ASSEMBLE

ASSOCIATION OF EUROPEAN MARINE BIOLOGICAL LABORATORIES EXPANDED

Acronym: ASSEMBLE Plus

Title: Association of European Marine Biological Laboratories Expanded

Grant Agreement: 730984

Deliverable [D32.3]

[3rd year access report on VA users]

[09][2020]

Lead parties for Deliverable: [VLIZ]
Due date of deliverable: M 36 [30/09/2020]
Actual submission date: M 36 [30/09/2020]

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: **48 months**

Deliverable number	D32.3
Deliverable title	3rd year access report on VA users
Submission due date	30/09/2020
Actual submission date	30/09/2020
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	☒
Websites, patent filling, etc.	☐
Ethics	☐
Open Research Data Pilot (ORDP)	☐
Demonstrator	☐
Other	☐

Dissemination Level

Public	☒
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)	NA
Version	1

Abstract

This is the first reporting on the virtual access to ASSEMBLE Plus resources: a reporting on the access of VA users up until the end of the 3rd year of the ASSEMBLE Plus project (the first two years have no associated reports). 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, publications, and virtual research environments (VREs) of ASSEMBLE Plus: what they are, who created them, who has accessed them, and recommendations for the future.

This report covers the timeframe from the start of ASSEMLE Plus until Sept 30, 2020. While this report is entitled "3rd year access report of VA users", the first and second reports were not written because it was initially considered that the VREs were to be the main component of virtual access, but they were only added in 2019. It was then decided to include all virtual access points under this work-package reporting.

2. The ASSEMBLE Plus virtual resources

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

- ASSEMBLE Plus created and collated datasets
- ASSEMBLE Plus created and collated publications
- Access to scientific data tools, aka Virtual Research Environments (VREs)
- The State of the Science Stories
- Access to the ASSEMBLE Plus Deliverables (those likely to be of public interest)

The first four of these access points are in fact created under the work-package WP4/NA2 (*Improving virtual access to marine biological stations data, information and knowledge*).

Note that we do not require user registration for those searching our catalogues of data and publications, or those accessing the VREs. Therefore we cannot include statistics of individual use of our virtual access.

2.1 ASSEMBLE Plus data

2.1.1 The data we collect

The data created under our JRA (Joint Research Activity) and TNA (Transnational Access) programmes are considered primary ASSEMBLE Plus data. We also collate data records from the ASSEMBLE Plus partners and marine stations. All data are archived and catalogued according to the ASSEMBLE Plus DMP (WP4, Task NA2.1); they are all linked together in the ASSEMBLE Plus collection in our metadata catalogue IMIS (Integrated Marine Information System).

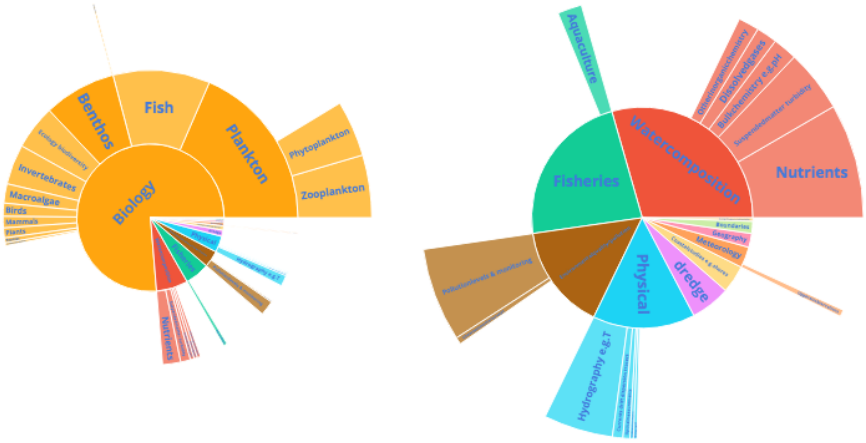
- All TNA users are requested to create a metadata record in the IMIS datasets catalogue, once their access has been completed. This record should describe the data they collected during the access, and these data should be uploaded to the MDA (Marine Data Archive), from where they can be linked to the 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.

