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 [D32.4]

[Final year access report on VA users]

[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.

GENERAL DATA

Acronym: **ASSEMBLE Plus**

Contract N°: **730984**

Start Date: **1st October 2017**

Duration: **60 months**

Deliverable number	D32.4
Deliverable title	Final year access report on VA users
Submission due date	30/09/2022
Actual submission date	31/10/2022
WP number & title	WP32 VA Virtual Access
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); Christina Pavloudi (JRA1 and NA2 partner), Hellenic Center of Marine Research (HCMR); Haris Zafeiropoulos (JRA1 and NA2 partner), Hellenic Center of Marine Research (HCMR), Antonis Potirakis (JRA1 partner), Hellenic Center of Marine Research (HCMR), Ilias Lagkouvardos (JRA1 partner), Hellenic Center of Marine Research (HCMR), Georgos Tsamis (JRA1 and NA2 partner), Hellenic Center of Marine Research (HCMR)

Dissemination Type

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

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

Dissemination Level

Public	X ☒

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 to ASSEMBLE Plus resources: a reporting on the access of VA users up until the end of the ASSEMBLE Plus project. This document explains what is covered by virtual access, the resources that are offered via this virtual access, the requests and hits that have been made by users to these resources, and recommendations for how better to do this in the future.

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 the Virtual Access that is provided by ASSEMBLE Plus via its website www.assembleplus.eu. While "virtual access" formally can include all resources provided by a website, in this report we will concentrate on virtual access to the data resources and publications, and on virtual access to analysis environments that are provided via ASSEMBLE Plus together with other European RIs.

This report is preceded by D32.3 and builds on that; for each main section, a response to the comments of this first report is included. Information has also been added for the "virtual access to analysis environments" since the first report. Finally, this report also draws heavily from D4.9 "First virtual access run to the analysis platform".

2. The ASSEMBLE Plus virtual resources

The main menu of the ASSEMBLE Plus website provides virtual access to the following Results of ASSEMBLE Plus:

- ASSEMBLE Plus outputs
  - Datasets created by ASSEMBLE Plus or collected during the ASSEMBLE plus project
  - Publications resulting from ASSEMBLE plus work and from ASSEMBLE Plus partners
- Access to scientific data tools via the MarineVRE
  - We note here that two additional virtual research environments are currently under development together with other RIs; these are not yet available and not linked from the website, but will be described in this report (as they are also in D4.9).
- The State of the Science Stories
- Access to the ASSEMBLE Plus Deliverables

The first three were created under WP4/NA2 (*Improving virtual access to marine biological stations data, information and knowledge*). The Deliverables page is created by the project management, and gives access to the ASSEMBLE Plus deliverables (which includes scientific outputs from some of the JRA programmes of ASSEMBLE Plus).

Note that we do not require user registration for those searching our catalogues of data and publications, or those accessing the VRE(s). Therefore, we cannot include statistics of individual use of any of these resources.

3. Virtual access to ASSEMBLE plus outputs

Data

The data created under the JRA (Joint Research Activity) and TNA (Transnational Access) programmes are considered primary ASSEMBLE Plus data. As ASSEMBLE Plus participated in the Open Data Pilot, these research data are expected to be made public via a metadata record in the IMIS (Integrated Marine Information System) datasets catalogue, with an open access licence. This is documented in the ASSEMBLE Plus DMP (D4.3).

- All TNA users are requested to create a metadata record in the IMIS datasets catalogue, once their access has been completed. This record should describe the data they collected during

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.

the access, and these data should be uploaded to the MDA (Marine Data Archive), from where they can be linked to their IMIS record. The record is made public immediately, and the data are to be made Open Access within no more than two years after their access. The TNA users are given guidelines as to how to do this on the ASSEMBLE Plus site.

- The JRAs are also required to upload their data to the MDA or to another community-used public archive, and to create a metadata record in IMIS with a link to the data. The record and the data are to be immediately Open Access (allowed opt-outs are described in the ASSEMBLE Plus DMP).

We also collate data records from the ASSEMBLE Plus partners and marine stations, where they are published already in IMIS. These are not ASSEMBLE Plus records and as such are not required to be open access, but by including them in the collection they are additionally "advertised. VLIZ, as the owner of IMIS, additionally has worked on improving the FAIRness of these records.

