



**BRAIN-be**  
(Belgian Research Action through Interdisciplinary Networks)

**AquaRES**  
**Aquatic species Register Exchange and Services**

**CONTRAT N° BR/132/A6/AQUARES**

**CONTRACT NR BR/132/A6/AQUARES**

**AQUARES**  
**Interoperabilité des Registres d'Espèces Aquatiques**  
**Mondiaux**

**AQUARES**  
**Uitwisseling van Data en Services tussen Aquatische**  
**Soortenregisters**

**ANNEXE I – SPECIFICATIONS TECHNIQUES**

**BIJLAGE I – TECHNISCHE SPECIFICATIES**

**ARTICLE 1: DESCRIPTION DU PROJET**

**1.1: Titre:** AQUARES - Interopérabilité des Registres d'Espèces Aquatiques Mondiaux (Aquatic species Register Exchange and Services)

**1.2: Description détaillée du PROJET**

**Summary**

The use of organism names is ubiquitous in a wide range of scientific, environmental management and policy domains. Specialist curated taxonomic databases and tools to query these data are therefore essential for ensuring the quality of biological data from collection and generation to data curation and management. Species information systems for monitoring status and trends of biodiversity and those dealing with policy concern – Natura 2000 species, commercial, invasive alien species and pest species – benefit from such high quality tools and databases ensuring the interoperability of the data.

The World Register of Marine Species (WoRMS), the Register of Antarctic Marine Species (RAMS) and the Freshwater Animal Diversity Assessment (FADA) database are three major Global Species Directories (GSD) hosted in Belgium. These data collections consist of authoritative taxonomic data, curated by international experts and contribute to initiatives such as Catalogue of Life (CoL) and LifeWatch, and potentially to the Pan-European Species directories Infrastructure (PESI) and other regional and national species lists. Most of these initiatives rely on a wide array of specialists' contributions to independent checklists and require extensive interactions with a wide expert network. Given the potential overlap in taxonomic specialists and the complex nature of the data, exchanging expertise and data among these initiatives is highly beneficial for all parties involved.

The main objective of this project is therefore to ensure and enhance the interoperability and public availability of these aquatic species databases through the development of a set of web services. Such services can guarantee the automatic and timely exchange of data between WoRMS, RAMS and FADA, but also expose the data for use in other initiatives and applications such as Encyclopedia of Life (EoL), Catalogue of Life (CoL), Global Biodiversity Information Facility (GBIF) and e-Science initiatives such as Biodiversity Virtual e-Laboratory (BioVeL) and LifeWatch.

To ensure the quality of the data exposed through those web services, we aim to improve the data import and exchange procedures into the partner databases and will develop a data entry interface to facilitate the entry of more complete distribution information. These procedures and tools will be tested and used during a hands-on workshop with taxonomic experts. To stimulate their involvement and advertise the free and open publication of their data, we will implement a tool for generating a checklist paper, which can be published in a scientific journal and provides more straightforward solution for properly citing and tracking citations of the data.

On top of the web services exposing the data from each of the databases, we will implement a joint data cache on which we can build a range of tools to perform data quality control of species related data. A central tool, which is relevant for both internal quality control of taxon data and for the curation of occurrence data or other species related data, is the so-called Taxamatch tool. This tool uses fuzzy matching algorithms for searching phonetically matching or very similar scientific species names and will be modified and improved for running on top of the joint data cache. Other tools, which are more specifically targeting species occurrence data include a service for mapping and validating occurrence data in comparison to expert checked distribution ranges and tools for checking technical errors in data files (e.g. incorrect date format, missing required fields). While these services will run as independent tools, they are highly relevant for a wide range of users and could, in the framework of the European Biodiversity Virtual e-Laboratory (BioVeL) project, be combined into a data workflow together with other web services available in the biodiversity informatics community.

Throughout this project, we will organise regular consultations with a wide range of potential users to document their requirements and get their feedback on the developed tools and services. Data from the FP7 BioFresh project, the European Ocean Biogeographic Information System (EurOBIS) and the Antarctic Biodiversity information Facility (AntaBIF) will be used as specific test cases to validate and improve the tools. Further tests with data from biological collections and ecological monitoring data are envisaged to ensure that these services are of interest to a wide range of institutes and researchers dealing with aquatic species data.

**Detailed description****1. Subject and objectives**

**Importance of scientific names.** While molecular methods have drastically changed the way biological research is conducted nowadays, the need for attributing a scientific name to a species is still universal in biological and ecological work. Scientific species names consist of a Latin binomial; the first part of the name identifies the 'genus' to which a species belongs and the second part, the 'specific epithet' or 'specific name' identifies the species within this genus, and the nomenclature is governed by various internationally agreed codes such as the International Code of Zoological Nomenclature (ICZN) for animals.

In addition to the use of species names in scientific and ecological management work, the importance of using correct organism names in the policy domain is not to be underestimated and is highly relevant for protected species, commercial, invasive alien and pest species.

