ASSEMBLE
ASSOCIATION OF EUROPEAN MARINE BIOLOGICAL LABORATORIES EXPANDED

ASSEMBLE
ASSOCIATION OF EUROPEAN MARINE BIOLOGICAL LABORATORIES EXPANDED

Acronym: ASSEMBLE Plus

Title: Association of European Marine Biological Laboratories Expanded

Grant Agreement: 730984

Deliverable [D4.9]

[First virtual access run to the analysis platform]

[10][2022]

Lead parties for Deliverable: [VLIZ]
Due date of deliverable: M 60 [30/09/2022]
Actual submission date: M 60 [31/10/2022]

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 727610. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 727610. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein.

GENERAL DATA

Acronym: **ASSEMBLE Plus**

Contract N°: **730984**

Start Date: **1st October 2017**

Duration: **60 months**

Deliverable number	D4.9
Deliverable title	First virtual access run to analysis platform
Submission due date	30/09/2022
Actual submission date	31/10/2022
WP number & title	WP4, Improving virtual access to marine biological stations data, information, and knowledge
WP Lead Beneficiary	VLIZ
Participants (names & institutions)	**Katrina Exter** (WP leader NA2), Flanders Marine Institute (VLIZ); **Georgios Kotoulas** (WP Leader JRA1), Hellenic Center of Marine Research (HCMR); **Dan Lear** (NA2 partner, EMBRC WGEI), Marine Biological Association (MBA); **Ivaylo Kostadinov** (NA2 partner), Max-Planck Institute for Marine Microbiology (MPIMM)

Dissemination Type

Report	X ☒
Websites, patent filling, etc.	☐
Ethics	☐
Open Research Data Pilot (ORDP)	☐
Demonstrator	☐
Other	☐

Dissemination Level

Public	X ☒
Confidential, only for members of the consortium (including the Commission Services)	☐

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 727610. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein.

Document properties

Author(s)	Katrina Exter
Editor(s)	Katrina Exter
Version	1

Abstract

This is the final reporting on the virtual access run on the analysis platform provided via ASSEMBBLE Plus as part of WP4 work.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 727610. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 727610. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein.

1. Introduction

This deliverable is to report on *virtual access run to the analysis platform*: we report on what the "virtual analysis platform" (VAP) is and how it is being used.

This is the first and also the final report on the VAP, since its development was not far enough advanced to write reports before this point.

2. The virtual analysis platform

The "virtual analysis platform" (VAP) is part of Task NA2.5 (WP4), and it was envisaged to be a platform on which standardised data from long-term biodiversity and genomics observatories could be analysed. Creating such a platform from scratch was not in scope, rather promotion of the long-term biodiversity and omics data from JRA1 (OSD and ARMS-MBON) on existing VAPs was to be the goal. Such GO (Genomics Observatory) data – consisting of eDNA and image data, and environmental measurements – have long been analysed by individual scientists in a local context (i.e. the data are located on their own computers), but (at the time the ASSEMBLE Plus project was proposed) it was less common to be able to find, access, and process data via a cloud environment. The aim with this task was to explore how to do this with the JRA1 data, which have the advantage of being fairly interoperable, long-term datasets.

Three avenues were eventually taken in this direction.