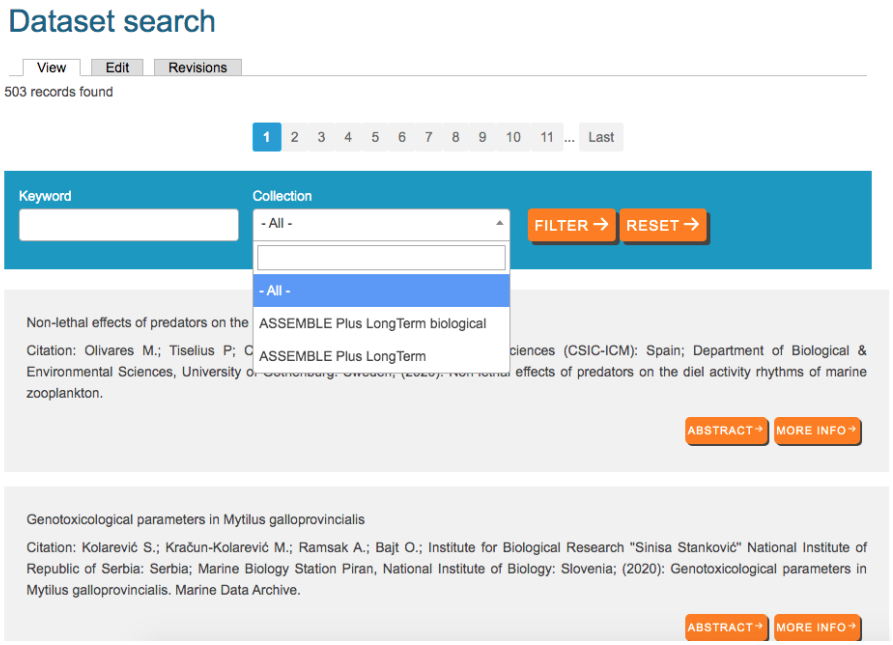
All the work on creating, maintaining, and improving the FAIRness of these various ASSEMBLE Plus data are reported on in deliverable D4.6. In summary:

- Up until mid-October, 2022, 30 TNA metadata records have been added to the IMIS datasets catalogue. Most have chosen CC BY licence, and while nine provide the data for download via the record, for the others a request to the indicated data contact can be sent by email.
- All the JRA1 data have been added to IMIS and the data can be downloaded directly from those records. For JRA2 four datasets have been archived in IMIS.
- About 500 records from the ASSEMBLE Plus partners are in IMIS and linked to the ASSEMBLE Plus datasets collection. Some work on improving the metadata in these records, to improve their FAIRness, has been undertaken throughout the ASSEMBLE Plus project. Approximately half the data records (and 2/3 of the LTEDS) are open access (with a licence most typically being CC BY or "unrestricted after an embargo period") although a great deal of those (e.g. ~70% of LTEDS) do not actually have a direct download link in the IMIS record (instead, an email can be sent to the indicated data contact).

Access to the (meta)data

IMIS is the metadata catalogue we use to make our data records public and FAIR. The JRA, TNA, and partner datasets are linked to the ASSEMBLE Plus collection within this catalogue, and this collection can be accessed from the search page on the ASSEMBLE Plus site.

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.



Data records in IMIS are described by a wealth of metadata, including keywords that are added by the data owner or by the IMIS experts: keywords for scientific topic, data type, type of experiment, etc. We additionally classify records by a "collection" type, which for ASSEMBLE Plus is our grouping of the records into those that are *long-term* and *long-term: biological*. Users can search on the keywords for each (sub)collection on our search page (see figure above). Finally, users can also filter on TNA and JRA[1,2,3,4,5] dataset records.

By clicking on "More Info" for any search result listing (see figure above), a user can view the metadata record. If the data have an open licence and are linked to the metadata record, the user can also directly download the data (when archived in the MDA) or follow links to the data (when provided via an external archive or portal).

The metadata can also be accessed programmatically using the webservices of IMIS, in xml, json, and eml formats.

Views on the data records

The landing page describing the ASSEMBLE Plus datasets collection, from where users can click onto the catalogue search page, has been available since March 2019. There have been 442 unique searches performed on the dataset search page between 2018 to Oct 2022. North America and Europe form the largest origin of these visits, with only 10% from outside these areas.