Here, species information systems consisting of specialist curated taxonomic databases can be an important aid to ensure that different parties are targeting the same species when using a specific name. The possibility to resolve the correct species names from synonyms and homonyms, and to highlight potential typographic errors in species names is for instance important for supporting data generation and linking in species collection management, ecological/environmental monitoring initiatives and synthetic, large scale analysis of heterogeneous (biodiversity) data, in which species names are often the backbone to which all other data are attached. There is thus a need for efficient and (semi-)automated tools that will clean raw databases, will catalogue synonyms and wrongly spelled species names in the correct species category, and will weed out homonyms.

**Databases of aquatic species names.** The World Register of Marine Species (WoRMS), the Register of Antarctic Marine Species (RAMS) and the Freshwater Animal Diversity Assessment (FADA) database are three major Global Species Directories (GSD) hosted in Belgium. These data collections consist of authoritative taxonomic data, curated by international experts and are well embedded in an international context. Overall, Europe is a stronghold in terms of taxonomic expertise and databases, with several major (database) partners from Belgium involved in these initiatives including the Catalogue of Life and the Pan-European Species directories Infrastructure (PESI) and its component databases (European Register of Marine Species, Fauna Europaea, and Euro+Med PlantBase).

The World Register of Marine Species (WoRMS) and its web portal [marinespecies.org](http://marinespecies.org) grew out of the European Register of Marine Species (ERMS), and its combination with several other species registers maintained at the Flanders Marine Institute (VLIZ). Rather than building separate registers for all projects, and to make sure taxonomy used in these different projects is consistent, VLIZ developed a consolidated database called 'Aphia'.

One of the initiatives building on the Aphia system is the Register of Aquatic Marine Species (RAMS), which covers Antarctic species from the three realms of the Southern Ocean: the sea floor (meio-, macro- and megazoobenthos; micro- and macrophytobenthos), the water column (phytoplankton, zooplankton, nekton) and the sea-ice.

The Freshwater Animal Diversity Assessment (FADA) started as a Belspo funded initiative, with the aim to provide an overview of the specific and generic biodiversity in freshwater environments. FADA organized a targeted workshop with ca. 150 taxonomists and this resulted in a special volume of *Hydrobiologia* (vol. 595, 2008), documenting an almost complete overview of the diversity of around 65 aquatic higher taxonomic groups. In collaboration with the Belgian Biodiversity Platform, this also initiated the creation of a database with freshwater checklists, which is accessible through the [fada.biodiversity.be](http://fada.biodiversity.be) website. The FADA results have attracted global interest and the papers from the FADA volume are very widely cited (10 of the 20 most cited *Hydrobiologia* papers in 2008 are FADA papers). The FADA database gained renewed interest during the EU FP7 project BioFresh (Biodiversity of Freshwater Ecosystems: Status, Trends, Pressures, and Conservation Priorities), which uses it as taxonomic backbone for its data portal.

Similarly, ERMS acts as a taxonomic backbone for the European Ocean Biogeographic Information System (EurOBIS) and RAMS is integrated in the Scientific Committee on Antarctic Research-Marine Biodiversity Information Network (SCAR-MarBIN; [scarmarbin.be](http://scarmarbin.be)) and the AntaBIF ([biodiversity.aq](http://biodiversity.aq)) project. These initiatives are therefore obvious users of this project's outcomes.

The main objective of this project is to ensure and enhance the quality, interoperability and public availability of the aquatic species databases FADA, WoRMS and RAMS and to offer tools to a wide user community to access and actively use the data.

In order to achieve this, we will build on existing standards, tools and services and expertise within the network. Data exchange will be based on the Darwin Core archive standard and existing tools implemented on WoRMS such as Taxamatch (see Methodology for details) will be adapted to call other datasets such as FADA.

To address the quality of the existing databases, we will both enhance and extend the import, quality control and data publication process and improve the interoperability, ensuring an efficient and automated exchange of data through web services. Such web services for exchanging data among the participating databases and external initiatives will be set up for FADA and improved for WoRMS and RAMS.

In addition to the services for exposing the data for the individual databases, we also envisage a number of services and tools that are targeting on one hand the taxonomic experts and data managers of the participating databases and on the other hand external users working with species (occurrence) data. The latter includes for example tools to validate data during publication & curation and data refinement of existing data (e.g. for use in modelling work).

By building modular and complementary web services, we ensure the possibility for efficient re-use of both data and services. This includes data exchange with initiatives such as Encyclopedia of Life (EoL), Catalogue of Life (CoL) and the Global Biodiversity Information Facility (GBIF). In addition, several of these services could be combined with existing services from other providers in a Biodiversity Virtual e-Laboratory (BioVeL) workflow and in LifeWatch e-Science tools. As such, we aim to ensure the relevance of the

databases and envisaged tools for a wide user community. We will therefore engage stakeholders at various stages in the project and envisage tools that are of direct relevance to European and international initiatives such as EurOBIS, SCAR-MarBIN, AntaBIF and BioFresh for the curation of species occurrence data published to the Global Biodiversity Information Facility (GBIF) network.