1. The Marine VRE is a web portal gathering marine-related analysis tools and workflows: providing a summary of each object and links to where they can be accessed. This is a portal that had already been developed at VLIZ in cooperation with LifeWatch Belgium, and it is owned, developed, and maintained by LifeWatch Belgium. ASSEMBLE Plus has used this portal to list a number of resources.
2. The LifeWatch Internal Joint Initiative Tesseract workflow on Non-native and Invasive Species (NIS). This is Tesseract workflow environment is an initiative of LifeWatch ERIC that began in 2019-20. It incorporates a number of individual workflows that each deal with different use-cases with different data, but with the overarching theme of science related to NIS. One of those workflows uses the ARMS-MBON data from JRA1, and we worked (and are still working) extensively with LifeWatch in the development and promoting of this workflow.
3. The ESOC-Life Open Call project PID 14324: development of an metagenomics workflow (MetaGOFlow) for marine genomics observatories, using OSD (and EMBRC's EMO BON) data. This project is funded by the Horizon project ESOC–Life, and is linked to ASSEMBLE Plus by the people and data involved. The emphasis of these EOSC-Life funded "open call" projects is to make data and/or workflows more FAIR, and that is where ASSEMBLE Plus is contributing to the project: providing the data with full provenance metadata (the R part of FAIR), and ingesting back scientific results from the workflow also with provenance metadata. This work uses the OSD developments (data, semantics, machine-accessibility) on the OSD and EMBRC's EMO BON Github repository.

The first – the Marine VRE – was set up in the first few years of ASSEMBLE Plus. The second two arose serendipitously, via personal and project interactions with LifeWatch and EOSC-Life. We took

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 727610. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein.

advantage of the common interests between all parties in developing FAIR workflows to work on data from JRA1, and subsequently also data from EMBRC's EMO BON project (which has benefited greatly from the work of JRA1).

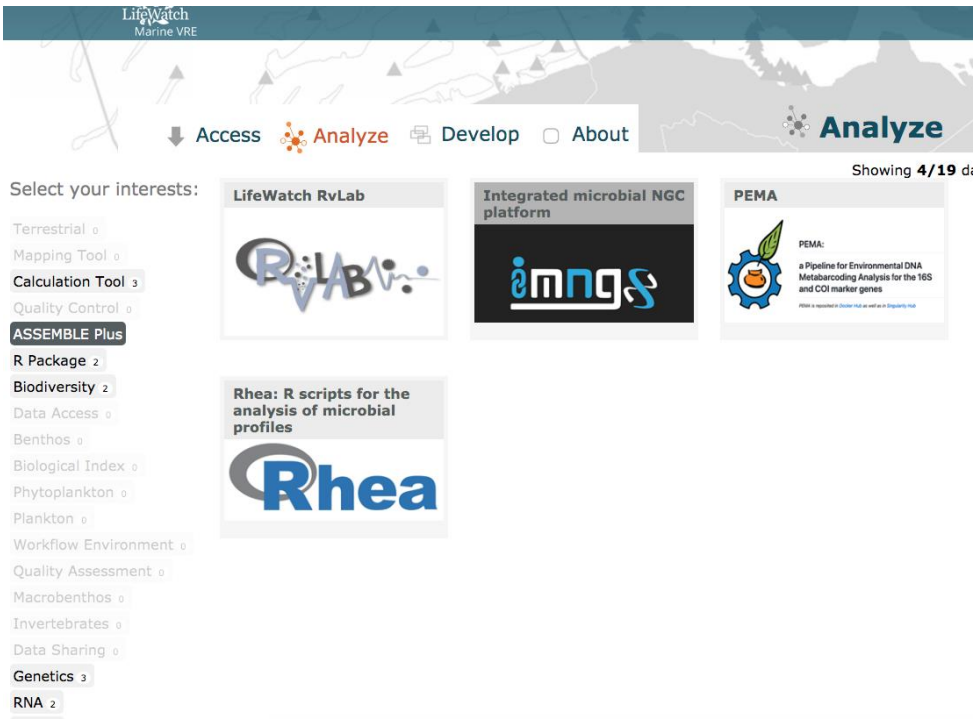
Task NA2.5 originally also envisaged setting up an authentication system for (tracking) use of the VAP(s). This part of the task has been dropped given how this "virtual analysis platform" has developed: rather than a single access and entry point, we have three different systems (1) a catalogue of marine-related online-accessible tools, (2) an entire workflow system running on the LifeWatch IT infrastructure, and (3) a pipeline running on the MGnify workflow (from EMBL-EBI). A single access point for authentication and for gathering user statistics is not possible with this arrangement.