Publications

The guidelines for data created or collated by ASSEMBLE Plus apply also to publications (books, conference proceedings, scientific publications, or data papers): these should archived in the IMIS publications catalogue and made Open Access (Green or Gold publishing). We collect publications from the JRA and TNA programmes, but we have not pro-actively collected publications from our partners if not linked to ASSEMBLE Plus. However, publications from the previous EMBRC-related project, Assemble Marine, have been added to the collection.

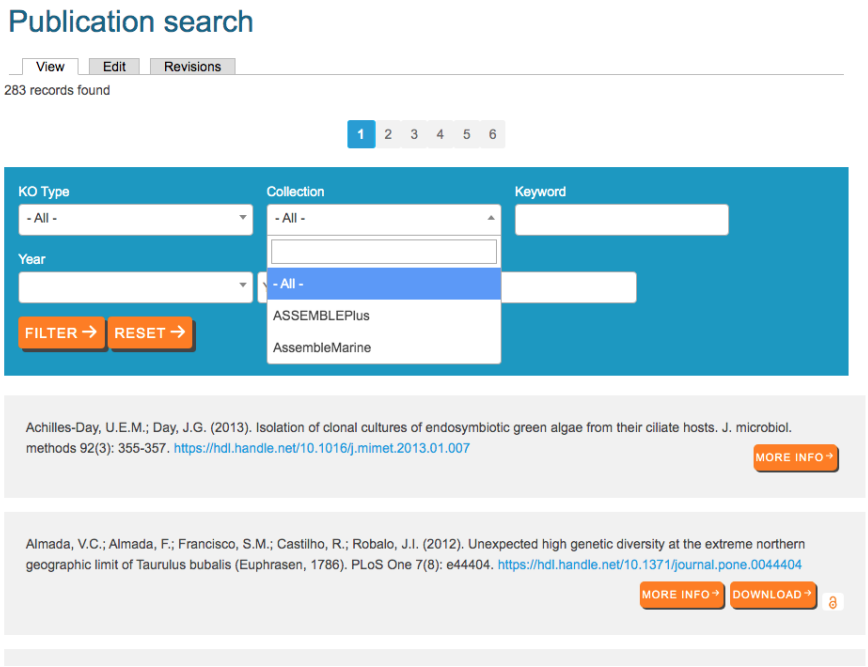
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 727610. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained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

The status of the publications collection is also reported in deliverable D4.6. In summary:

- By looking for an acknowledgement to ASSEMBLE Plus in publications, we determined that at least 60 of the publications arise from TNA projects. An additional 41 publications are from the JRAs (JRAs 1,2 and 3). Publications from JRAs 4 and 5 are certainly present, but we do not have the information to identify which ones these are.
- Of the 60 known TNA publications, only 8 are not open access or available via the ASSEMBLE Plus Open Repository, and we can consider this to be a success. All the JRA publications are open access or accessible via the Open Repository.

Access to the (meta)data

Once in our collection, publications can be search from the search page on the ASSEMBLE Plus site. This page allows users to search on keyword, date, author, sub-collection, and KO (knowledge output) type. The keywords describe the publication (e.g. scientific topic, experiment type, species, etc) and are added by the record creator or by our library specialists. The sub-collections here are *ASSEMBLE Plus* and *AssembleMarine*. The KO types used are: scientific publications, book, case study, software/modelling, prototype, services, and exploitable scientific result, as well as JRA[1,2,3,4,5] and TNA to allow for a filtering on these subcomponents of ASSEMBLE Plus outputs.



From the search results the user can click to download the publication (if open access) and can click to look at the metadata record. To accommodate those publishing as Green access, i.e. for which the journal is willing to provide a version of the publication for access only via an Open Repository, we have created the ASSEMBLE Plus Open Repository: these publications can be downloaded to be read only after clicking an agreement to not distribute the PDF any further. Of the 31 ASSEMBLE Plus publications, 22 are open access/open repository.

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.

Views of the publications

The landing page describing the ASSEMBLE Plus publications collection, from where users can click onto the catalogue search page, has been available since March 2019. There have been 350 unique searches performed on the publications search page between 2018 to Oct 2022. North America and Europe also form the largest origin of these visits.

Response to panel's comments (D32.7)

In D33.7, a reporting by an international panel on the previous version of this deliverable (D33.3), comments were made on the virtual provision of ASSEMBLE Plus data and publications. We respond to those comments here.