The overall aim of this project is to facilitate access to scientific potential and data collections related to aquatic species managed at Belgian institutions including the Royal Belgian Institute of Natural Sciences. The developed tools and services will be optimally adapted for a wide range of users including the general public, scientists, biodiversity data publishers and aggregators, environmental managers and policy makers, providing access to these aquatic species data collections and ensure a maximal (re)use and exploitation.

In an international context, this project clearly supports the Convention on Biological Diversity (CBD) Aichi Target 19 on biodiversity knowledge. This target highlights the fact that effective conservation and management of biodiversity depends in large part on our understanding of taxonomy. Given its relevance to the developments in Biodiversity Virtual e-Laboratory (BioVeL) and LifeWatch, this project contributes to the e-Science developments in the EU. It is also complementary to the activities of the Belspo funded Belgian Biodiversity Platform in terms of data publication to the Global Biodiversity Information Facility (GBIF).

## ***2. Relevance to society***

Given the importance of species names for scientific, environmental management and policy domains as outlined above, the availability of high quality inventories of aquatic species has an important societal relevance. Aquatic environments face dramatic biodiversity losses, and improved understanding of the status and trends of biodiversity contributes to effective conservation and management in order to halting this decline. The aquatic species information systems included in this project are especially relevant for standardising information on species of policy concern, including Natura 2000 species, commercial, invasive alien species and pest species. Invasive species for instance are estimated to represent an estimated cost of 12 Billion euro to the EU and increased knowledge of these species supports a more effective management.

## ***3. Methodology***

FADA, WoRMS and RAMS rely on relational database systems. FADA uses a PostgreSQL database and the website providing access to the data ([fada.biodiversity.be](http://fada.biodiversity.be)) is built using the Ruby programming language and is hosted by the Belgian Biodiversity Platform. WoRMS and RAMS are part of the consolidated Aphia database running on SQL server at VLIZ. The websites providing access to these databases are built using PHP and are hosted by VLIZ ([marinespecies.org](http://marinespecies.org), [scarmarbin.be](http://scarmarbin.be)).

Web services provide a means for communication between machines over the web. Such services allow for instance data from one website/machine to be shown on another website without having to duplicate the data. This machine-to-machine interoperability is also a means to expose publicly available data for re-use in other initiatives and is highly suitable for exchanging data from taxonomic databases. For these web services, we envisage the use of widely accepted standards such as the REST and/or SOAP-compliant web services and provide extensive on-line documentation of the web API (application programming interface).

For data exchange, we will primarily use the Darwin Core standard from Biodiversity Information Standards (TDWG). The Darwin Core is a widely adopted, flexible standard for exchanging both taxon and occurrence related data.

One of the major tools that will be implemented and further improved is the Taxamatch tool for matching phonetically identical or nearly identical names. The 'TAXAMATCH' algorithms were originally developed by Tony Rees at CSIRO Marine and Atmospheric Research, Australia (<http://www.cmar.csiro.au/datacentre/taxamatch.htm>). The WoRMS Taxamatch tool uses the TAXAMATCH fuzzy matching algorithm by Tony Rees, and is complemented by the PHP/MySQL port of TAXAMATCH by Michael Giddens (<http://code.google.com/p/taxamatch-webservice/>) and the Scientific Names Parser (<https://github.com/GlobalNamesArchitecture/biodiversity>) by Dmitry Mozzherin. The web interface for entering distribution data and visually checking occurrence data will be build using open source JavaScript libraries such as OpenLayers to load, display and render maps.

An important means to improve the quality of the data collections, test the input tools and ensure that the developed tools meet the needs of the (potential) users we plan a number of meetings and workshops as outlined in the description of Tasks 2.5, 3.3 and 3.4. Both VLIZ and RBINS have experience with such meetings with taxonomic experts and consider these as an effective way to motivate the specialists involved to perform updates, complete any missing information and keep them engaged with the community.

The services developed during this project will be registered in the BiodiversityCatalogue <https://www.biodiversitycatalogue.org> and their documentation will be made available on-line under a CC-BY license. The source code for the different components is freely available on request.

## ***4. Complementarity and added value of the project with respect to international activities and initiatives (existing or in preparation) and***

### opportunities for new international collaboration

This project will result in a higher complementarity among the participating databases FADA, WoRMS and RAMS. Each of these databases is the de-facto standard and aggregator in its topic area and is (or will –as part of this project– be) contributing to the international initiatives Encyclopedia of Life (EoL), Catalogue of Life (CoL) and Global Biodiversity Information Facility (GBIF). All three taxonomic databases also act as taxonomic backbone for the respective thematic initiative mobilising occurrence data; BioFresh, (Eur)OBIS and SCAR-MarBIN and the developed tools will contribute to the quality assurance of the data published through these topical initiatives. The services and tools developed in this project will build on existing ones and will in turn contribute to the Biodiversity Virtual e-Laboratory (BioVeL) project and LifeWatch, leading to the wider dissemination and use of these tools.

### 5. Expected research results

The expected outcomes of this project include more complete, high quality databases and highly interoperable databases on freshwater, marine and Antarctic marine species, which will be made publicly available through its websites and exchanged with GBIF (Global Biodiversity Information Facility) and other international initiatives such as Encyclopedia of Life (EoL) and Catalogue of Life (CoL).

After this project, each of the databases will sport an improved input interface, quality procedures and data publication capabilities which will further contribute to the sustainability and international significance of the databases.

The tools developed on top of the combined databases providing the option to check species names, validate occurrence data and check data format(s) have a wide user community in mind and its functionalities will be chosen and evaluated in close collaboration with the follow-up committee.

#### ARTICLE 2: TACHES DU PROJET

Les tâches spécifiques du PROJET sont les suivantes :

#### ARTIKEL 2: PROJECTTAKEN

De specifieke taken van het PROJECT zijn de volgende:

### **1. Data exchange and web services to ensure the interoperability among the participating databases FADA, WoRMS and RAMS**

#### *1.1 Mapping the database structures and documenting requirements in terms of export functionalities, web services and central data cache [C: 1 pm, P2: 0.50 pm, P3: 0.25 pm]*

While the World Register of Marine Species (WoRMS) and the Register of Antarctic Marine Species (RAMS) rely on a common database system (Aphia), the Freshwater Animal Diversity Assessment (FADA) database is independent from these two and not entirely organised in the same manner. But, being taxonomic databases and making data available through the Darwin Core standard, these databases obviously share a wide range of similar features. As an initial task in this project, we will **map the database structure (1.1.1)** to ensure a common understanding of the envisaged content and underlying philosophy, select the required fields for data exchange and evaluate the feasibility for adopting a common schema. At this stage we will take the requirements from regional, national (3.4) and international (3.3) users and major international initiatives (4.1) into consideration when selecting the **requirements** in terms of the **web services and data cache (1.1.2)**. These requirements may also trigger the need for modifications to the database structure to ensure the interoperability between FADA, WoRMS and RAMS and with partner initiatives. Such changes could for instance include an improved mechanism for recording and storing the environment of the organisms and a timestamp indicating when the record and dataset was last updated.

#### *1.2 Improve and set up exchange web services [C: 2.25 pm, P2: 1 pm]*

An important means to ensure data synchronisation between FADA, WoRMS and RAMS for selected organism groups (see further; Task 1.4) is **setting up web services (1.2.1)** to exchange the data. While such web services are already in place for the Aphia database, the web services for the FADA database have to be designed and implemented to ensure this data exchange. Where necessary, the Aphia web services (<http://www.marinespecies.org/aphia.php?p=webservice>) will be improved and refined to match the requirements identified in Task 1.1. An important addition to these web services is for example the possibility to obtain data for non-marine environments only.

#### *1.3 Design and implement central cache for common web services [C: 0.5 pm, P2: 2 pm, P3: 0.25]*

Following the analysis under Task 1.1, we will **design (1.3.1)** and **build (1.3.2)** a **central data cache** linking the three databases FADA, WoRMS and RAMS. This data cache will be hosted at VLIZ and is primarily meant to act as an internal system for running common web services in terms of taxon matching and data cleaning & refinement (Task 3.1 and 3.2). Each of the databases will retain its import, update mechanism and quality control, which will be further refined as outlined in WP2. The central data cache will synchronise with the component dataset(s) using web services (Task 1.2).

#### *1.4 Synchronise updates for taxonomic groups managed in one of the participating databases [C: 1.25 pm, P2: 1 pm]*

For groups that are predominantly freshwater or marine, but contain a relatively small number of species relevant for other

environments (e.g. Rotifera, Macrophytes, Diptera, Collembola, Porifera, Isopoda), we envisage that they would preferably be managed in FADA or WoRMS respectively and data exchange for the relevant species will be organised and automated using the web services under 1.2. By setting up this exchange, we ensure that the data are managed in a more efficient way and that the thematic databases can act as more complete registers, which also include species groups with relatively few representatives in the marine or freshwater environment. This **data synchronisation** task (1.4.1) will encompass identifying gaps and overlaps in the organism groups covered in the different databases (which will feed in Task 2.5), data integration and updating. For organism groups with considerable overlap, and for which check lists exist in both FADA and WoRMS, we rely on expert input during the workshop as part of Task 2.5 to suggest the most efficient solution and resolve potential conflicts.

## 2. *Tools/services for improving taxonomic checklist data*

### 2.1 *Improving input and data publication services for individual datasets. [C: 3 pm, P2: 2 pm, P3: 0.25]*

Providing user-friendly input and quality control mechanisms for the contributing experts is the best approach to ensure the quality of taxonomic data in FADA, WoRMS and RAMS. The project partners have a wide experience in this respect and will exchange expertise in terms of working with Excel templates, modifying data through a web interface, validation of newly imported and updated checklists. To **improve** the procedure for **data import (2.1.1)** into the FADA database, we envisage an on-line tool for uploading and validating Excel-worksheets. During this process, technical issues, omissions and potential typos (based on comparison of names with existing databases) will be flagged, so they can immediately be corrected by the taxonomic editor.

The FADA website features a tool for generating species checklists from the imported data, which can be used as a data paper and can thus act as an incentive for the taxonomic editors to make their data openly available. Such a **checklist publication tool (2.1.2)** will also be implemented for WoRMS and RAMS and improved for FADA. This tool will be demonstrated and promoted during the workshop with taxonomic editors (Task 2.5).

### 2.2 *Perform data quality control and updating procedures [C: 4 pm, P2: 3 pm, P3: 2 pm]*

Similar to the procedures for importing new or updated data, we want to ensure the quality of the available data by applying rigorous quality control procedures on the databases. As part of this task we will document and improve tools and procedures for checking and annotating taxonomic lists.

As part of the scientific **data management (2.2.1)**, we will perform quality control by matching existing checklists (especially those covered during the workshop with taxonomic experts (Task 2.5)) with those available through other sources (see Task 2.3). This is especially relevant for taxonomic groups occurring in both FADA and WoRMS. Checklists and records will be annotated to document the results of these cross-checking exercises and potential conflicts will be flagged and reported to the experts for feedback.

In addition, we will explore the possibility to improve the data quality by cross-checking the entries with the literature e.g. through text mining of resources available on-line through the Biodiversity Heritage Library and BioStor initiatives.

### 2.3 *Taxamatch tools and implementation of fuzzy matching algorithms [C: 1.75 pm, P2: 1 pm]*

An indispensable component of the quality control procedures (including those mentioned under Task 2.1 and 2.2) is the possibility to match taxa with alternative (e.g. non-thematic) species directories. As changes in the gender of the 'Genus' name are often erroneously not reflected in the 'specific epithet' and long or complex taxonomic names are prone to typos, name matching requires not only exact matching, but also fuzzy matching algorithms to detect names that are very similar or match phonetically. A set of algorithms specific for Latin species names was worked out in the Taxamatch tool (see Methodology for details). As the implementation of such a tool is far from straightforward, but is already running on the Aphia database, the most efficient solution is to adapt this tool so it can run on top of the central data cache and can be called from the FADA, BioFresh and AntaBIF website.

Running the **Taxamatch tool (2.3.1)** as web service on the joint data cache would allow users to call either all or individual datasets. In addition, for species not found in FADA, WoRMS or RAMS, the search would be extended to other on-line sources including the Catalogue of Life (CoL), the Integrated Taxonomic Information System (ITIS), the Interim Register of Marine and Non-Marine Genera (IRMNG), Index Fungorum and FishBase (and each of these databases could also be called individually if desired). While a taxon name lookup is already in place for CoL and ITIS, as part of a LifeWatch tool developed at VLIZ, this will be extended with fuzzy matching capabilities. In addition to the FADA-database that will be included in the Taxamatch service, running the Taxamatch on the IRMNG database will also be a newly developed feature under this project.

In order to offer fuzzy match **search** options on the FADA, BioFresh and AntaBIF **websites (2.3.2)**, the possibility to call the Taxamatch web service will be added to the search functionalities. In addition, each of the database websites will promote this web service and tool.

### 2.4 *Develop interface for entering and validating distribution information for species [C: 3 pm, P2: 4 pm, P3: 0.75 pm]*

Both the WoRMS and FADA taxonomic experts are requested to document species distribution, respectively by assigning locality information and specifying faunistic regions. For WoRMS, this is currently done by selecting and linking one or more place names to a taxon. Each place name is linked to Marine Regions, a standard list of marine geo-referenced place names and areas, with a hierarchical structure.

Within the project, we will develop a more user friendly, visual interface to enter, validate and update this distribution information. Experts will be presented with a clickable map, where multiple relevant regions can easily be identified and saved in one action, thereby greatly simplifying the input effort of the editors. The input interface will take into account scaling: several standard levels of regions will be available, depending on the scale the editor wants to work on (e.g. global versus regional).