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 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).
- The metadata records we collate from the ASSEMBLE Plus partners are those that had already been created in IMIS by these partners. As these data do not "belong" to ASSEMBLE Plus, the data licence is the decision of the data owner. However, these records are being made FAIR within WP4 (Task NA2.4), with a particular focus on the long-term biodiversity data and the genomics observatory data of JRA1, and we encourage open access licences.

By the end of the ASSEMBLE Plus project, the aim is to have all data produced under the project to be made FAIR via the ASSEMBLE Plus collection in IMIS, and the data to be saved for the long term in the MDA or another appropriate archive.

There are currently 526 metadata records in the ASSEMBLE Plus collection. These cover a wide range of scientific topics – biology is clearly the most popular, followed by water composition, fisheries, and physical datasets – as can be seen in the starburst charts below:

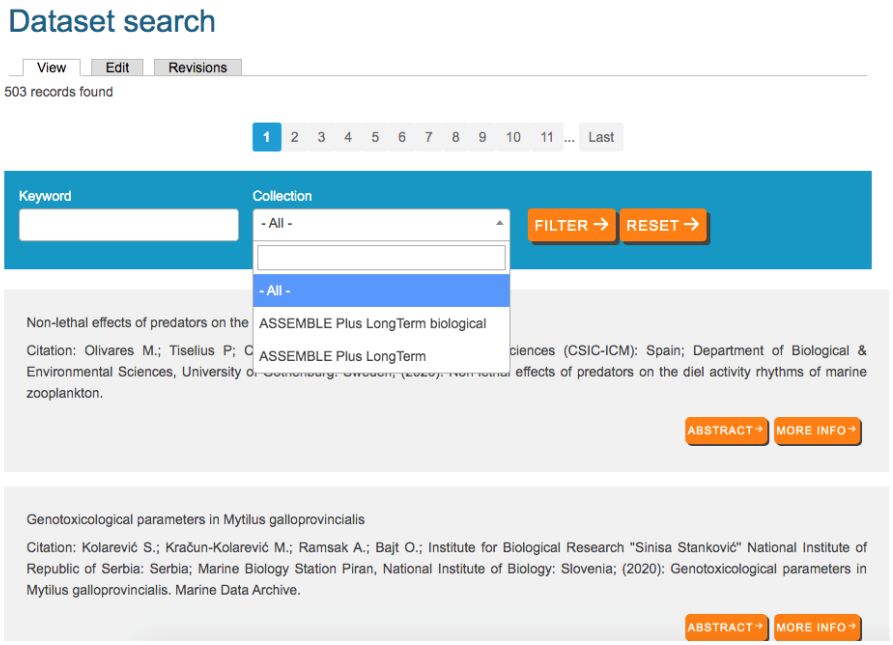


The bulk of this collection (500) are the records that come from the ASSEMBLE Plus partners; of these, we have 183 that are classified as *long-term*, where 161 of those *are long-term: biological*. We have 20 data records that have been created by TNA users since 2018, with a wide range of scientific topics. There is one data record from the JRA1 (Genomics Observatories) programme: the ARMS-MBON 2018 data. To this we will shortly add the data from the Ocean Sampling Day project, which is also a JRA1 activity: we are working together with MPI (via GFBio and its Pangea metadata catalogue) on the management of these data. While the collected sampling measurements and associated data of ARMS and OSD are archived in the MDA, the genetic sequences are archived in ENA (European Nucleotide Archive). Wherever the data are archived, they are always linked together via their IMIS or Pangea metadata record.

2.1.2 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). The intention for later in 2020 is to add further collection types to allow users to search only on data records from each ASSEMBLE Plus work-package, in particular the JRA1 to 5 and TNA activities.

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 partner's data portal, EurOBIS, etc).

Of the 526 records in the ASSEMBLE Plus collection, 260 have an open access licence (CC BY, CC 0, or "unrestricted/moratorium period"), although ~90 of those have no link to a dataset or data portal, and 180 records have no licence at all; improving this is work for Task NA2.4 for the final year of the project. Of the 20 records from TA users, five have CC BY[1] licences and the rest are under moratorium until publication.