1. *The D32.3 report clearly defines the different types of data and metadata that have been collected, FAIRified, and made accessible through the ASSEMBLE Plus portal. It also mentions the various public repositories that data produced by ASSEMBLE Plus projects can use to store the actual data.* ***However, it is not clear in the report to what extent ASSEMBLE Plus data producers can expect to be assisted in this task by dedicated ASSEMBLE Plus staff.*** *Even if the report accurately emphasises in the recommendations section that achieving data FAIRification can be considered as potentially time-consuming and sometimes non-rewarding activities.* The responsibility of creating FAIR datasets (both in terms of the data themselves and in terms of their publications) and FAIR scientific publications fell on the data producers, being the TNA users and the JRA scientists. The ASSEMBLE Plus DMP (D4.3) explained how this could be done. Assistance from WP4 came in the form of (i) advice, documentation, a workshop, and eventually an online FAIR course, and (ii) access to the IMIS metadata catalogue and the Marine Data Archive. For IMIS and the MDA, this assistance included the standard helpdesk that is provided by VLIZ (who created and runs these resources). WP4 also gave the JRAs, TNAs, and the partners specific advice for the FAIRification of their datasets and how to improve their metadata records in IMIS. This is explained in D4.6.
2. ***It is not clear whether there is a means to be more specific about the metadata fields where the keyword has to be searched for.*** *It would improve search efficiency if the user was allowed to narrow it down using a more advanced search form. In particular, it would allow for searches covering a specific geographical area and/or a given timespan. As this is metadata that users have entered when creating the IMIS record, it could be helpful to enable them to search using these criteria. Another relevant search criterion would be the licence of the dataset described in the metadata record, or the actual dataset availability. As outlined in the report, there is a significant amount of records for which actual data is not readily available, and/or lacking an open access licence, or simply lacking a link to the dataset.* It was not made clear that there is an advanced search form that can be used for browsing the IMIS dataset and the IMIS publications catalogues (click on the words for the links). It is possible to search on: author, title, year, keywords, geography, taxonomy, etc.
3. ***Also when searching by clicking on a component of the burst chart, the number of returned results doesn't match the figure of the component's tooltip*** *(for example, as of this writing, the "Fish" component shows 177 whereas the search returns 173 records).* This is because of a difference in the metadata that the basic search box "Keyword" searches through and that which is counted for the figure. As the two pieces of code were developed differently, it is too late now to change them.
4. *Finally, a factor favouring (meta)data re-use is the ability for third party projects to* ***programmatically access the published resources through a well-defined API*** *providing*

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.

search and retrieval possibilities. The report does not mention whether this has been considered and planned or discarded or postponed. The IMIS metadata catalogue has webservices that allow programmatic access to metadata and data. Indeed, it was an oversight not to mention this. See https://www.vliz.be/en/imis?module=webservices for more information.

5. ***Proposal**: **A minimum set of information could be designated without which a IMIS record will be considered invalid and removed from the repository.** Alternatively, such meta-data poor records could be only included in searches when actively requested. Such a minimum set of meta-data fields need not be as comprehensive as those currently designated obligatory for deposition in IMIS, but just sufficient to weed out data sets that are of such poor quality that they are of little use to users.* This was done and explained to the attendees of the workshop in 2019 and explained later in personalised emails sent out to each station with datasets records in IMIS. It is also explained in the IMIS metadata/data submission form. It is proving very difficult to figure out how to get people to read these instructions, other than a complete redesign of the submission form (which is beyond the financial capabilities of VLIZ). However, based on the experience working with projects such as ASSEMBLE Plus, IMIS is working on standardised and semantically annotating some of the metadata fields, including licence, keywords, author details, parameters, and instruments.
6. ***Proposal: Partner institutes should be encouraged to provide training on how to collate and manage project DMPs and provide assistance through data specialists where available.** TNA final reporting should specifically reference the actions included in the DMP at time of review.* This would be lovely. "A course on FAIR data for marine biologists" was developed under ASSEMBLE Plus with the aim of providing teaching material that partners could push onto their students (staff?). The final DMP (D4.3) includes a summary of the actually-performed data management activities in ASSEMBLE Plus

4. Access to ASSEMBLE plus VREs

Under Task NA2.5 (WP4) *Set up virtual platform for data analysis* we provided access to scientific tools and workflows that were created by ASSEMBLE Plus partners or those that will be created within the ASSEMBLE Plus project. This is reported on in D4.9. Three virtual platforms were eventually provided, although only the first is linked from the ASSEMBLE Plus website as the others are still in development.