3. The MarineVRE

The tools that that were added by ASSEMBLE Plus to the Marine VRE are:

- Data access tools. We have added a description of, and links to, the OSD and ARMS metadata records in IMIS, and to the larger ASSEMBLE Plus datasets collection in IMIS.
- Data analysis tools. Four tools are listed in this part of the Marine VRE.
    1. The LifeWatch RVLab (IMBCC-HCMR with FORTH-ICS and LifeWatch Greece), for statistical analysis of marine data
    2. PEMA (IMBCC-HCMR) for metabarcoding analysis, and as used by OSD and ARMS
    3. Rhea (IMBCC-HCMR) for analysis of microbial profiles
    4. IMNGS (IMBCC-HCMR with TUM), the integrated microbial NCG platform

    For each tool, the user is directed to an information page, from where they can click on a link to be directed to the page from where the tool can be downloaded (Rhea, PEMA) or accessed online (RvLab, IMNGS).



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 727610. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein.

3.1. Views, hits, and use of these tools

The number of unique page views since July 2020 (when tracking started) are the following

- The VRE page on the ASSEMBLE Plus site: 122
- OSD "data access" on the MarineVRE: 79. The ARMS record on the MarineVRE was only created later in 2022 and no views are recorded.
- ASSEMBLE Plus LTEDs collection on the MarineVRE: 20
- Rhea, IMNGS, PEMA, and RvLab on the MarineVRE: 142, 55, 102, and 41

Because the four tools are all hosted on their own websites, we cannot track viewers moving from the MarineVRE site to the site of each individual tool. However, we can report on the traffic statistics produced by the individual hosting sites.

- PEMA: PEMA is available via Dockerhub, and there have been 1400 pulls from there since it was placed on the site (12/04/2019). There have been 27 citations (and over 8000 views) of its 2020 publication in Giga Science (DOI 10.1093/gigascience/giaa022).
- RvLab: RvLab can be run via a web-interface, and it has run 358 jobs runs for 14 registered since 2020, and in the year 2022 to date, there have been 1431 views of the RvLab page on the HCMR site. There are currently 10 citations of its 2016 publication in Biodiversity Data Journal (DOI 10.3897/bdj.4.e8357), which also has had over 4000 views.
- IMNGS: Between 2019 and 2022 there have been 289, 211, 237, and 237 new users added, and 1699, 1851, 1023, 1762 tasks were run. There are currently 259 citations (and over 6400 views) of its 2016 publication in Scientific Reports (DOI 10.1038/srep33721).
- Rhea: Rhea is provided via Github, and the traffic collecting of Github only extends to a 2-week period. The number of unique visitors per day over the 2-week period when checked in 2020 and again in 2022 were 9 and 123, with 196 views for the 2022 period. There are currently 237 citations of its 2017 publication in PeerJ (DOI 10.7717/peerj.2836).

4. The LifeWatch NIS workflow

The LifeWatch Internal Joint Initiative Tesseract workflow on Non-native and Invasive Species (NIS): this LifeWatch initiative is to develop a workflow environment in which investigations on data of NIS can be undertaken. A set of five use cases were developed simultaneously, one of which is the "ARMS" workflow.

The ARMS workflow starts with the data collected by the ARMS-MBON project, offering an overview of all the ARMS sampling events to-date and allowing people to select which parts of those data they want to process through the workflow. This overview is taken from the ARMS-MBON GitHub repository. Eventually it will be possible to process both the ARMS-MBON image data and the sequence data through this workflow, but as of Sept. 2022, only the sequence analysis pathway has been developed. The user can choose which sequences they wish to process via the PEMA pipeline, they can upload or create the necessary PEMA parameter file, and then they can launch the job. This job consists of:

1. Running PEMA on the chosen sequences with the chosen parameter file.
2. Taking one of the PEMA output files (the "final table", which contains the OTU/ASVs and the taxonomic identifications assigned by PEMA), and adding to that the AphiaIDs from WoRMS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 727610. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein.