2.1.3 Views on the data records

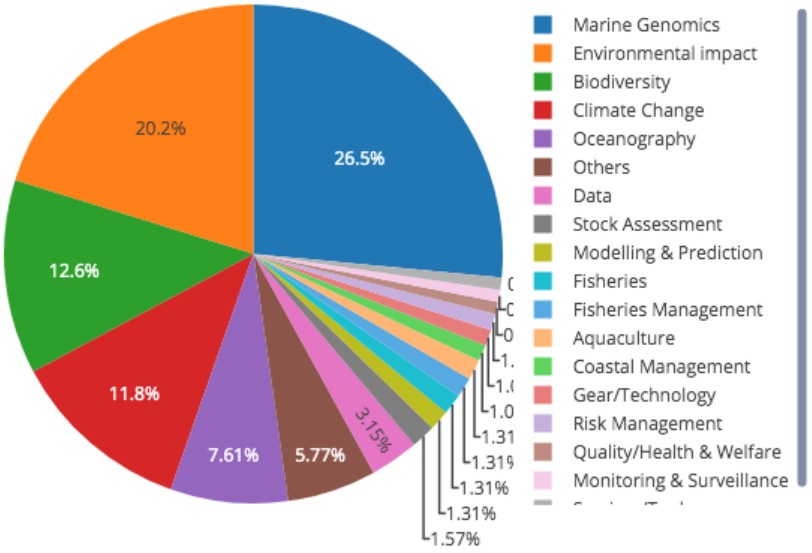
The landing page describing the ASSEMBLE Plus datasets collection, from where users can click onto the catalogue search page, has been available since March 2019. There have been 284 unique pageviews between then and Sept 30, 2020. North America and Europe form the largest origin of these visits, with only 10% from outside these areas. For the data records in the ASSEMBLE Plus collection, there were 442 views in 2019 (somewhat skewed by the ~100 views associated with the FAIR data management workshop, which focussed on these records), and 450 so far in 2020.

[1] CC BY is essentially an open access licence. This licence allows others to distribute, remix, tweak, and build upon the licensed work, including for commercial purposes, as long as the original author is credited

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.

2.2 ASSEMBLE Plus publications

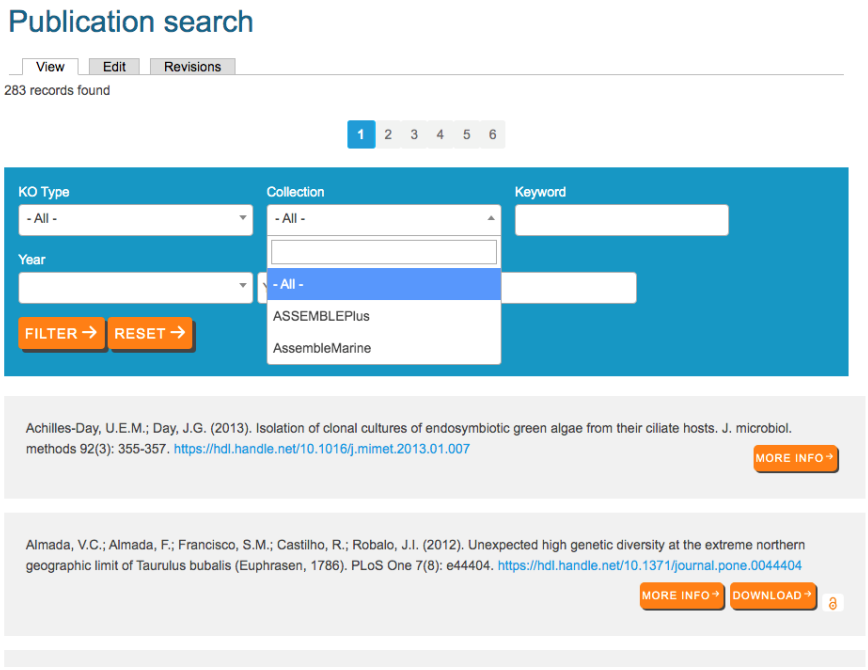
The guidelines for data created or collated by ASSEMBLE Plus apply also to publications (books, conference proceedings, scientific publications, or data papers): these should archived in the IMIS publications catalogue and made Open Access (Green or Gold publishing). We collect publications from the 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. In total there are 283 publications archived in the ASSEMBLE collection in IMIS, of which 31 have been created by the ASSEMBLE Plus project (18 from the TNA programme and the rest from JRAs). The publications cover a wide range of scientific topics – the largest field is marine genomics (mostly Assemble Marine), following by environmental impact and biodiversity – as can be seen in the pie-chart below:



2.2.1 Access to the publications

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. The intention for later in 2020 is to add further collection types for the work-package that the data were created under (JRA1 to 5 or TNA). Other classifications may also be added.

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.



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.

2.2.2 Views on 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 205 unique pageviews between then and Sept 30, 2020. North America and Europe also form the largest origin of these visits. For the publications in the ASSEMBLE Plus collection, there were 127 views in 2019, and 194 so far in 2020.

2.3 The Virtual Research Environments

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. The access point is the VREs page of the ASSEMBLE Plus site, from where we link to LifeWatch Belgium MarineVRE website where our tools are catalogued. This website uses keyword "tags" to help users search for tools: our tools have all been tagged with the keyword *ASSEMBLE Plus*, plus keywords appropriate to each particular tool.

2.3.1 Current VREs

The VREs that are currently provided are:

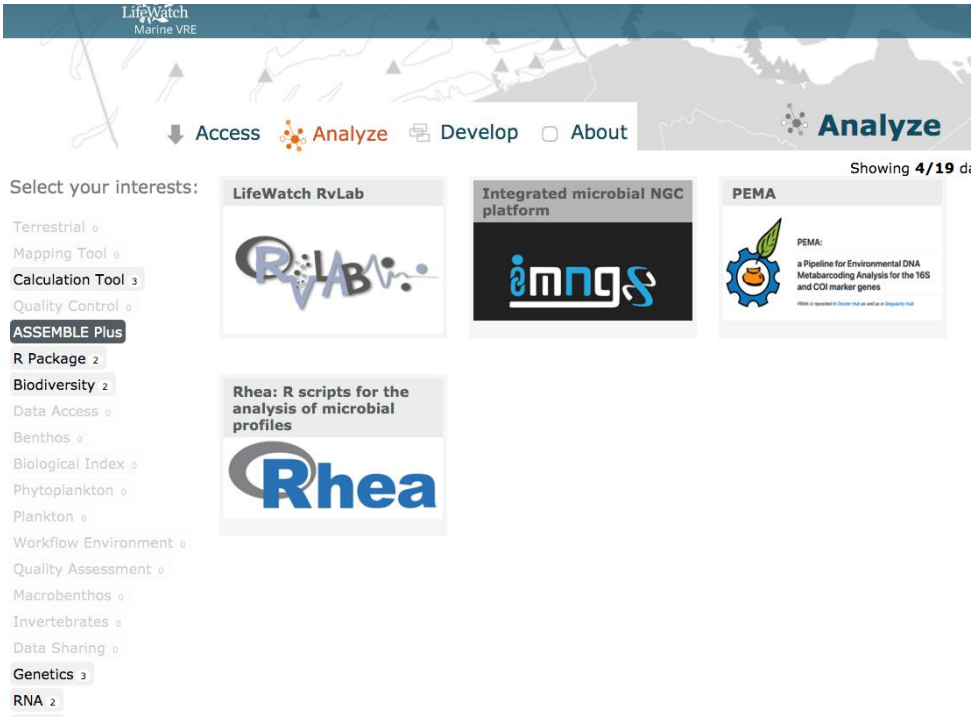
- Data access tools. At present these are a description of, and link to, the OSD 2014 github site and its Pangea record, and the search page of the ASSEMBLE Plus datasets collection in IMIS.

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 intention is to add links directly to the ARMS and OSD IMIS records directly, once these records are complete.

- Data analysis tools. Four tools are currently included.
    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).



2.3.2 Future developments

We are currently working on two new data analysis tools to be added to the current four:

- A DarwinCore explorer. DwC Archive is a popular format for biodiversity data, and our JRA1/genomics observatory data will be formatted into the DwC-A-OBIS event format for eventual publishing via EurOBIS. This format allows a variety of data measurements from individual sampling events to be linked to each other, with standard vocabularies and data formats to ensure that the data are interoperable (the I of FAIR). However, the DwC files are not very user-friendly, especially when they contain data from 100s of sampling events, with results from image analysis, visual observations, DNA sequence analysis, and more all linked together. VLIZ are therefore developing a tool to explore DwC files (starting with those from our ARMS and OSD programmes), to allow a user to browse the mass of data therein, visualise the geographic, temporal, and parameter space, and extract subsets of the data. This tool will be provided as a standalone script (via github), and it will also be incorporated (as a webservice) in the second new VRE, which we describe next.

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.

- A workflow for ARM data. LifeWatch have launched an initiative to create workflows to do scientific analysis on data of non-indigenous and invasive species (NIS). One of these workflows will be focussed on the ARMS data from JRA1. This workflow will start from the metadata record in IMIS and the data archived in the MDA. The workflow will also used two of the tools that are on the MarineVRE site: PEMA and RvLab. As well as the data and these tools, ASSEMBLE Plus are providing the detailed scientific use-case, the input and output data specifications for each step, details of what the user interface should provide, and finally we will contribute to the user testing of the workflow. This workflow will be available for general use in 2021, possibly sooner.

2.3.3 Views, hits, and uses of the VRE tools

The VREs are described on the ASSEMBLE Plus website and also on the MarineVRE website, and so we report separately on the page views for these two sites: 61 unique page views to the ASSEMBLE Plus VRE page since Sept 2019; and 6 (Rhea), 7 (IMNGS), 13 (PEMA), 6 (RvLab) unique pages views to the tools on the MarineVRE site since July 2020 (when tracking began).

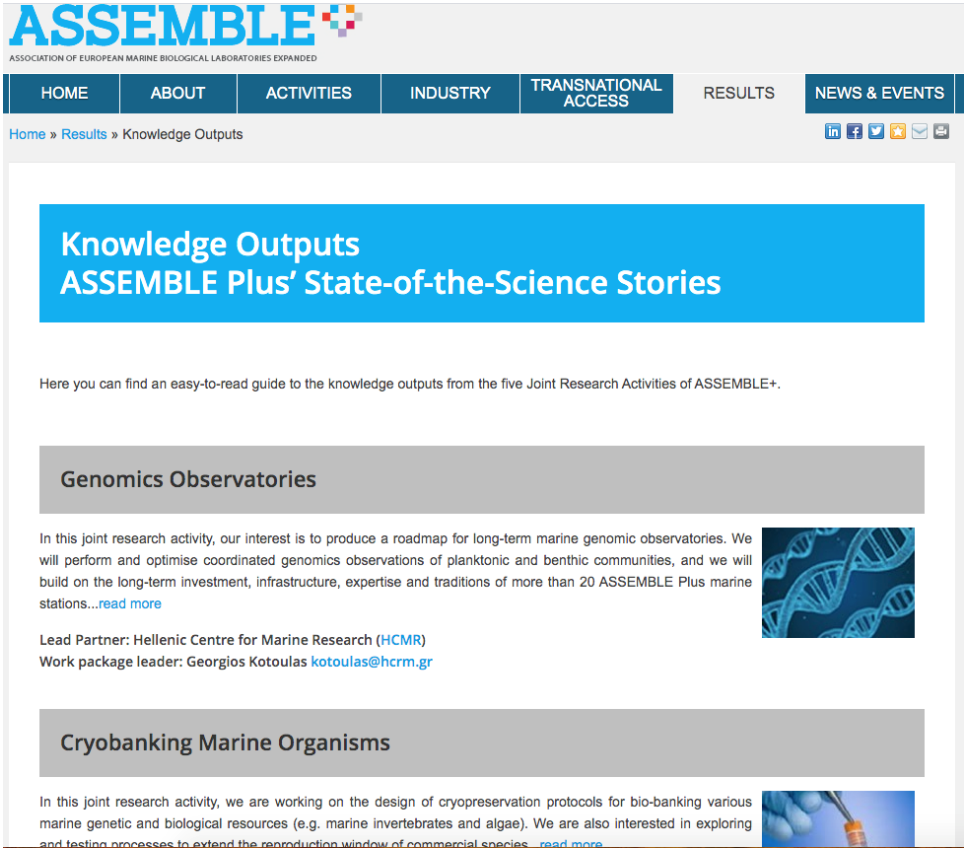
Because the four tools are all hosted on their own websites, we cannot track viewers moving from the MarineVRE site to the site of each individual tool. However, we can report on the traffic statistics produced by the individual hosting sites.