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 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.

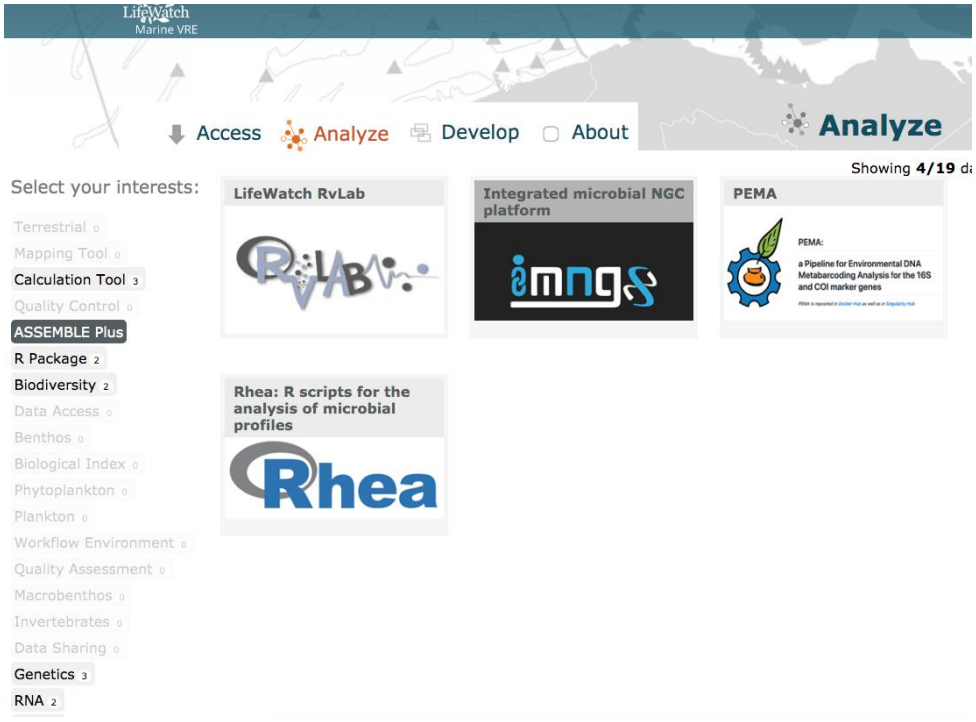
This project is funded by the Horizon project ESOC–Life, and is linked to ASSEMBLE Plus by the people and data involved. The emphasis of these EOSC-Life funded "open call" projects is to make data and/or workflows more FAIR, and that is where ASSEMBLE Plus is contributing to the project: providing the 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 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).



Views, hits, and use of these tools

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

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.

- 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 page requests for 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).

The LifeWatch NIS workflow (NEW in this report)

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

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). 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.

The EOSC Life Open Call project (NEW in this report)

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 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. This project will end at the end of 2022, and work on the provenance part thereof is still underway. The workflow will then be offered for anyone to use, especially on EMBRC GO data.

Response to panel's comments (D32.7)

In D33.7, a reporting by an international panel on the previous version of this deliverable (D33.3), comments were made on this MarineVRE. We respond to those comments here.

The MarineVRE is a development of LifeWatch Belgium, and as such ASSEMBLE Plus cannot control the development of the website. Note that the MarineVRE is not a full metadata catalogue (i.e. it does not have all the functionalities that a catalogue offers), rather it is a webpage in which resources are described and tags are added. A number of relevant suggestions were made by the panel that we have passed on to LifeWatch Belgium, who have responded positively. A revamp of the scope and approach taken by this site will be undertaken in 2023, and these comments will feed into that.

1. ***Proposal on the keywords provided for the resources****: It could be relevant to order these keywords either alphabetically or by classes*. This is an excellent suggestion and will be implemented in 2023. It is likely that the ordering will be done primarily on class and then alphabetically.

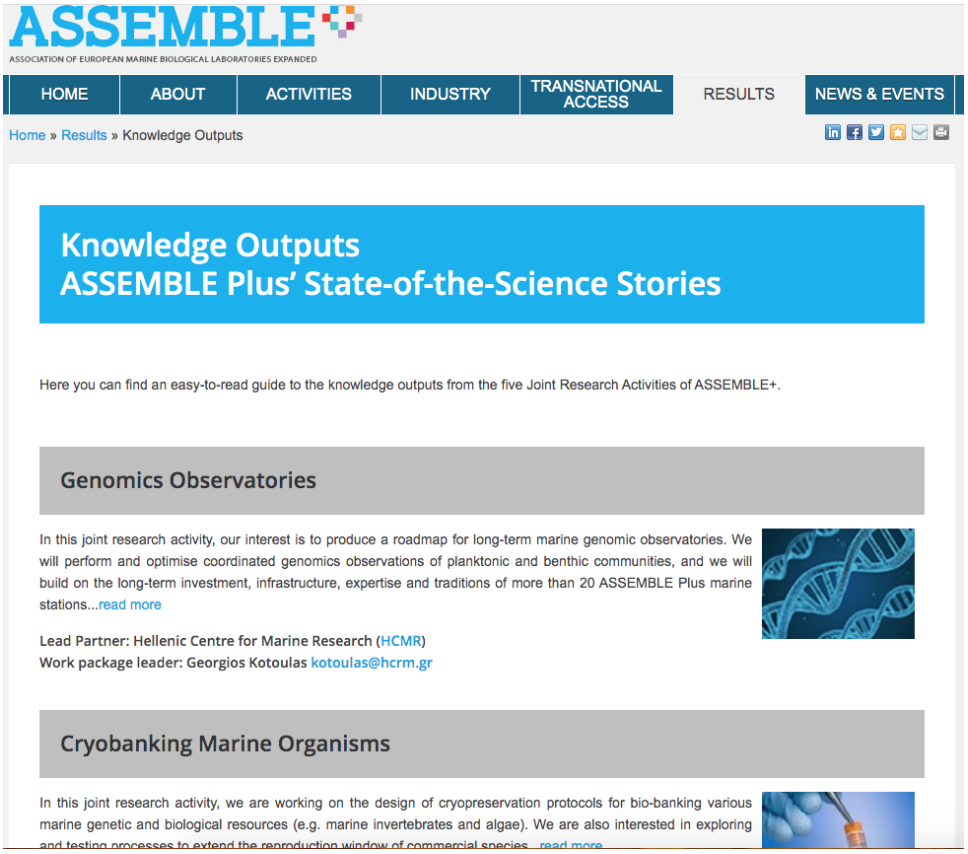
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 727610. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained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

2. ***Proposal on the resources listed on the website****: The list of portals accessing the dataset does not seem to be exhaustive. So far many LifeWatch ERIC resources are described there, but it would be advisable to motivate the partners of the project to promote the access to their own portals in this catalogue. This enrichment would allow a more precise categorization of the accesses (general data, physico-chemical data, data on large eukaryotes, data on protists, data on microorganisms, genetic data, etc.)…. For the analysis of environmental data many pipelines are available in the European and national Galaxy instances (...). An instance specifically dedicated to environmental data ... is hosted on the European instance*… The first set of tools on this website were LifeWatch-focussed because the MarineVRE was a development of LifeWatch Belgium. However, the intention is to not keep it so limited. Since D33.3, a LifeWatch ERIC metadata catalogue has been developed in which software, workflows, and VREs are catalogues in a machine-accessible way (i.e. this is a full metadata catalogue rather than just a website). LifeWatch Belgium do intended to perform an updated review of the public marine tools that can be described via the MarineVRE and, where appropriate, also via the LW ERIC catalogue.
3. ***Proposal on the categories (keywords) of the described resources****: It would be relevant to enrich this catalogue with additional VRE, particularly with regard to the analysis of genetic data (metaB and metaG).... (i)* Part of the work for 2023 would be to adopt or create a machine-actionable vocabulary to better describe and categorise the resources listed on the MarineVRE. (ii) It can be difficult to convince the owners of the resources to check and update their entries, but this is also on the roadmap for 2023.
4. ***Proposal****:* ***It is not clear the source or rationale behind these keywords.*** *Ideally, they should come from an existing controlled vocabulary, or form the basis of a new controlled vocabulary to provide consistent keywording. The images, whilst visually appealing do not contribute to the user experience. It would be more useful to present a short summary or descriptive text relating to the VRE. A free-text search for both Access points and Analysis resources would help users identify relevant resources.* Indeed, a review of the keywords is on the roadmap for 2023 and the suggestions for the look of the pages will be taken on board.
5. ***Proposal****:* ***It is not clear what the prioritisation or rationale for selecting/presenting these access points has been****. Do they represent a unique or significant resource? Some indication of the reason for inclusion would help the naive user. Additionally, an indication of "maturity" or the expected longevity and temporal and/or spatial range of the resource would help guide the user. this could be achieved through additional, controlled vocabulary-based keywords or short descriptive text alongside/in place of, each thumbnail.* A review of the scope of the MarineVRE is the starting point of its 2023 roadmap, and following that the website could be revamped following these suggestions.
6. ***Proposal****:* ***The same issues with keywords and thumbnails exist as described above.*** *In addition, the available keyword changes between the Access and Analyze section. Whilst this may reflect availability of resources, it is not clear why keywords are ordered and presented in a seemingly random order.* Response as above.