The first task will consist of the **selection** and testing of suitable **geographical regions** and define the precise requirements for this tool (**2.4.1**). For marine species, the interface will offer polygons of both the Exclusive Economic Zones (EEZ), and the International Hydrographic Organization (IHO) areas. For freshwater species we will, in addition to the polygons of the faunistic regions, include the HydroBASINS catchment delineation shape files, but would have to explore which grain (organised in hierarchical levels) is most appropriate. These HydroBASIN units are based on HydroSHEDS and were further refined and completed with support from the BioFresh project. For Antarctic species, potential geographical units include the areas defined by the Food and Agriculture Organization (FAO) and the Commission for Conservation of Antarctic Marine Living Resources (CCAMLR).

The **development of this visual data entry interface (2.4.2)** will be closely linked to the efforts on improving the data input procedures (Task 2.1). In addition to the possibility to assign regions to species through a clickable map, the taxonomic editors will be able to visualize the occurrences from either OBIS for WoRMS and from GBIF/BioFresh for FADA. Through this option, editors will be able to spot and flag possible errors in the occurrence databases and notify the data publishers or providers. As data providers to OBIS and BioFresh will, as outlined under Task 3.2, be able to plot the expert-validated distribution ranges together with their own data, this could aid the detection of possible distribution gaps in WoRMS, RAMS and FADA. The combination of these services and their use by several involved parties (editors and users) will provide a mutual benefit as this interaction will enable data holders to improve the quality of both taxonomic and occurrence datasets.

### 2.5 *Workshop to test tools and procedures for checking taxonomic lists, tools to complete distribution data and address identified gaps and overlaps between datasets [C: 2 pm, P2: 1 pm, P3: 3 pm]*

Based on the gaps and overlaps in the taxonomic datasets identified during the synchronisation of the participating datasets (Task 1.4) we will select a number of focal organism groups for which we will invite taxonomic experts. Where relevant, we will identify alternative sources of data and experts that could increase the coverage of the component datasets (e.g. for phytoplankton and other algal groups). We have preliminarily identified Crustaceans such as Decapoda, Branchiura, Copepoda and Cladocera as relevant groups, which are represented in FADA, WoRMS and RAMS. The aim of the workshop is double; first we want to demonstrate and test the data entry and publication tools and engage the editors in completing distribution information at the most appropriate grain (resolution) and secondly, we want to take the opportunity to resolve conflicts for overlapping groups (which will also guide the selection of taxonomic experts).

Convening a meeting with taxonomic experts is an effective way to motivate to perform updates, complete any missing information and keep them engaged with the community of taxonomic editors. During this meeting we will offer hands-on (technical) support for the testing the input and data publication tools by the experts. On issues of taxonomy and species range, such as whether related taxa in different environments or continents are the same species or not, workshop discussion among experts will help reach best-possible solutions on the basis of current knowledge. The data management support under Task 2.2 will allow us to actively support the data holders in preparing their data and information in the time leading up to the workshop and ensure a maximal output by actively following on their input afterwards.

We plan to have a 3 to 4 day workshop with up to 20 participants. As these experts do not receive any remuneration for their contribution, we envisage covering at least part of their travel and accommodation expenses.

## 3. *Tools/services to validate species distribution/occurrence data*

### 3.1 *Produce and improve services to perform quality control on species occurrence data [C: 0.5 pm, P2: 1 pm, P3: 0.25]*

Next to the importance of developing tools and services for the internal and expert-based quality control, data improvement and refinement, it is equally important to meet the needs of our users. The services that will be developed and refined under Work package 2 will – directly or indirectly – have their relevance towards the user communities, and this on two levels. First of all, these services will help users in the preparation of their data for publication. Secondly, they can be of help in applications requiring data refinement, specifically when dealing with heterogeneous data from various sources that will be used in large-scale analyses.

Initially, the currently available functionalities of the services available in WoRMS and Lifewatch will be **reviewed** and compared to the **quality control procedures** and recommendations from the BioFresh project (**3.1.1**). Where relevant, such recommendations will feed into the process of improving the ‘*General quality control and data format checking services*’ (below; C), possibly with feature requests specific for freshwater datasets.

Following this review, the developed tools under Work package 2 will - where relevant - be further **improved** and **integrated** as easily accessible **web services (3.1.2)**. The different web services will be available on the LifeWatch portal ([www.lifewatch.be](http://www.lifewatch.be)) and be linked to and advertised through the websites of the partner databases. Evidently, these web services will significantly contribute to the European LifeWatch project and are highly relevant in the context of the BioVel tools (in particular the data refinement workflow).

The offered services will cover the following items:

*A. Taxon match services*

This service will allow users to match their taxonomic list to available online standards, including FADA, WoRMS, RAMS, the Catalogue of Life (CoL), the Integrated Taxonomic Information System (ITIS), the Interim Register of Marine and Non-Marine Genera (IRMNG), Index Fungorum and FishBase. As specified under Task 2.3, users will have the option to search all of the listed taxonomic standards or just a selection. Matching species lists to accepted standards helps to improve the overall quality and greatly enhances the operability of occurrence data. Next to individual users (scientists, students...), these tools will also be of use to large, European initiatives and data systems, such as e.g. PESI, the Pan-European Species directories Infrastructure.

*B. Occurrence checking services*

This web service will enable users to plot their sampling locations on a map for running a quick visual quality check. Through this service, the user will be able to detect possible errors in the coordinates (such as the switching of latitude and longitude or the lack of a minus sign to indicate West or South). This kind of flaws can easily be fixed by the user, again improving the overall quality of the data. Taking this further, users will also be able to compare their occurrences with the documented distributions in the taxonomic databases (WoRMS, RAMS, FADA). This service is described in detail in Task 3.2.

*C. General quality control and data format checking services*

These services include e.g. mapping of the uploaded field names with a standard set of fields, highlighting non-matches or missing required fields, and checking of the data format of e.g. the date-related fields. These quality control steps are primarily targeting data providers to allow them to easily check the format and content of their data before submission, thereby mitigating the tasks of the data manager. Currently, such quality control services are specifically being developed for data that will contribute to (Eur)OBIS.

Within this project, the different data formats used for WoRMS, FADA, SCAR-MarBIN, AntaBIF and BioFresh will be compared and where possible mapped to a common standard (e.g. Darwin Core) in order to build more generic web services for checking the quality and format of these data.

**3.2 Build tool for visually checking and validating occurrence data based on available distribution information [C: 0.5 pm, P2: 2.5 pm, P3: 0.5 pm]**

The simplified input tool for distributions in WoRMS, RAMS and FADA (Task 2.4) will stimulate the taxonomic experts to add the known distribution range of species to these data systems and make the currently available distribution information even more complete.

This information will be exposed to the users by **building a tool** for visually checking occurrence data (**3.2.1**). Through this tool, users will be able to upload their species occurrence data and compare them to the documented and expert-validated distribution range. A quick visual inspection and checking of the assigned quality flags will allow validating or questioning the accuracy of the occurrence data. Double-checking possible anomalies will lead to an improved quality of the overall content of the dataset or it can indicate a gap in FADA, WoRMS or RAMS as appropriate. In the latter case, the identified gap can be communicated with the relevant experts, who will evaluate this notification and take action when necessary.

This tool will not only prove its value to individual users in checking and validating their data, but it will also be a major contribution to the validation of the occurrence records stored in large biogeographic data systems such as the (European) Ocean Biogeographic Information System ((Eur)OBIS) and the Global Biodiversity Information Facility (GBIF).

**3.3 Validate tools during workshops with users: BioFresh, EuroBIS, AntaBIF [C: 1.5 pm, P2: 0.5 pm, P3: 1 pm]**

Once these tools are in place, they will be put to the test with data from the user community during **hands-on workshops (3.3.1)**. In contrast to the workshop planned under Task 2.5, these will be smaller one-day workshops. We plan one joint workshop with participants from the European Ocean Biogeographic Information System (EurOBIS), the Antarctic Biodiversity Information Facility (AntaBIF) and the Biodiversity of Freshwater Ecosystems consortium (BioFresh) and several ad-hoc workshops with individual users. With the feedback from the user community, the tools and services can be refined to better meet their needs.

**3.4 Consultations with Follow-up Committee and its members [C: 1.5 pm, P2: 0.5, P3: 0.5]**

Throughout this project, we envisage **regular consultations** with the Follow-up Committee and its members (**3.4.1**). At the start of the project we will organise a consultation with potential users of the data and the tools that will be improved and developed within this project. In addition to the international users (BioFresh, EurOBIS and AntaBIF), we also aim to



specifically target Belgian users (including the Research Institute for Nature and Forest-INBO and Département de l'Etude du milieu naturel et agricole-DEMNA) to ensure the relevance of the data and services in a national and regional context. During this consultation, we will document the user requirements in terms of databases and services. Once the tools under Work package 3 are available, we will demonstrate the tools and document how - for instance - specific regional requirements could be taken into account during further developments of the component databases and joint services.

#### 4. Data exchange with international initiatives

##### 4.1 Establishing and improving data exchange with international initiatives [C: 1, P2: 0.5, P3: 0.5]

In an early phase, the existing and planned **web services** within this project will be **mapped** with the known requirements of major (international) initiatives such as Encyclopedia of Life (EoL), Catalogue of Life (CoL), Global Biodiversity Information Facility (GBIF), Biodiversity Virtual e-Laboratory (BioVeL) and LifeWatch (**4.1.1**). This mapping will allow thorough assessment and comparison of what is planned and what is needed within the international community, leading to the most workable and widely employable services.

Both WoRMS (including RAMS) and FADA partly share their content with CoL and GBIF. WoRMS also provides content to EoL, BioVeL and LifeWatch and we plan to set up similar exchanges for the FADA database within this project.

In parallel to **streamlining** this data **exchange (4.1.2)**, the range of tools and services developed during this project will result in an increased quality of the content we share. This will result in greater visibility and credibility to the component databases and the initiatives to which they contribute at the international level.

