AQUA EXCEL 2020

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

D3.6: Elixir-DigitalFish case study

François Allal, Adelino Canario

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: Digital Fish assembles in a single platform data and metadata on isogenic, divergent lines or more generally well characterized fish lines of carp, sea bass, trout, and Atlantic salmon.

Rationale: The use of common standards among RIs is essential for interoperability and the collaboration with the EMBRIC project partner ELIXIR-European Bioinformatic Institute (EBI) services to evaluate the standards for data exchange in AQUAEXCEL^{2020} Digital Fish as a practical case study.

Main Results:

During a workshop in Cambridge with ELIXIR-EBI followed by e-mail exchanges, an evaluation of Digital Fish needs in terms of interactions with existing EBI services was carried out with subsequent recommendations.

Authors/Teams involved:
Ifremer:
François Allal.

Table of content

Executive summary……..……..…2

Table of content……...……..……..…3

1. Rationale……..…..……...4

2. Results……..…..…….……4

3. Implementation……..……..……5

4. Conclusion……..……..……7

1. Rationale

1. The Digital Fish is a bioinformatics platform that stores the data and metadata on isogenic, divergent fish lines or more generally well characterized fish lines of carp, sea bass, trout, and Atlantic salmon collected by AQUAEXCEL2020 and beyond. The final aim is to assemble and make available to potential users existing data and metadata related to those lines of interest in a single platform. This data comes from several sources: published data in the public domain and from AQUAEXCEL2020 experiments carried out in WP7. The types of data included in Digital Fish are phenotypic data (i.e. growth traits, feed efficiency, fillet yield, age at first maturation, sex ratio, diseases resistance, metabolic efficiency, feed conversion ratio, lipid deposition, proximal composition, behavioural traits...), "omics" data (RNAseq, RAD sequencing, microsatellites, SNP array genotyping, whole genome sequencing, methylome sequencing...) and analytical data (from genome wide association studies, QTL mapping, genetic mapping...).The recording of such data is carried out using EBI services. Therefore, the present case study aimed to identify the relevant services of EBI that could be recruited to populate the Digital Fish database. In addition to fulfilling a specific task in AQUAEXCEL2020 WP3, it was included as a case study in the WP4 of the EMBRIC project, in which the practical usage of EMBRIC configurator services was tested.

2. Practical use of EMBRIC services to record the data and metadata in Digital Fish.

A EMBRIC/AQUAEXCEL2020 workshop was organized at the ELIXIR-European Bioinformatics Institute (EBI), Cambridge in February 2017 to introduce the Digital Fish objectives and architecture. As a result, a Digital Fish case study was identified by ELIXIR and the procedure to gather advice was established.
In August 2017, a detailed description of the case study was done by partner IFREMER. The case-study was defined as follows:

Study description:

IFREMER is running a pilot study on Sea Bass divergent isogenic lines for fasting tolerance. This pilot study focusses on 600 fish from two divergent isogenic fish lines (high feeding efficiency and low feeding efficiency). For each of these 600 fish the following data is being generated: phenotypic data (sex, growth indexes, fasting tolerance and risk-taking behaviour for example), raw genomic data (Illumina 3KSNPchip and potentially RNAseq data for 54 samples), and analysed genomic data (genetic map, QTL mapping, RNA differential expression data).
The scientific question behind this study is "can we evaluate individual feed efficiency of fish, and can we find genomic clues of evaluation of this phenotype for leading genomic selection of feed efficiency in breeding programs".
This case study will take advantage of the EMBRIC Configurator service, to provide an appropriate description of data sustainability infrastructure for AQUAEXCEL2020, and will serve as an EMBRIC case study, for which we will offer support in Digital Fish.

Following a questionnaire provided by EMBRIC, the type and volume of data was evaluated (phenotypic data, raw genomic data, and analysed data). The case study was analysed and a list of recommendations was provided by two ELIXIR consultants.

Recommendations:

Capturing rich metadata and archiving of data into established, sustained data resources

As the data are complex and varied we can offer the services of a data coordinator at the European Nucleotide Archive (ENA) to ensure submission of data to the repositories below:

• Samples and related phenotypic data should be submitted to ENA (http://www.ebi.ac.uk/ena/submit)[1]. We can discuss phenotypic data submission and set up a list of attributes for sample submission to ensure high quality metadata around these samples.

• Raw data from genomics and transcriptomics should be submitted to ENA (http://www.ebi.ac.uk/ena/submit).