5. State of the Science stories

The research being carried out by the JRAs have been described in layman's terms in our State of the Science Stories pages. A description of what and why for each of these JRAs is provided. The products and outputs, patents, publications, data, public-outreach material, will all be linked to these pages as and when they are produced. It is envisaged that most of this material will only be produced in the final year of the project; links to the datasets and publications that have been added to the ASSEMBLE Plus collection are listed on these JRA stories.

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.



6. Recommendations for future virtual access programmes (NEW in this report)

The aim of the virtual access provision of ASSEMBLE Plus is explained in WP4/NA2:

1. To provide access to data created by our marine biological stations to give them a lifetime beyond the initiating project
2. To provide public and lasting access to data and publications created under ASSEMBLE Plus (JRA and TNA)
3. To specifically promote the uptake of long-term biological datasets gathered by our marine biological stations by the larger scientific community
4. To provide access to workflows and VREs for analysing our genomics observatory data (JRA1) and long-term ecological datasets, to allow them to be exploited more fully and long term

For these ambitions to be reached, it is necessary that data (whether JRA, TNA, or marine stations' outputs) are FAIR. The data must be archived for the long term and must be findable in a metadata catalogue (F,A). These metadata should be rich, complete, and standardised (I). In order for the data to actually be re-used by others, it is necessary that the data are interoperable – are in standard formats and use standard vocabularies – and have their provenance describe together with clear access information (I,R).

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.

To allow data to be accessed by and used within data analysis workflows and VREs, it also is necessary that these data are FAIR, with a strong emphasis on the I (interoperable, i.e. standardised data formats, data content, and metadata) and the R (re-useable, i.e. with provenance and clear access rights).

Clearly, the overriding theme is that FAIRness is a necessary step before data can be usefully provided to a scientific audience via a virtual access point. This is the approach we have been adopting in ASSEMBLE Plus. Here we list the most significant problems we have encountered while making our data FAIR, with recommendations on how to overcome some of these problems.

Making scientific outputs reusable

Providing FAIR access to existing datasets, include long-term biological datasets (points 1 and 3 above). As explained above, we have data records in the ASSEMBLE Plus collection that were already catalogued in IMIS by the various ASSEMBLE Plus partners. The diversity of data types is wide, and a good number are *long-term* (183 records) and *long-term biological* (161 records) in nature. It was the intention to work on FAIRifing these data via Task NA2.4. To this end, we performed a review of the metadata in those records to create recommendations as to what could be improved, and these were sent to the ASSEMBLE Plus partners. Unfortunately, only a few records were improved as a result. When questioned, it was reported that this was usually due to difficulties in being able to contact the original data record creator (be they a person, institution, or project). To address this problem, better data management practises at the data-creating institutes would be necessary.

- It should be clear to whom (the person, the lab, the project, the institute) ownership of the dataset belongs.
- Tracking of where the scientist/lab/institute has catalogued and archived (i.e. published) their datasets should be the norm, so that updates can be instituted by the labs themselves when desired.
- Having templates for the metadata that any dataset created in that lab/institute will ensure good metadata management, and would better allow for future updates to the metadata in the catalogue and/or archive where those data are published.
- A clear licence policy should be instituted and conformed to.

