AQUA EXCEL 2020

AQUAculture infrastructures for EXCELlence in European fish research towards 2020 — AQUAEXCEL2020

# D3.4 Usable ELIXIR services, standards for data exchange and data modelling

*USB*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 652831. This publication reflects only the view of the author, and the European Commission cannot be held responsible for any use which may be made of the information contained therein

# Executive Summary


**Objectives**

The objective was to introduce the ELIXIR, its services, databases and standards to the AQUAEXCEL2020 member to support them in the standardization, data exchange, data processing and data archiving activities.

**Rationale:**

The current boom of the new experimental devices and methods in the natural sciences generates exponentially increasing volume of experimental data. The data need to be stored organized, standardized, processed and archived. A lot of community-developed standards, inhouse databases and data processing tool exist. It was mainly developed based on the need of the research activities and used by particular researchers. To not waste the money to develop all the tools again and to enable free access to the already measured and processed data, some initiative was needed. Bioinformaticians and life science researchers in both academic and industrial settings need open access to technically and scientifically excellent data resources for effective data discovery, deposition, and re-use. They also need confidence in the sound governance, life cycle management, and long-term sustainability of those data resources.
One of the initiatives providing this kind of resources is ELIXIR organization which brings together life science resources from across Europe. Many researchers have heart about ELIXIR but just few of them are using the resources available from different reasons.

**Main Results:**

The result is the document describing selected ELIXIR services and the process of relevant service identification and implementation of the results.
It briefly describes the ELIXIR organization, the structure and what does it do. The list of services, databases, tools and standards relevant to the AQUAEXCEL2020 consortium is provided together with the links to them. The process of identifying the best services matching user needs is described as the EMBRIC use case together with example of AQUAEXCEL2020 use case.


**Authors/Teams involved:** Dr. Petr Císař (USB)

# Table of Contents

# 1. ELIXIR

ELIXIR is an intergovernmental organisation that brings together life science resources from across Europe. These resources include databases, software tools, training materials, cloud storage and supercomputers. The goal of ELIXIR is to coordinate these resources so that they form a single infrastructure. This infrastructure makes it easier for scientists to find and share data, exchange expertise, and agree on best practices.

ELIXIR includes 21 members and over 180 research organisations. It was founded in 2014, and is currently implementing its first five-year scientific programme.

ELIXIR is organised using a 'Hub and Nodes' model, where the Hub's role is to coordinate the work done by the Nodes.

ELIXIR uses the resources at all levels of the organizational structure. The main focus is on the usage of the already build national infrastructures and resources that are specialized and can be accessed by other researchers. The infrastructures (research institutes) usually provide the tools for data processing and analysis or the devices for experimental data collection. The national resources are mainly used in the form of national data centres.
At the level of international resources, the consultation services or the specialized databases are available for the users.

# 2. ELIXIR services

The model of the usage of the ELIXIR services is based on the open access to the provided infrastructure, to the data processing and analysis tool or to the data storage (databases). Usually the access to the infrastructure (measurement devices) is charged to the user. There are several possibilities of the free or discounted access using the Transnational access (small research projects – equivalent of the TNA projects in AQUAEXCEL2020) or the national Large infrastructures projects. The different ELIXIR services are show on the Fig. 1.

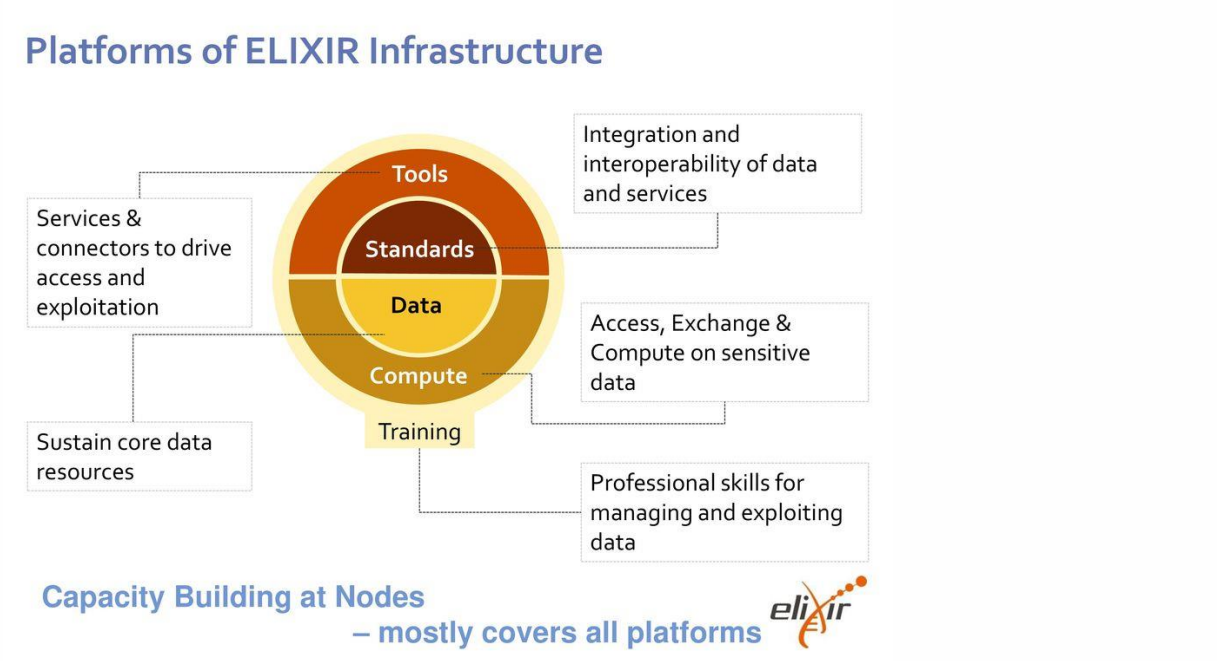


**Figure 1 Source: Elixir capacity building – ISMB/ECCB Prague,2017. J.Vondrášek, B. Persson, B. Leskošek**

The services are divided into five main groups: Tools, Standards, Data, Compute and Training. The training is dedicated to the how to use the other services provided by ELIXIR and is not described in this document. More information about the training can be found here: https://www.elixir-europe.org/services/training

## *The tools platform*

The tools is a collection of the software tools to access, study and compare the data. The tools are listed in the ELIXIR Tools Platform which serves to improve the discovery, quality and sustainability of software resources. The platform helps life scientists find, deploy and benchmark software tools, including workflows. It also helps software providers and developers better describe and develop software tools and integrate them into workflows.
The tools are accessible through the webpage: https://www.elixir-europe.org/services/tools

Where the tool can be sorted by nodes. There are three main groups of tools:

- **bio.tools**: software tools and data resources for life sciences.
- **Biocontainers:** software you can run on any operating system.
- **OpenEBench:** tools for benchmarking and monitoring software.

Selected sublist of tools relevant to AQUAEXCEL2020 project:

| Name of service | Description | ELIXIR Node |
|---|---|---|
| CSC Chipster | Chipster is a user-friendly analysis software for high-throughput data. It contains hundreds of analysis tools for next generation sequencing (NGS), microarray, proteomics and sequence data. | ELIXIR Finland |
| g:Profiler | g:Profiler is a bioinformatics toolkit for characterising and manipulating gene lists of high-throughput genomic data. | ELIXIR Estonia |
| ISMARA | Models genome-wide expression data in terms of genome-wide annotations of regulatory sites. For a given expression data-set it infers the key transcription regulators, their sample- dependent activities, and their genome-wide targets. | ELIXIR Switzerland |
| META-pipe | The META-pipe pipeline for annotation and analysis of marine metagenomics samples provides insight into phylogenetic diversity and metabolic and functional potential of environmental communities. | ELIXIR Norway |

## *The Compute Platform*

The ELIXIR Compute Platform was established in 2015 to build and integrate cloud, compute, storage and access services for the life-science research community.

Today, thousands of science laboratories across the world generate massive amounts of data that they make available to collaborators directly or place in public archives for open access. In this situation, the traditional method of a researcher downloading, and analysing data locally is no longer viable due to both the data size and scope of the analysis activities. The data needs to be managed as a federation, where data providers work as a single infrastructure providing mechanisms where researchers can bring their analysis to where the data is located. The ELIXIR Compute Platform infrastructure will allow life scientists to easily access, share and analyse data from different sources across Europe.

The objective is to combine all components of the ELIXIR Compute services into a seamless workflow. A researcher may use the ELIXIR Authorisation and Authentication services to securely create a scientific software analysis environment, and use the environment to access large biological data resources stored in a cloud.

The tools are accessible through the webpage:
https://www.elixir-europe.org/services/compute
Where the tool can be sorted by nodes.
Selected sublist of platforms relevant to AQUAEXCEL2020 project:

| Name of service | Description | ELIXIR Node |
|---|---|---|
| Computerome | Computerome is the Danish National Supercomputer for Life Sciences. It serves all life science research groups within Denmark and is also open for international collaboration. | ELIXIR Denmark |
| French Academic Cloud | IFB is setting up a High Throughput Computing infrastructure focusing on Life Science data sharing, analysis and integration that is complementary to the national High Performance Computing centres that mainly emphasize computing. | ELIXIR France |
| sciCORE | sciCORE provides a high-performance computing infrastructure, large-scale storage resources, scientific software and databases, server infrastructures and user support. It also provides expertise to scientific research groups. | ELIXIR Switzerland |
| Vital-IT | Vital-IT supports and collaborates with life scientists in Switzerland and beyond. It provides expertise, training and maintains a high-performance computing (HPC) and storage infrastructure. | ELIXIR Switzerland |

## *The Data core resources*

The goal of the ELIXIR Data platform is to drive the use, re-use and value of life science data. It aims to do this by providing users with robust, long-term sustainable data resources within a coordinated, scalable and connected data ecosystem.
The ELIXIR Data platform promotes Open Access as a core principle for publicly funded research. ELIXIR resources ideally reflect this commitment and have terms of use or a licence that enables the reuse and remixing of data.

Services offered:

- Core Data Resources: European data resources that are of fundamental importance to research in the life sciences and are committed to the long-term preservation of data.
- ELIXIR Deposition Databases: repositories recommended for the deposition of life sciences experimental data.
- Database services listing: this list is updated as Nodes finalise or review their Service Delivery Plans.

To achieve its goals the Platform works in four groups.

- **Core Data Resources and Deposition Databases**
  - Goal: To administer and support the Core Data Resource and Deposition Database portfolio.

- **Literature-Data Integration**
  - Goal: To build a comprehensive, connected data ecosystem across ELIXIR, with deep integration to the scientific literature via the ELIXIR Core Data Resource, Europe PMC.

- **Scalable Curation**
  - Goal: To maximise the ability of expert human curators to enrich the ELIXIR knowledgebases through providing trans-resource, scalable curation solutions.

- **Long Term Sustainability**
  - Goal: To ensure the long-term financial sustainability of the ELIXIR Core Data Resources by contributing to the establishment of a global, internationally shared, sustainable funding model for Core Data Resources.

The two main services within the data platform are: Data Resources and Deposition databases.

**ELIXIR Core Data Resources** are a set of European data resources of fundamental importance to the wider life-science community and the long-term preservation of biological data. Identification of the ELIXIR Core Data Resources involves a careful evaluation of the multiple facets of the data resources.

The details of the selection criteria are described in the F1000R ELIXIR track article 'Identifying ELIXIR Core Data Resources'. The initial Core Data Resource list was defined in July of 2017. The list will be reviewed regularly - further rounds of selection are planned, going forward.

ELIXIR is committed to Open Access as a core principle for publicly funded research. ELIXIR Core Data Resources should reflect this commitment and have terms of use or a licence that enables the reuse and remixing of data. The Creative Commons licenses CC0, CC-BY or CC-BY-SA are all conformant with the Open Definition (http://opendefinition.org/licenses/), as are equivalent open terms of use.

Selected sublist of data resources relevant to AQUAEXCEL2020 project:

| Core Data Resource | Data type |
|---|---|
| ArrayExpress | Functional Genomics Data from high-throughput functional genomics experiments. |
| Ensembl | Genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. |
| Ensembl Genomes | Comparative analysis, data mining and visualisation for the genomes of non-vertebrate species. |
| PRIDE | Mass spectrometry-based proteomics data, including peptide and protein expression information (identifications and quantification values) and the supporting mass spectra evidence. |

**ELIXIR Deposition Databases:**

The purpose of this Deposition Databases list is to provide guidance to those who formulate policy and working practices about the appropriate repositories for publishing open data in the life sciences. An ELIXIR Deposition Database is defined as being part of the ELIXIR Node portfolio of services that accepts deposition of experimental data from an international community of researchers beyond the funding envelope of the database itself. The ELIXIR Deposition Databases meet the technical quality and governance criteria expected of ELIXIR Core Data Resources (see the F1000R article “Identifying ELIXIR Core Data Resources”), which align with the FAIR principles, but may be at an earlier stage of development, meeting an emerging scientific requirement, or maybe narrower in scope. Consequently some, but not all, of the ELIXIR Deposition Databases also appear in the ELIXIR Core Data Resources list.

Selected sublist of deposition databases relevant to AQUAEXCEL2020 project:

| Deposition Database | Data type | International collaboration framework |
|---|---|---|
| BioModels | Computational models of biological processes. | |
| BioSamples | BioSamples stores and supplies descriptions and metadata about biological samples used in research and development by academia and industry. | NCBI BioSamples database |
| BioStudies | Descriptions of biological studies, links to data from these studies in other databases, as well as data that do not fit in the structured archives. | |
| EMDB | The Electron Microscopy Data Bank is a public repository for electron microscopy density maps of macromolecular complexes and subcellular structures. | |

All the databases provide standardized data connection interface. To deposit the data, the user contacts the database through the web form and the data curator is assigned to the user. The data curators review the dataset which should be stored and decide about the possibility to store into the selected database. The data and metadata are checked by data curator for the completeness and consistency and stored in the database for open access to the data.

# 3. Standards

Life science research produces data in a variety of formats. The diversity of formats means that scientists cannot easily find and compare datasets from different sources, so their ability to make new discoveries is hampered.

The standards for the data exchange in ELIXIR are concentrated in Interoperability Platform. The platform created and updates the list of standards and also defined the workgroups for the identification and promotion interoperability best practices for data providers and data integrators and delivery the interoperability services that underpin ELIXIR Communities and Platforms.

The goal of the Interoperability Platform is to help people and machines to discover, access, integrate and analyse biological data. It encourages the life science community to adopt standardised file formats, metadata, vocabularies and identifiers.

The Platform works both within Europe and globally. For example, through organisations such as Research Data Alliance (RDA), which promotes data sharing and exchange around the world.

The platform offers three main services:

- FAIRsharing: a curated resource on data and metadata standards, inter-related to databases and policies.
- Identifiers.org: a resolving system to reference data in both a location-independent and resource-dependent manner.
- Ontology Lookup Service (OLS): a portal for biomedical ontologies, providing access to the latest ontology versions.

Selected sublist of standards relevant to AQUAEXCEL2020 project:

| Name of service | Description | ELIXIR Node |
|---|---|---|
| FAIRsharing | An information and educational resource of inter-related data standards, databases and policies. | ELIXIR UK |
| ISA Tools Commons | A suite of open source tools and formats to enable standards-compliant collection, curation, management, publication and reuse of experiments. | ELIXIR UK |

| Name of service | Description | ELIXIR Node |
|---|---|---|
| The Ontology Lookup Service (OLS) | The Ontology Lookup Service (OLS) provides a web service interface to query multiple ontologies from a single location with a unified output format. | EMBL-EBI |

# 4. Use case - EMBRIC

The list of the services provided by ELIXIR is large and it can be complicated to find the best solution for data repository or the tool for data processing. One of the ways how to orient in the list of services is to participate on the online training/standard training events organized by ELIXIR platforms. The events are listed here: https://www.elixir-europe.org/events and are organized by the Hub, Nodes or the institutes belonging to the national node.

Another way of the support provided to orient in the ELIXIR services and to find the best solution is to ask to use EMBRIC configurator. EMBRIC, to which AQUAEXCEL2020 partners through INRA, CCMAR and HCMR is the European Marine Biological Research Infrastructure Cluster (EMBRIC) designed to accelerate the pace of scientific discovery and innovation from marine Bio-Resources. EMBRIC aims to promote new applications derived from marine organisms in fields such as drug discovery, novel foods and food ingredients, aquaculture selective breeding, bioremediation, cosmetics and bioenergy.
One of the services launched within the EMBRIC project is the EMBRIC configurator. It is an consultancy service which helps to fins specific solution for the user defined problem in the area of data storage, management or processing. The service provides an entry point into the existing informatics data resources, especially around molecular biology. Targeting those embarking on the design of new marine projects, clients of the service provide a scientific description into an online form and, following several rounds of structured discussion and review involving input from informatics and scientific experts, are provided with a project-specific ‘configuration’. This configuration includes a description of the elements of infrastructure (such as databases, standards, formats, curation groups, analysis methods and cloud compute capacity), advice on accessing and setting these elements up for the project and data management guidelines.
Support is provided to those operating selected EMBRIC use cases in the set up and operation of configurations of infrastructure. Support draws on the recommendations from the EMBRIC Configurator service and may involve, inter alia, standards development, curation, data coordination and helpdesk services. Such support is offered to a subset of users of the EMBRIC Configurator service, where the case in question has been accepted as a full EMBRIC use case, referred to hereon as “EMBRIC UC”.

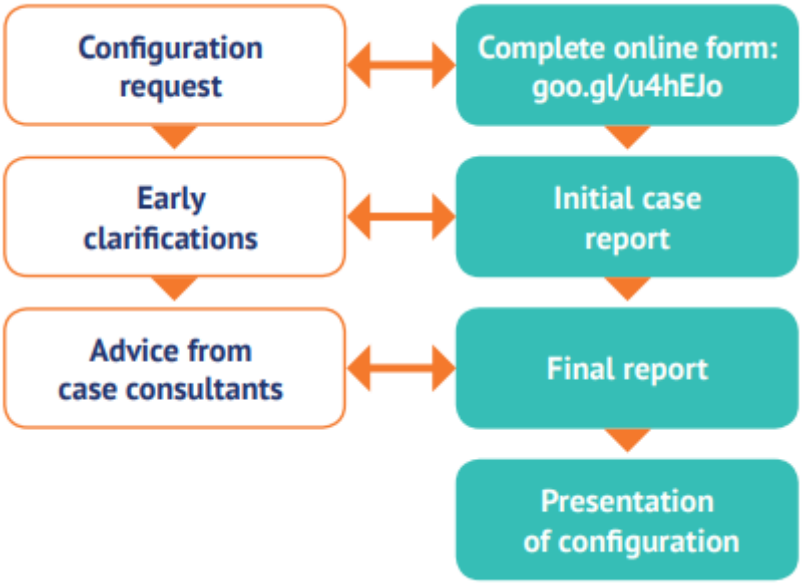


**Figure 2. EMBRIC configurator workflow. Source:**

**http://www.embric.eu/sites/default/files/EMBRIC_Configurator_A4_print2.pdf**

To use the configurator service, the user has to first describe the case study where the overall information about the data, processing and storage are provided, see Fig. 2. The case officer and the experts review the case and ask for the additional questions to identify the best possible solution. The outcome of the service is the report providing the information about the recommended tools for data processing, guidance on data management together with recommended data storage databases (ELIXIR data storage) and other case specific advices.

The configurator was used for the three use cases of AQUAEXCEL2020 project of WP3: FishandChips Tool, Digital fish and AQUAEXCEL repository. The sustainability aspects of all three tools was considered in the case studies. The example of the AQUAEXCEL repository use case is provided in annex n. 1: EMBRIC_CCD_AQUAEXCEL2_final.pdf

**Example of AQUAEXCEL repository as EMBRIC use case:**

At the first step there was the idea to use ELIXIR repositories for the long-term data archiving contained in the AQUAEXCEL repository. The motivation was to sustain the data after the project finish. The data storages established and operated during the active period of the project are great for the project partners collaboration, but the operation is usually limited after the project end (it is not the case of AQUAEXCEL repository because it will be operated after the end of the project by the University of South Bohemia). The data archiving is resources demanding and can be done through the ELIXIR services.
The first discussion about the possibility of the data archiving was made at the AQUAEXCEL2020 – ELIXIR EBI meeting. It was decided that the EMBRIC configurator can be used for the identification of the best possible solution within ELIXIR portfolio.
As it was mentioned EMBRIC configurator is the consultancy service which deals with the individual request, use the questionnaire and the interview with the user to describe the needs and to recommend specific solution.

First the request was formalized:
*One of the main goals of AQUAEXCEL is to have a long-term solution for the exchange, archive and access of data, metadata and experiments related to isogenic fish lines. Isogenic fish lines are important and widespread experimental material to study fish sensitivity to*

*changes in diet or environmental conditions. Therefore, it is crucial that there is a platform for the long-term management and sharing of standardized data and metadata on these lines among not only the AQUAEXCEL[2020] members but also the broader international community. At the moment there are two types of AQUAEXCEL[2020] repositories planned: 1) The "Digital Fish" repository, which is for data and metadata on isogenic fish lines (development in discussion); 2) the "AQUAEXCEL[2020]" repository, which is for storage, sharing and visualisation of protocols, metadata and data generated by the AQUAEXCEL[2020] consortium and the Transnational Access (TNA) projects funded by AQUAEXCEL[2020] (developed). These repositories will be sustained for the life time of the project and there is a need for more long-term storage that goes beyond that.*

The request was filled into the EMBRIC configurator form (see Fig.3): https://docs.google.com/forms/d/1NYdbiZV_DxyyoahaWgmyXvVq54WpRVAJBFOjaFW0R9U/viewform?edit_requested=true
The form is the entry point for the EMBRIC configurator use case.

Based on the description the response from the case officer was obtained:
*In order to ensure sustainability of the data generated by AQUAEXCEL[2020] we will: 1) collect information on the AQUAEXCEL[2020] repository and Digital Fish; 2) collect information on data types supported in these systems); 3) identify ELIXIR services appropriate for the long-term sustainability*
*of data and metadata from isogenic fish lines; and 4) explore the possibility of linking the AQUAEXCEL[2020] repositories with ELIXIR services.*
*This case study will take advantage of the EMBRIC Configurator service, to provide an appropriate description of data sustainability infrastructure for AQUAEXCEL[2020], and will serve as an EMBRIC UC, for which we will offer support in the establishment of the described infrastructure.*

In the next step the response report describing the description of the problem, the proposed solution and the identified ELIXIR services was delivered. You can find the full text of the report as annex 1. of this document below.

Finally, the teleconference was organized to clarify the aspects of the report. The user can ask any question regarding the proposed solutions and ELIXIR services.

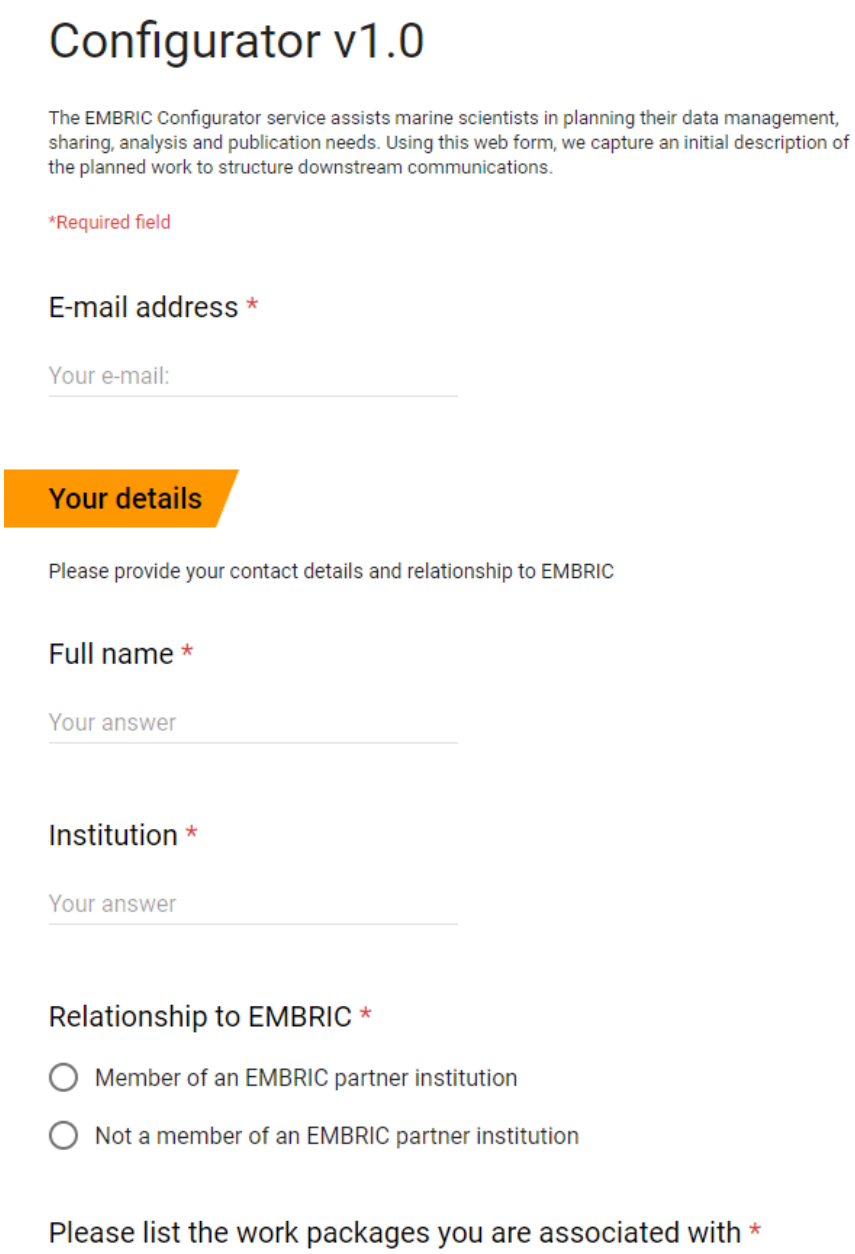


**Figure 3. Example of EMBRIC configurator form.**

**Implementation phase:**

The final report from the EMBRIC use case provided three recommendations for the AQUAEXCEL repository:

1) Increase user base. It was mentioned that 20 estimated users are not enough to get further support in ELIXIR services because of low number of the users. The solution to increase the number of users is to use the Central repository described in the deliverable D 3.2. Central system for metadata sharing. The Central repository is an extension of the AQUAEXCEL repository which enables to share the data with public (external users) under the specific access right. The Central repository itself does not increase the number of users. Therefore, the data from Digital fish (D3.6 – Elixir-Digital Fish case study), cryobanking (D3.3 – Booklet of cryopreservation procedures for the cryobanked species) and work package 7 - Fish lines will be stored first in

AQUAEXCEL repository and shared with public users using the Central repository. The availability of the datasets should increase the number of users of the AQUAEXCEL repository.

2) Data archiving – because of the low number of AQUAEXCEL repository it cannot be archived directly using the ELIXIR services. The possible way is to use existing ELIXIR repositories dedicated to the specific data types. Several archives were recommended for parallel data storage in AQUAEXCEL repository and data archive. There is no automatic way how to archive the data in the dedicated archive, the user has to do the archiving himself.
   To support separate data archiving in the ELIXIR archives the link in the AQUAEXCEL repository can be used. The protocol used in the AQUAEXCEL l repository for metadata description can contain a link to the external webpage. This link can be used to point to the data archived in one of the ELIXIR archives. ELIXIR usually provide the link to the archived data which can be used for direct access to the dataset.
   The new application interface (API) of AQUAEXCEL repository was also implemented to support the metadata export from AQUAEXCEL repository to external archive. The NAACCR XML standard was implemented for metadata export because it is used for the metadata representation in the BioSample repository operated by ELIXIR. More information about implemented API is described in milestone MS6 Standards applied to data repositories.

## Annex 1. EMBRIC Configurator case document: AQUAEXCEL2

### AQUAEXCEL2020 data repository sustainability

Case Study Contact People: Petr Cisar
Case Officer: Isabel Santos Magalhaes

**This document**

The AQUAEXCEL2 case presented itself as an exploration of sustainability of a project-specific data repository. In this document, we therefore assess the sustainability of the AQUAEXCEL2020 data repository, surrounding databases, data and workflows and make recommendations as to steps that can be taken to enhance sustainability beyond the current funding programme for the repository.
The assessment follows consultation with the group that has developed the repository. Sustainability is defined here as the ability to continue to operate the core functions supported by the repository, which may include work on the maintenance and ongoing operation of content, software, services and human workflows. In the text, we outline the repository and its context, highlight areas of particular value to be stressed in sustainability planning and provide recommended practical steps that can be taken by the developers to drive greater opportunities for sustainability in the future.

**Description of the AQUAEXCEL2020data repository**

The AQUAEXCEL2020 data repository is a platform for data and metadata storage from the work packages of the AQUAEXCEL project. The platform also supports cooperation between project partners and offers data storage for the project's Trans-National Access activities (TNA). The repository will ultimately contain various data types from different research areas within aquaculture, including from the "Digital Fish" initiative (see below) and the varied research themes that emerge from the project's TNA calls.
The developers of the AQUAEXCEL2020 data repository have demonstrated the functionality and value of the repository using the Digital Fish initiative. This initiative focuses on four species (Atlantic salmon, Rainbow trout, Common carp and Sea bass), with data types covering phenotypic traits, 'omics data and the outputs of genomic data analysis. Further data types are expected that include video recordings of fish behaviour, fish mortality statistics, 3D fish trajectories and internally sensed fish physiology data.

The core mission for the AQUAEXCEL[2020] data repository is to secure data generated in the AQUAEXCEL project. The team are exploring options to provide continuity in data availability after the project has reached its end.

**Recommendations:**

**1) Increase the user base**

With 20 users predicted until 2020 and 100 after that, the case for support for new funding to continue operations will be challenging. It is essential therefore to increase the size of the user base to gain engagement and weight in the community – e.g. offer and promote access to the repository beyond the AQUAEXCEL domain, for example to the EMBRIC community. This will also provide greater user feedback on needs and utility and will guide developments and functionalities.

**2) Archiving of data deposited in the AQUAEXCEL[2020] repository into established sustained data resources**

**Data submission:**

We recommend that the data submitted to the AQUAEXCEL repository are also archived – in parallel - in appropriate established public repositories in order to ensure sustainability of the data beyond 2020. Data archived in these resources can be kept confidential for a period, so short-term privacy requirements of users of the AQUAEXCEL[2020] Repository need not be compromised. We note below, though, the benefits of making data public as early as possible..

• Samples and related phenotypic data should be submitted to the European Nucleotide Archive (ENA) (http://www.ebi.ac.uk/ena/submit )[1]

• Raw data from genomics and qualitative transcriptomics should be submitted to the European Nucleotide Archive (http://www.ebi.ac.uk/ena/submit ).

• Genome and transcriptome assemblies should be submitted to the European Nucleotide Archive (http://www.ebi.ac.uk/ena/submit).

• Microarray data should be submitted to ArrayExpress (https://www.ebi.ac.uk/arrayexpress/submit/overview.html ).

• Quantitative RNA-seq data can be submitted to either ArrayExpress (https://www.ebi.ac.uk/arrayexpress/submit/overview.html ) or to the European Nucleotide Archive (http://www.ebi.ac.uk/ena/submit), but not both.

• SNP data should be submitted to the European Variation Archive (EVA; see https://www.ebi.ac.uk/eva/?Submit%20Data).

• Methylome, meiotic mapping studies, GWAS data and all protocols can be submitted to BioStudies ( http://www.ebi.ac.uk/biostudies/ submit.html ). All data relating to a single study should be referenced in BioStudies.

1 Sample record accessions (SAMEAXXXXXXX) in ENA should be cited in subsequent data submissions; in ENA their accessions can be cited directly; for ArrayExpress, please use the “Contact Us” tab on the ArrayExpress page linked above to indicate that the data are to be associated with samples that have been preregistered in ENA; for EVA, you will ultimately complete a spreadsheet with several fields of information – the sample accessions can be cited here.

**Accessibility of data deposited in the AQUAEXCEL[2020] repository**

Submitting data to the above ELIXIR resources will ensure the availability of the data after the funded phase of AQUAEXCEL has ended. We can provide advice around sample metadata if required, and as the data are quite complex we can also offer the services of a data coordinator to ensure the data is submitted to the repositories mentioned above.

We would encourage early data release as this has many advantages:

• Increased exposure of data: users finding datasets of interest in the databases will find the name and research centre associated with it, allowing for potential collaborators to contact the submitters,

• Connectivity with other ELIXIR data resources, enabling data to be propagated to secondary and tertiary databases, such as UniProt (http://www.uniprot.org/ ) and Ensembl

(https://www.ensembl.org/index.html ), in turn generating added information on top of the data, such as annotations, and

• integration with similar datasets which can be used to do further analyses on a more comprehensive overall dataset.

# 5. Definitions

**AQUAEXCEL[2020]**: AQUAculture Infrastructures for EXCELlence in European Fish Research towards 2020

**Creative Commons (CC) licenses** - is one of several public copyright licenses that enable the free distribution of an otherwise copyrighted work. A CC license is used when an author wants to give people the right to share, use, and build upon a work that they have created. CC provides an author flexibility (for example, they might choose to allow only non-commercial uses of their own work) and protects the people who use or redistribute an author's work from concerns of copyright infringement as long as they abide by the conditions that are specified in the license by which the author distributes the work.

**FAIR principles** – the list of 15 principles for discovery of, access to, integration and analysis of scientific data data. https://www.force11.org/group/fairgroup/fairprinciples

**Local repository** - is a specialized software that includes a local database, application communication interface, visualization framework, web services, web user interface and other APIs running on specific hardware (HW, especially disk array, etc.). There are two access points to the repository: Protocol manager and web interface. The local repository is designed to store and manage the experimental data and metadata.

**AQUAEXCEL repository** – local repository dedicated for the AQUAEXCEL[2020] project

**Central repository** - is a specialized software that includes a central database, application communication interface, visualization framework, web services, web user interface and other APIs running on specific hardware (HW, especially disk array, etc.).

**Protocol** – electronic protocol. Filled protocol template with experimental data stored in the database. The protocol is the complete description of the experiment or data processing. It contains metadata and data.

**Metadata** – data describing the experimental conditions of experimental data processing. It is an information needed for the reproducibility of the experiment or data processing

# 6. References

Because the document summarizes the information about ELIXIR organization and services whose are already well documented, the full text, lists and images from the https://www.elixir-europe.org/ web page are used. The source of the images used from another sources is referred directly in the image caption.

# 7. Document information

| | | | |
|---|---|---|---|
| **EU Project N°** | 652831 | **Acronym** | AQUAEXCEL2020 |
| **Full Title** | AQUAculture Infrastructures for EXCELlence in European Fish Research towards 2020 | | |
| **Project website** | www.aquaexcel2020.eu | | |

| | | | | | |
|---|---|---|---|---|---|
| **Deliverable** | **N°** | D3.4 | **Title** | Usable ELIXIR services, standards for data exchange and data modelling | |
| **Work Package** | **N°** | 3 | **Title** | Common standards and tools | |

| | | | | |
|---|---|---|---|---|
| **Date of delivery** | **Contractual** | 30/09/2018 (Month M36) | **Actual** | 12/11/2018 (M38) |
| **Dissemination level** | X | **PU Public, fully open, e.g. web** | | |
| | | **CO Confidential, restricted under conditions set out in Model Grant Agreement** | | |
| | | **CI Classified, information as referred to in Commission Decision 2001/844/EC.** | | |

| | | | | |
|---|---|---|---|---|
| **Authors (Partner)** | USB | | | |
| **Responsible Author** | **Name** | Dr. Petr Cisar | **Email** | cisar@frov.jcu.cz |

| **Version log** | | | |
|---|---|---|---|
| **Issue Date** | **Revision N°** | **Author** | **Change** |
| 17/10/2018 | V1 | Petr Cisar | First draft |
| 22/10/2018 | V2 | Petr Cisar | Second draft |
| 05/11/2018 | V3 | Petr Cisar | Third draft |
| 11/12/2018 | V4 | Petr Cisar | Final document |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |

# Annex: Check List

Deliverable Check list (to be checked by the "Deliverable leader")

| | Check list | | Comments |
|---|---|---|---|
| BEFORE | I have checked the due date and have planned completion in due time | X | *Please inform Management Team of any foreseen delays* |
| | The title corresponds to the title in the DOW | X | *If not please inform the Management Team with justification* |
| | The dissemination level corresponds to that indicated in the DOW | X | |
| | The contributors (authors) correspond to those indicated in the DOW | X | |
| | The Table of Contents has been validated with the Activity Leader | X | *Please validate the Table of Content with your Activity Leader before drafting the deliverable* |
| | I am using the AQUAEXCEL2020 deliverable template (title page, styles etc) | X | *Available in "Useful Documents" on the collaborative workspace* |
| ***The draft is ready*** | | | |
| AFTER | I have written a good summary at the beginning of the Deliverable | X | *A 1-2 pages maximum summary is mandatory (not formal but really informative on the content of the Deliverable)* |
| | The deliverable has been reviewed by all contributors (authors) | X | *Make sure all contributors have reviewed and approved the final version of the deliverable. You should leave sufficient time for this validation.* |
| | I have done a spell check and had the English verified | X | |
| | I have sent the final version to the WP Leader, to the 2nd Reviewer and to the Project coordinator (cc to the project manager) for approval | X | *Send the final draft to your WP Leader, the 2nd Reviewer and the coordinator with cc to the project manager on the 1st day of the due month and leave 2 weeks for feedback. Inform the reviewers of the changes (if any) you have made to address their comments. Once validated by the 2 reviewers and the coordinator, send the final version to the Project Manager who will then submit it to the EC.* |