#### ARTICLE 3: CALENDRIER DES TACHES DU PROJET

3.1: Le commencement et l'achèvement des tâches décrites à l'article 2 de la présente annexe correspondent respectivement au DEBUT OPERATIONNEL et au TERME OPERATIONNEL.

3.2: Les délais d'exécution des tâches sont les suivants :

#### ARTIKEL 3: TIJDSHEMA VAN DE PROJECTTAKEN

3.1: De aanvang en het einde van de taken omschreven in artikel 2 van deze bijlage, stemmen respectievelijk overeen met de AANVANG DER WERKZAAMHEDEN en de BEEINDIGING DER WERKZAAMHEDEN.

3.2: De uitvoeringstermijnen van de taken zijn de volgende:

	Semester	Year 1		Year 2		Year 3		Year 4		Total months <sup>(1)</sup>	Man-
		1	2	1	2	1	2	1	2		
Work Package 1: Data exchange and web services to ensure the interoperability among the participating databases FADA, WoRMS and RAMS											
Task 1.1 Mapping the database structures and documenting requirements in terms of export functionalities, web services and central data cache											
Task 1.1.1: Map database structure	C									0,5	
	P2									0,25	
	P3										
Task 1.1.2: Requirements web services and data cache	C									0,5	
	P2									0,25	
	P3									0,25	
Task 1.2 Improve and set up exchange web services											
Task 1.2.1: Setting up web services	C									2,25	
	P2									1	
	P3										
Task 1.3 Design and implement central cache for common web services											
Task 1.3.1: Design central data cache	C									0,5	
	P2									0,5	
	P3									0,25	
Task 1.3.2: Build central data cache	C										
	P2									1,5	
	P3										
Task 1.4 Synchronise updates for taxonomic groups managed in one of the participating database											
Task 1.4.1: Data synchronisation	C									1,25	
	P2									1	
	P3										

		Year 1		Year 2		Year 3		Year 4		Total months <sup>(1)</sup>	Man-
Semester		1	2	1	2	1	2	1	2		
Work Package 2: Tools/services for improving taxonomic checklist data											
Task 2.1 Improving input and data publication services for individual datasets											
Task 2.1.1: Improve data import	C										2,75
	P2										0,25
	P3										
Task 2.1.2: Checklist publication tool	C										0,25
	P2										1,75
	P3										0,25
Task 2.2 Perform data quality control and updating procedures											
Task 2.2.1: Data management	C										4
	P2										3
	P3										2
Task 2.3 Taxamatch tools and implementation of fuzzy matching algorithms											
Task 2.3.1: Taxamatch tool	C										
	P2										0,75
	P3										
Task 2.3.2: Search on website	C										1,75
	P2										0,25
	P3										
Task 2.4 Develop interface for entering and validating distribution information for species											
Task 2.4.1: Selection of geographical regions	C										0,75
	P2										0,25
	P3										0,5
Task 2.4.2: Development of data entry interface	C										2,25
	P2										3,75
	P3										0,25
Task 2.5 Workshop to test tools and procedures for checking taxonomic lists, tools to complete distribution data and address identified gaps and overlaps between datasets											
Task 2.5.1: Workshop	C										2
	P2										1
	P3										3

		Year 1		Year 2		Year 3		Year 4		Total months <sup>(1)</sup>	Man-
Semester		1	2	1	2	1	2	1	2		
Work Package 3: Tools/services to validate species distribution/occurrence data											
Task 3.1 Produce and improve services to perform quality control on species occurrence data											
Task 3.1.1: Review quality control procedures	C										0,5
	P2										0,25
	P3										0,25
Task 3.1.2: Integrate and improve web services	C										
	P2										0,75
	P3										
Task 3.2 Build tool for visually checking and validating occurrence data based on available distribution information											
Task 3.2.1: Building tool	C										0,5
	P2										2,5
	P3										0,5
Task 3.3 Validate tools during workshops with users: BioFresh, EuroBIS, AntaBIF											
Task 3.3.1: Hands-on workshops	C										1,5
	P2										0,5
	P3										1
Task 3.4 Consultations with steering committee and its members											
Task 3.4.1:	C										1,5

Regular consultations during design and testing	P2									0,5
	P3									0,5

	Semester	Year 1		Year 2		Year 3		Year 4		Total Man-months <sup>(1)</sup>
		1	2	1	2	1	2	1	2	
Work Package 4: Data exchange with international initiatives										
Task 4.1 Establishing and improving data exchange with international initiatives										
Task 4.1.1: Mapping web services	C									0,5
	P2									0,25
	P3									0,25
Task 4.1.2: Applying improvements	C									0,5
	P2									0,25
	P3									0,25

Total Man-months <sup>(1)</sup>	
C	24,25
P2	20,5
P3	9,25

<sup>(1)</sup> Only for persons for whom funding is requested