• RNA-seq data can be submitted to either ArrayExpress (https://www.ebi.ac.uk/arrayexpress/submit/overview.html) or to ENA(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).

• Genetic maps, mapping studies 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.

Accessibility of data

Submitting data to the above ELIXIR resources will ensure the availability of the data after the funded phase of AQUAEXCEL2020 has ended.

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.

3. Implementation

Following the recommendations, a Study was created in ENA (PRJEB35091/ERP118087). Samples from the study were recorded in ENA (i.e. SAMEA6135427, SAMEA6135439...) and related genomic data, including RNAseq raw sequences and variant call to ENA and EVA (i.e. ERR3623684...), respectively. The related experiments were also recorded to Digital Fish using the web-based bioWES repository (Figure1).

[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 pre registered in ENA; for EVA, you will ultimately complete a spreadsheet with several fields of information – the sample accessions can be cited here.

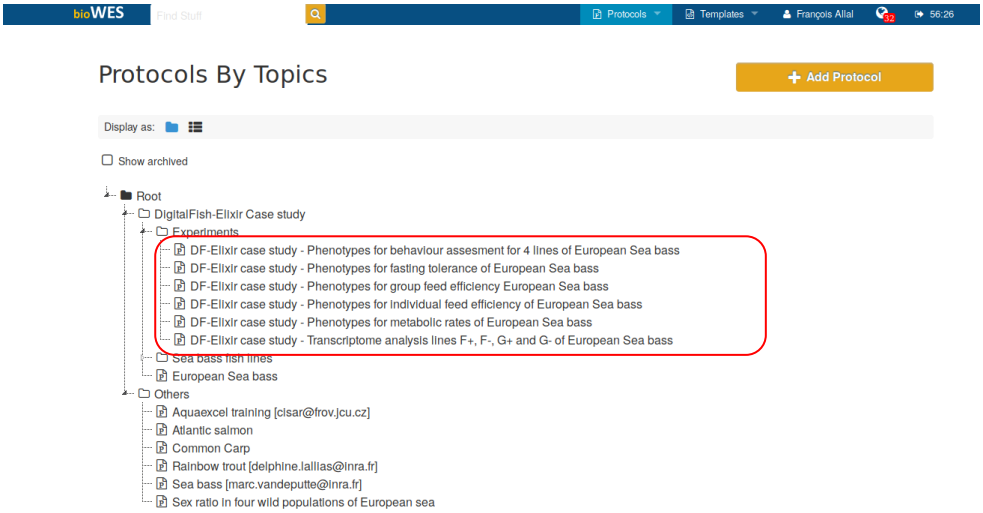

Figure 1. case study records in DigitalFish database

The accession numbers obtained from EBI (ENA/EVA) were linked to the specific protocols recorded in DigitalFish (Figure2).

Figure 2. Example of EBI-Digital Fish cross linked accession numbers

The usage of accession numbers to cross link Digital Fish to EBI-Elixir references will be generalized and is promoted throughout the AQUAEXCEL2020 project. The usage of Elixir services ENA and EVA will be done for all species targeted in Digital Fish (Salmon, Carp, Trout and Sea bass) for genomic relevant traits.

4. Conclusion

Through a collaboration with the European infrastructure for life sciences data (ELIXIR) it was possible to include a series of standards which makes Digital Fish storage of complex data interoperable with public databases. At the same time Digital Fish promotes early publication of data making data more visible and fostering collaboration.

Document information

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

Deliverable	**N°**	D3.6	**Title**	Elixir Fish case study
Work Package	**N°**	3	**Title**	Common standards and tools

Date of delivery	**Contractual**		30/09/2019 (Month 48)	**Actual**	26/11/2019 (Month 50)
Dissemination level	PU	**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)				
Responsible Author	**Name**		**Email**	

Version log			
Issue Date	**Revision N°**	**Author**	**Change**
dd/mm/yyyy			Ex: first version/first review by WP leader etc/accepted version

Annex 1: 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		*Please inform Management Team of any foreseen delays*
	The title corresponds to the title in the DOW		*If not please inform the Management Team with justification*
	The dissemination level corresponds to that indicated in the DOW		
	The contributors (authors) correspond to those indicated in the DOW		
	The Table of Contents has been validated with the Activity Leader		*Please validate the Table of Content with your Activity Leader before drafting the deliverable*
	I am using the AQUAEXCEL^{2020} deliverable template (title page, styles etc)		*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		*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)		*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		
	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		*Send the final draft to your WPLeader, 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.*