(World Register of Marine Species), where a match to that omics-taxonomic species name can be found (using the taxon-match tool of WoRMS).

3. Again using the WoRMS subcollection WRIMS (World Register of Introduced Marine Species) and the latitude, longitude of the sample site where each sequence came from, a check on the NIS status ("known to be native", "known to be introduced", "known to be present", "no information available") of the species listed in the PEMA output is done. Flags are added to the PEMA output to convey this information.
4. These modified outputs files, together with the standard output of PEMA, can be downloaded by the user from the workflow, but they are also retained within the user's account on the workflow.

This Tesseract workflow environment has been created and the workflows situated therein. Fine-tuning of the five individual workflows is still in progress, and hence they are not yet fully available to the general public, although the ARMS workflow has been tested by some of the ARMS-MBON scientists (it may be possible to access the Beta version on: https://51.210.38.65/personal-space). The contribution of VLIZ, HCMR, and UGOT to this, as part of NA2.5, is the following:

- VLIZ provided the initial workflow requirements and a design for the user interface to allow those requirements to be met
- UGOT provided much of the scientific input required for the workflow design, both for the input and output requirements (i.e. "what will people want to do")
- VLIZ provided examples of all datasets that would be processed by and created within the workflow (where arising from PEMA, WoRMS, or WRIMS)
- HCMR provided the PEMA code, RvLab, and worked with the developers to incorporate these in the workflow
- VLIZ, HCMR, and UGOT performed user testing of the developing workflow, with numerous telecons with the development teams to sort out problems
- VLIZ provided the code to interrogate WoRMS and WRIMS and to update the PEMA output files accordingly
- VLIZ provided machine-interoperable (turtle and CSV) files with the overview of the ARMS data to be ingested as the starting point of the workflow, and worked with the developers on the ways to present these data in the workflow
- VLIZ, UGOT, and HCMR participated in workshops promoting this workflow among the ARMS and other scientists.

It is expected that a first public release will be demonstrated at a workshop to be held at the International conference on Ecological Sciences which will be held in Metz on 21-25 November 2022. As the workflow is not yet public, there are no user statistics to report.

Advertising of this workflow, once ready for "the public" will be done via LifeWatch and EMBRC.

5. The EOSC Life Open Call project

The Open Call programme of EOSC Life funds project to work on the development of life-science data and tools (workflows, catalogues, portals, ontological services, etc) to make them more FAIR. A team of EMBRC scientists proposed the development of a "workflow for marine Genomic Observatories (GO) data analysis". The aim of the project (PID 14324, aka MetaGOFlow) is to modify an existing workflow

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 727610. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein.

in MGnify to work specifically on EMBRC's GO data (shotgun metagenomics), with the JRA1 data from OSD as the first example datasets, to later to complemented by data from EMBRC's EMO BON project. This new workflow pathway will allow researchers to deal better with the increasing amount of data arising from GOs, will make the data produced by the GOs more easily interpretable by providing the taxonomic inventories of each sample in a timely manner and in a non-technical format, and will also provide a complete accounting of the provenance of the data that are processed.

VLIZ have provided support for this project in the area related to

1. Providing access to the OSD data. These are the data included in the OSD GitHub pages, as described in CSV and turtle format: from this starting point, users can see what OSD data are available to be analysed. Later, the EMBRC EMO-BON data will be provided for the workflow in the same way.
2. Providing recommendations for how to work with provenance for datasets and workflows. As the data that will be accessed from GitHub are semantically annotated and packaged in Ro-Crates, provenance information regarding the input data are included in those data packages. Recommendations were given to the EMBRC team on how to describe and organise the outputs of the workflow also in Ro-Crates, from where they can be placed back on the OSD GitHub site to be made available hence also outside of Mgnify.

This project will end at the end of 2022, and work on the provenance part thereof is still underway. As the workflow is not yet public, there are no user statistics to report.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 727610. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein.