheller AL, Bowker GC, Borgman CL (2011) Science friction: Data, metadata, and collaboration. *Social Studies of Science* 41:667–690.

Forró, L., L. De Meester, et al. (2003) An update on the inland cladoceran and copepod fauna of Belgium, with a note on the importance of temporary waters. *Belgian Journal of Zoology* 133: 31-36

Heilpern S (2015) Biodiversity: Include freshwater species. *Nature* 518: 167.
doi:10.1038/518167d

Hudnell HK, Steffensen DA (2008) Economic cost of cyanobacterial blooms. *Cyanobacterial Harmful Algal Blooms: State of the Science and Research Needs* (Hudnell HK, ed.), pp. 855–865. Springer, Heidelberg.

Huisman J, H.C.P. Matthijs, P.M. Visser 2005. *Harmful cyanobacteria*. Springer aquatic ecology, Series 3, Springer, Dordrecht, The Netherlands.

Louette, G. & L. De Meester (2005) High dispersal capacity of cladoceran zooplankton in newly founded communities. *Ecology*, 86: 353-359.

Louette, G., L. De Meester & S. Declerck (2008) Assembly of zooplankton communities in newly created ponds. *Freshwater Biology*, 53: 2309-2320.

Teferi, M., S. A. J. Declerck, et al. (2014) Strong effects of occasional drying on subsequent water clarity and cyanobacterial blooms in cool tropical reservoirs. *Freshwater Biology* 59: 870-884.

Yilmaz P, Kottmann R, et al. (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIXS) specifications. *Nature Biotechnology* 29:415–420.

ANNEX 1

UNPUBLISHED Report on Task 3.1. Horizon scanning of developments in standardization of data in molecular studies on environmental samples

This task is part of the Workpackage 3. Standardized publication of (freshwater) microbial ecology research data of the BELSPO-BRAIN Saving Freshwater Biodiversity Research Data (SAFRED) project

Authors: Yannick Lara, Maxime Sweetlove, Elie Verleyen, Annick Wilmotte

Over the last two decades, evolutionary biology including environmental studies benefited from the tremendous amount of genetic sequences that have been accumulated (Parr et al., 2012). For instance, the knowledge concerning prokaryotic evolution was blown by “the wind of change” thanks to the use of 16S rRNA gene sequences as taxonomic marker (Olsen et al., 1994). In parallel, the number of publications mentioning the terms ‘molecular’ and ‘phylogeny’ followed an exponential growth since the eighties of the previous century (Pagel, 1999; Page, 2007). This increasing trend in the number of sequences has accelerated in recent years with the advent of High Throughput Sequencing (HTS) techniques such as Illumina MiSeq and 454 pyrosequencing (Luo et al., 2012).

In order to publish studies that include molecular data in peer-reviewed journals, a consensus was reached. This included the obligation that authors need to deposit the sequences in one of the International Nucleotide Sequence Database Collaboration (INSDC) partner databases (incl. EMBL-EBI, NCBI-GenBank, DDBJ) (Cochrane et al., 2011). The classical deposition process implied that the depositor should include information about the nature and origin of the sequences. However, the responsibility to add these details was left to the good will of the depositor. Therefore, sequences are often poorly documented in INSDC databases. Indeed, the percentage of sequences without any taxonomic qualifier increased each year (Parr et al., 2012). Therefore, it has become a very arduous task to look for a specific query such as organisms’ habitat, biogeography, or toxin production.

In addition, microbial ecological data often consist of a complex mixture of genetic data obtained using different analytical cultivation-dependent (strain isolation and characterization) or -independent techniques (e.g. DGGE, TGGE, clone libraries, ARDRA, DNA chips, diverse sequencing methods), possibly in combination with microscopy as well as environmental data. However, the link between molecular and other types of data is not available.

During this SAFRED task, a horizon scanning of the developments in standardization of data in molecular studies on environmental samples was conducted in 2016. Strategies were subsequently elaborated in order to deposit freshwater genetic data in such a way that it could be clearly identified, is visible for the scientific community and re-usable in future studies.

1. Molecular data

Briefly, molecular data can be obtained using two main approaches. The cultivation-dependent method consists of the isolation of a target organism before DNA extraction, sequence amplification, and sequencing. It produces a unique sequence originating from a unique organism. The second approach, the so called ‘cultivation- independent method’, is performed on mixtures of organisms, and is used to characterize microbial communities directly taken from their environment. This approach can be based on amplification of small targeted genomic fragments (‘amplicons’, often based on the SSU rRNA gene), or complete fragmented genomes, and can use cloning, or direct high throughput sequencing.

Molecular techniques are in constant evolution. The rapid improvement of sequencing technology has mainly contributed to the accumulation of sequences in the databases but also in the researchers’ computational devices. The nucleotide database size was multiplied by 4.64 since the release of the first next generation sequencer in 2005 (Figure 1). Indeed, before the development of high throughput sequencing, microbial ecology researchers generated up to a maximum of hundreds to thousands of sequences per study, whereas it is now possible to generate 400 million sequences with a run of an Illumina NextSeq series system. Therefore, deposition strategies have evolved in order to store the gigantic amount of data generated.

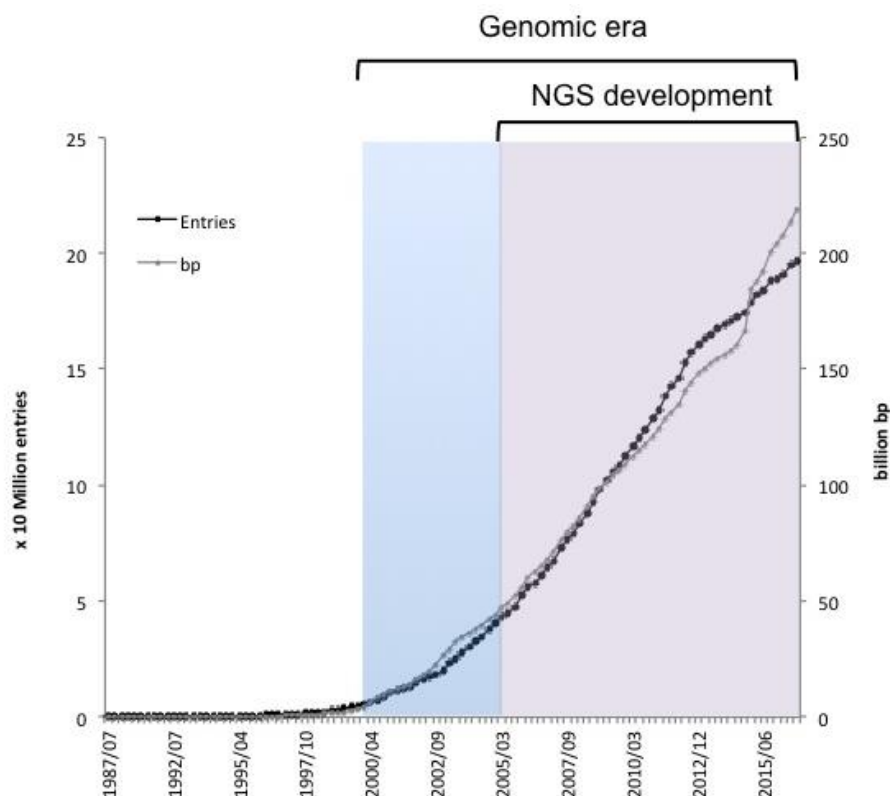


Figure 1. Nucleotides and Entries accumulation in EMBL-EBI, NCBI-GenBank, DDBJ database.

2. Sequence Databases for deposition and The Genomic Standards Consortium

The International Nucleotide Sequence Database Collaboration (INSDC; <http://www.insdc.org>) started from an initiative taken by three organizations in 1987, the DNA Databank of Japan (DDBJ) at the National Institute for Genetics in Mishima, Japan; the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) in Hinxton, UK; and the National Centre for Biotechnology Information (NCBI) in Bethesda, Maryland, USA. Its main goal is to gather molecular data and annotations in a comprehensive, accessible and unified database. The deposition of a sequence through one of the three partners is sufficient, and data are accessible using the same accession number via each platform. Fig. 2 shows a view of the Genbank fields associated with a freshwater cyanobacterial strain sequence.

Snowella litoralis 1LT47S05 partial 16S rRNA gene, strain 1LT47S05
 GenBank AJ781041.1
[FASTA](#) [Graphics](#)

[Go to](#)

LOCUS AJ781041 1444 bp DNA linear BCT 22-FEB-2006
DEFINITION Snowella litoralis 1LT47S05 partial 16S rRNA gene, strain 1LT47S05.
ACCESSION AJ781041
VERSION AJ781041.1
KEYWORDS 16S ribosomal RNA; 16S rRNA gene.
SOURCE Snowella litoralis 1LT47S05
ORGANISM [Snowella litoralis 1LT47S05](#)
 Bacteria; Cyanobacteria; Synechococcales; Coelosphaeriaceae;
 Snowella.

REFERENCE 1
AUTHORS Rajaniemi-Macklin, P., Rantala, A., Mugnai, M.A., Turicchia, S.,
 Ventura, S., Komarkova, J., Lepisto, L. and Sivonen, K.
TITLE Correspondence between phylogeny and morphology of Snowella spp.
 and Woronichinia naegeliana, Cyanobacteria commonly occurring in
 lakes
JOURNAL J. Phycol. 42 (1), 226-232 (2006)
REFERENCE 2 (bases 1 to 1444)
AUTHORS Rajaniemi, P.
TITLE Direct Submission
JOURNAL Submitted (30-JUN-2004) Rajaniemi P., Applied chemistry and
 microbiology, Helsinki University, P.O. Box 56, Viikinkaari 9,
 00014 Helsinki, FINLAND

FEATURES Location/Qualifiers
 source
 1..1444
 /organism="Snowella litoralis 1LT47S05"
 /mol_type="genomic DNA"
 /strain="1LT47S05"
 /isolation_source="Lake Trasimeno"
 /db_xref="taxon:371758"
 /country="Italy"
 <1..>1444
 /gene="16S rRNA"
 <1..>1444
 /gene="16S rRNA"
 /product="16S ribosomal RNA"

[gene](#)
[rRNA](#)

ORIGIN
 1 ttgctcagga tgaacgtgg cggatgctt aacacatgca agtcgaacgg aatcttcgga

Change region shown
Customize view
Analyze this sequence
[Run BLAST](#)
[Pick Primers](#)
[Highlight Sequence Features](#)
[Find in this Sequence](#)
Related information
[Full text in PMC](#)
[Taxonomy](#)
LinkOut to external resources
[Ribosomal Database Project II](#)
 [Ribosomal Database Project II]
[SILVA SSU Database](#)
 [SILVA]
Recent activity
[Turn Off](#) [Clear](#)
 Snowella litoralis 1LT47S05 partial 16S
 rRNA gene, strain 1LT47S05 Nucleotide
 Snowella (10)
 Nucleotide
[See more...](#)

Fig. 2. Genbank information example

To face the impressive increase of the data volume, INSDC partners also provided new strategies of storage, such as the SRA (Short Read Archive) and DRA (DDBJ Sequence Read Archive). Recently, in order to better describe the data available, the INSDC integrated BioProject and BioSample data to the database (Cochrane et al., 2011).

The Genomic Standards Consortium (GSC) was formed in 2005. As an answer to the avalanche of genomic, metagenomic, and amplicons data, the GSC recognized the need to integrate contextual data in a machine-readable way (Garrity et al., 2008). Simultaneously, the GSC started the Genomic Rosetta Stone project that aimed to map genomes identifiers across INSDC databases (Van Brabant et al., 2008).

In order to facilitate the future comparative analyses of genomic data, the GSC elaborated a checklist to give the ‘Minimum Information about a Genome Sequence’ specification (MIGS) report (Field et al., 2008). Soon after that, the MIGS standards have been extended to metagenomics data (Garrity et al., 2008). The checklist aimed to describe project specifications, organisms, descriptors (collection date, environment, geographic location,

investigation type, ...), specific descriptors such as assembly details, isolation and growth conditions. The applicable environmental packages gather measurements and observations specific to a type of environment (*e.g.*, marine, terrestrial, water...). In 2011, Yilmaz et al. elaborated the ‘Minimum Information about a Marker gene Sequence’ (MIMARKS) and the ‘Minimum Information about any Sequence’ (MIXS) specifications. The latter allows documenting the genetic diversity data from diverse sources (technical or sample locations) (Figure 3). Finally, vocabulary terms were reviewed, so that MixS standards can be aligned with the Darwin Core biodiversity data standard (Tuama et al., 2012).

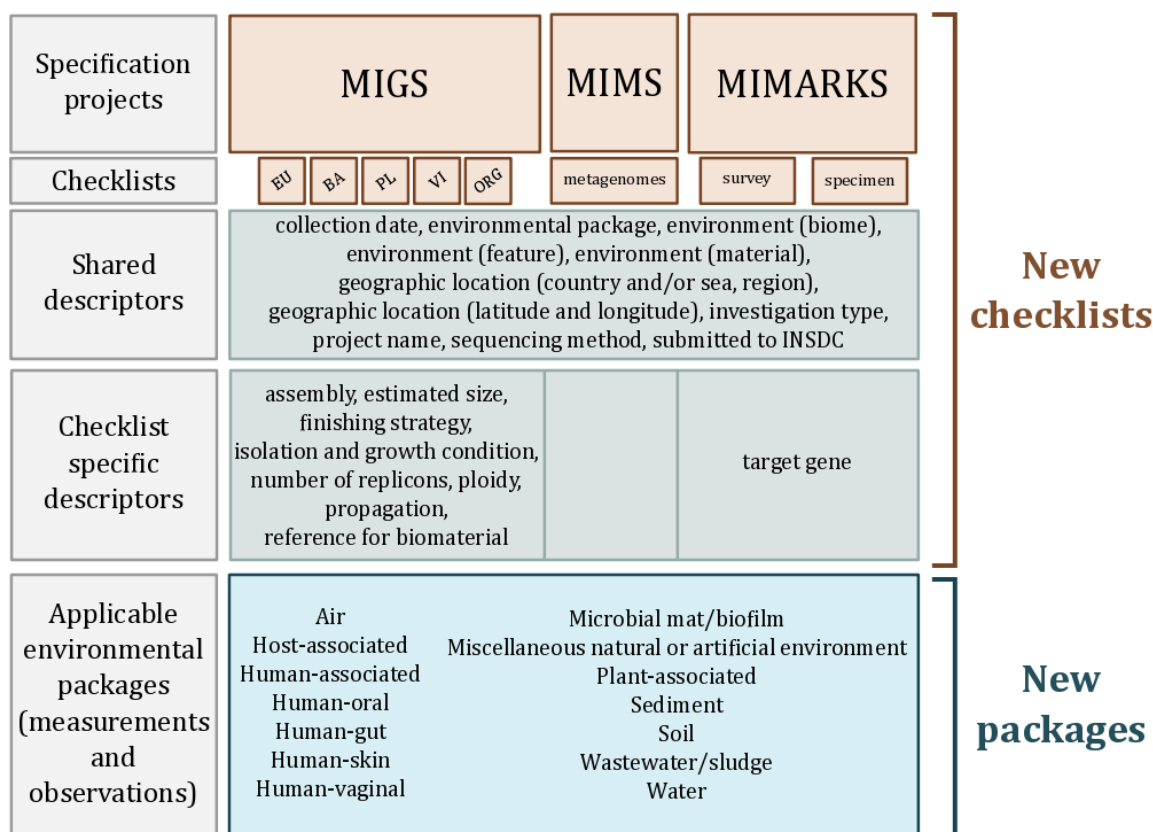


Figure 3. MIXS description (from Yilmaz et al., 2011)

3. Data resources management initiative

It is now common knowledge that microorganisms are key components of our ecosystem as they play a major role in biochemical fluxes (*e.g.* carbon, nitrogen, sulphur, oxygen cycles). With the increasing amount of data available since the advent of high throughput sequencing and the ‘-omics’ era, several initiatives have been launched to integrate multiple sources of data, such as environmental data, geographic location, biological observations, and molecular

data. Among them, the megx.net (<http://mb3is.megx.net/>) is a web portal that provides a collection of georeferenced marine prokaryotic genomes and metagenomes, the so called MegDB for ‘Microbial Ecological Genomics Database’ (Lombardot et al., 2006). The latter arises from an effort to standardize and organize the data storage, and a centralization of data access and interpretation. The MegDB database integrates geo-localization data (GIS), environmental parameters, gene functions, 16S/18S rRNA gene sequences, and (meta)genomes for a set of prokaryotes and their contextual data (MiMarks/MIGS, MIMS). The main objective of this initiative is to provide necessary tools and data to address environmentally relevant questions such as microorganisms’ adaptations to oceanic regions (Lombardot et al., 2006). For instance, using the megx.net server, it is now possible to map functions or microorganisms. This is possible through the integration of environmental data using a Global Information System and a Geographic-BLAST (Kottmann et al., 2010).

In a similar fashion, the Microbial Antarctic Resource System (mARS: http://mars.biodiversity.aq/site_pages/home) is designed to organize and centralize molecular diversity (meta)data generated by Antarctic researchers, in a way that allows its visibility, access and the analysis of geo-referenced (meta)data. The mARS platform proposes to describe and to reference datasets using GBIF depositions, MiMarks/MIGS/MIMS/MIxS templates and ‘sequence set template’. The latter was provided by the mARS team and guarantees access to molecular data by listing GenBank accession numbers or data repositories (Figure 4). In addition, technical information such as primer sequences, or sequencing technologies are listed.

Structured Comment Name	Field Descriptor	Sequence_Set_1	Sequence_Set_2	Sequence_Set_3	Sequence_Set_4	Sequence_Set_5	Sequence_Set_6
unique_sequence_set_id	Unique Sequence Set ID	SWI_Bact_20Aug2002	SWI_Arch_20Aug2002	SWI_MG_20Aug2002	SWI_Bact454_20Aug2002	SWI_Bact454_17Jan2002	SWI_Bact454_17
collection_date	Sample collection date	20/08/02	20/08/02	20/08/02	20/08/02	17/01/02	
study_type	Study type (marker gene, genome, metagenome, metatranscriptome, metaproteome)	marker gene	marker gene	metagenome	marker gene	marker gene	marker gene
target_gene	target marker gene (e.g.: 16S rRNA)	16S rRNA	16S rRNA	na	16S rRNA	16S rRNA	16S rRNA
target_taxa	Target taxa (e.g.: Bacteria, Archaea, Eukarya)	Bacteria	Archaea	Bacteria, Archaea, > 1.6 micron Eukarya	Bacteria	Bacteria	Bacteria
region_targeted	Region targeted (e.g. V6)	na	na	na	V6	V6	V6
forward_primer	Forward primer (e.g.: BACT27F)	BACT27F	Arch4aF	na	967F	967F	967F
reverse_primer	Reverse primer (e.g.: UNIV1391R)	UNIV1391R	Univ1391R	na	1046R	1046R	1046R
forward_primer_sequence	Forward primer sequence	agagtttgatcctgcctcag	tcaggttgatcctgcctcag	na	CNACGCGAAGAACCTTACC, CAACGCGAAGAACCTTACC, CAACGCGAAGAACCTTACC, ATACCGGARGAACCTTACC, CTACCGGARGAACCTTACC, CGACAGCCATGCANACCT, CGACAGCCATGCANACCT, CGACAGCCATGCANACCT	CNACGCGAAGAACCTTACC, CAACGCGAAGAACCTTACC, CAACGCGAAGAACCTTACC, ATACCGGARGAACCTTACC, CTACCGGARGAACCTTACC, CGACAGCCATGCANACCT, CGACAGCCATGCANACCT, CGACAGCCATGCANACCT	CNACGCGAAGAACCTTACC, CAACGCGAAGAACCTTACC, CAACGCGAAGAACCTTACC, ATACCGGARGAACCTTACC, CTACCGGARGAACCTTACC, CGACAGCCATGCANACCT, CGACAGCCATGCANACCT, CGACAGCCATGCANACCT
reverse_primer_sequence	Reverse primer sequence	gacggcgctgwgtrca	gacggcgctgwgtrca	na	967	967	967
primer_5	5' base of primer target in marker gene	27	27	na	967	967	967
primer_3	3' base of primer target in marker gene	1391	1391	na	1046	1046	1046
genbank_contigs	Number of sequences or contigs in Genbank data set or raw reads in short read archive	639	505	16561	12662	23733	
metagenome_sequencing_approach	Sequencing approach for metagenome (e.g.: direct shotgun, small or large insert library)	na	na	large insert fosmid library	na	na	na
sequencing_technology	Sequencing technology (e.g.: 454, Illumina, Sanger,...)	Sanger	Sanger	Sanger, 454 and Illumina	454	454	454
seq_yr	Output filetype for next generation sequence data (eg: SFF, iseq, qseq)	na	na	abi, SFF, fastaq	FASTA	FASTA	FASTA
run_type	Is the next generation sequence output file available to link to?	na	na	no	yes	yes	yes
output_filetype_for_next_generation_sequence_data	Sequence data repository ID(s) (e.g.: GenBank, SRA,...)	GenBank	GenBank	GenBank; IMG-M	VAMPS	VAMPS	VAMPS
has_next_generation_sequence_out_data	Website where data is deposited (if GenBank, not necessary)			http://img.jgi.doe.gov/cgi-bin/m/main.cgi	http://vamps.mbl.edu/utlis/login.php	http://vamps.mbl.edu/utlis/login.php	http://vamps.mbl.edu/utlis/login.php
sequence_data_repository	GenBank accession number(s), accession number range, or other project identifier(s) (e.g.: GsDno, EF069336 to EF069388) (aD)	GU234688-GU235327	GU234182-GU234687	Genbank-ADKQ00000000; IMG: 2008193001, 2040502004, 2077657020	CAM_0003_2002_08_20	CAM_0001_2002_01_17	CAM_0002_2002_01_17

Figure 4, Example of a sequence set template downloaded via the mARS platform

4. Belgian freshwater molecular data deposition strategy

Modern microbial ecology studies in freshwater ecosystems often consist of a mix of environmental analyses, combined with observations or counts and molecular analyses. While metadata and occurrence datasets may be uploaded through GBIF, molecular analyses and minimum information about molecular datasets is uploaded in the GenBank repository (or another INSDC). Genbank and GBIF are the two most important biodiversity data repositories. However, there is currently no link between the GenBank and the GBIF entries. Such combination would represent a promising tool to study the distribution of freshwater species, as well as to describe species' niches. For example, in the case of the lichen fungus *Usnea longissima*, it was possible to show that this species was not found in tropical regions and the Southern hemisphere by using a combination of GBIF and GenBank entries and drawing a predictive niche model (Smith et al., 2016).

Albeit it is on a third platform, the mARS portal is providing access to both GBIF and GenBank for the same dataset and would be ideally suited in the framework of SAFRED, to

deposit high throughput sequencing data on (sub-)polar organisms. In conclusion, depending on the project, we propose to store sequence templates and MiMarks/MIMS/MIGS/MiXS in open source data repositories, such as GenBank and GBIF. These records will contain the metadata information and link to both occurrences and molecular datasets with standard compliant contextual data.

References

Cochrane G, Karsch-Mizrachi I, Nakamura Y (2011) The international nucleotide sequence database collaboration. *Nucleic Acids Research* 39: D15-18.

Field D, Garrity G, Gray T, Morrison N et al. (2008) The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnology* 26:541-547.

Garrity GM, Field D, Kyrpides N, Hirschman L and others (2008) Toward a standards-compliant genomic and metagenomic publication record. *OMICS* 12:157-160.

Kottmann R, Kostadinov I, Duhaime MB, Buttigieg PL and others (2010) Megx.Net: Integrated database resource for marine ecological genomics. *Nucleic Acids Research* 38: D391-395.

Lombardot T, Kottmann R, Pfeffer H, Richter M, Teeling H, Quast C, Glockner FO (2006) Megx.Net--database resources for marine ecological genomics. *Nucleic Acids Research* 34: D390-393.

Luo, C., Tsementzi, D., Kyrpides, N., Read, T., & Konstantinidis, K. T. (2012). Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. *PloS one*, 7(2), e30087.

Olsen GJ, Woese CR, Overbeek R (1994) The winds of (evolutionary) change: Breathing new life into microbiology. *Journal of Bacteriology* 176: 1-6.

Page R (2007) Toward a taxonomically intelligent phylogenetic database. *Nature Precedings* DOI: 10.1038/npre.2007.1028.

Pagel M (1999) Inferring the historical patterns of biological evolution. *Nature* 401:877-884.

Parr CS, Guralnick R, Cellinese N, Page RD (2012) Evolutionary informatics: Unifying knowledge about the diversity of life. *Trends in Ecology and Evolution* 27:94-103.

Smith BE, Johnston MK, Lucking R (2016) From GenBank to GBIF: Phylogeny-based predictive niche modeling tests accuracy of taxonomic identifications in large occurrence data repositories. *PLoS One* 11:e0151232.

Tuama EO, Deck J, Droge G, Doring M and others (2012) Meeting report: Hackathon-workshop on darwin core and mixs standards alignment (February 2012). *Standards in Genomic Sciences* 7:166-170.

Van Brabant B, Gray T, Verslyppe B, Kyrpides N and others (2008) Laying the foundation for a genomic rosetta stone: Creating information hubs through the use of consensus identifiers. *OMICS* 12:123-127.

Yilmaz P, Kottmann R, Field D, Knight R and others (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology* 29:415-420.