Providing FAIR access to all the data and publications produced under ASSEMBLE Plus (point 2 above). As a participant in the Open Data Pilot, it is required that the data and publications funded by ASSEMBLE Plus are published and made (immediately or eventually) Open Access, and for data additionally to be FAIR enough that being Open Access is useful (given that data which are not FAIR, are effectively not very findable, understandable, or re-useable). At a practical level this requires a pro-active data management by each data-creator/institute that begins when the experimental work begins: each stage of the data life-cycle should be described, standard vocabularies should be used for the metadata and in the data, and the data should be in standard file types with standard data formatting. This applies to all data that are going to be made public – which ideally should include raw data, quality controlled and finalised data, and data products. This was outlined in the ASSEMBLE Plus Data Management Plan.

However, doing pro-active data management still not a common practice in many institutes, and FAIR data and data management is something that is not commonly taught as part of a university course. This probably explains why the level of FAIRness of what we have been receiving from the TNA and

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.

JRA scientists has been very uneven. While everyone understood the principles, in practice it was a different matter: in particular, there was a misunderstanding that publishing data within their scientific publications (as images, tables, figures, or attached data) was *publishing data*, with no additional need to archive and catalogue the datasets separately. While in some cases, contacting the data creators (TNA and JRA) did then lead to data publications, in more cases this did not; we had no means to police/enforce this policy.

We had more success with the Open Access requirements on scientific publications, with the JRAs and with the TNA users: cases where TNA users published in closed access journals seem to have been so done because of the impact factor of the journal, and/or because the TNA work was only a small part of a project and the TNA user did not have the deciding vote in where the work was published. In any case, several such publications could be added to our ASSEMBLE Plus Open Repository.

Some collected recommendations:

- *Practical FAIR data management is a mandatory component of all university courses.* Within ASSEMBLE Plus, we have carried out one FAIR data management workshop and written extensive documentation, and have produced an online self-teaching course funded by ASSEMBLE Plus (FAIR Data for Marine Biologists).
- *Institutes invest in systems that make data management an integral and instinctive part of the scientific process, via training, development of clear policies, provision of data formatting templates, e-labs, etc.* An investigation of the existing e-labs, and how they help with the FAIR part of data management, would be a useful product from the EOSC community. It could also be initiated at the RI level.
- *Systems to help remind scientists to make their data open access once they have published their results,* and to link those publications to the metadata record of the related datasets, would also be useful.
- *Encouraging and enabling FAIR data management has to be a bottom-up process.* If individual scientists are not engaged, it will never be done fully or completely. This is something that can be instigated at the RI level, but has to be carried out with the active and willing participation of everyone at the university/institute/lab level.
- *If it is considered important that data are archived, catalogued, made interoperable, and made open access, then there should be requirements and incentives at the university/institute level, as well as at the researcher level.*

VREs

VREs work on standardised data – the data inputs need to conform to specified standards so that the VRE knows how to read them and knows which parts of the data to access. In ASSEMBLE Plus WP4, Task NA2.5, the intention was to develop/adopt VREs for both long-term biodiversity datasets in the ASSEMBLE Plus collection, as well as the genomics observatory (being also long-term biodiversity) data produced by JRA1.

The JRA1 data were included in the development of two VREs – via LifeWatch ERIC and the EOSC Life project with EMBRC – as reported in D4.9. It was only with these extra resources that it was possible to develop new workflows for these data: the development time for these projects was on scale of year(s). When the goal is to develop a new workflow for the analysis of genomics observatory data, it is a serious undertaking, one that requires dedicated ICT resources and dedicated scientific input. This is something that we could not have produced in ASSEMBLE Plus alone: we simply do not have the ICT

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.

resources to do this. But working together with LifeWatch/EOSC-Life/EMBRC, we could combine the strengths of all respective partner to produce something new and useful. Our *conclusion here is that RIs working together to tackle common scientific gaps and to create new scientific tools, is very much to be recommended*.

Similar developments were not possible for the data in the ASSEMBLE Plus long-term datasets collection, because these records were not FAIR enough, not enough of them provided open, machine-accessible, standardised data, for this to be a useful activity within *this* task. In principle this *is* a useful activity for others to pursue, but with a more specific focus on the most FAIR data in such a collection. In any case, those data within the collection that are published via EurOBIS can be accessed with the tools of EMODNet.

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.