- PEMA: PEMA is available via Dockerhub, and there have been 729 pulls from there since it was placed on the site (12/04/2019). There have been 3 citations (and 18 tweets) 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 users so far in 2020. There are currently 7 citations of its 2016 publication in Biodiversity Data Journal (DOI 10.3897/bdj.4.e8357).
- IMNGS: In 2019 289 new users were added and 1699 queries were fielded. In 2020 there have been 211 new users with 1851 queries. In total there are 1453 users for IMNGS. There are currently 116 citations 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. We have now implemented a procedure to collect these fortnightly traffic reports to allow a more complete reporting next time. For now we can say that over the 2 weeks starting 10 Sept., there have been an average of 9 unique visitors per day. There are currently 94 citations of its 2017 publication in PeerJ (DOI 10.7717/peerj.2836).

2.4 The 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.

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. Recommendations for future virtual access

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, certain things are necessary. In the same order:

1. Providing meaningful public access to data requires that the data are FAIR. At a minimum the data must be archived for the long term and must be findable in a metadata catalogue. These metadata should be rich, complete, and standardised. In order for the data to be re-used by others, it is necessary that the data are interoperable – are in standard formats and use standard vocabularies – and are re-usable – provenance of the data is described, a data licence is clearly stated, and a data download link or contact information are provided.
2. Providing access to the JRA and TNA data requires the same as for point 1 – that the data are FAIR. This requires pro-active data management on the part of the data creators: to make

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 FAIR from when they are first created is much less work than making them FAIR after-the-fact.

3. To promote the uptake of long-term biological datasets, it is clearly also necessary that they are FAIR, as described in point 1.
4. To create or provide workflows and VREs, it is necessary that the data they should work on 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 also open access). In addition to this, if workflows are to be developed and not just adopted, it is necessary to have a good development team working closely with the scientists who will use the workflows.

An 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. We additionally comment on problems with and recommendations for user tracking and VREs.

3.1 Providing FAIR data and publications for public re-use

Adhering to the FAIR principles (findable, accessible, interoperable, and re-useable) is a pre-requisite to providing data that others can find and re-use, in particular via a virtual environment (i.e. online, via a portal or via a cloud setup). In our experience with ASSEMBLE Plus, there are a number of issues that make the process of FAIRifying existing or new data difficult.

Providing FAIR access to existing datasets, include long-term biological datasets (points 1 and 3 above). As explained in Sec. 2.1, we have data records in the ASSEMBLE Plus collection that were added by the ASSEMBLE Plus partners at some point in the past. We have further classified records that are *long-term* (183 records) and *long-term biological* (161 records) in nature. All of these records are being FAIRified via Task 2.4. For the long-term records, we are additionally asking the partners for extra information to help better understand which records belong in these collections, and also asking if they have any other records they would like to add to the collection.

To FAIRify metadata records for data collected in the past (from even as little as a few years ago) is a laborious process. To add the new necessary metadata fields (to conform with modern standard and vocabularies), to improve the descriptions, to add full information about the geographic, temporal, taxonomic, and parameter scope of the data, requires being able to contact the original data creator. If this person has since left the institute, this can be a near-impossible task. The process of requesting that the data become open access is often made difficult because it is unclear who own the data, it is unclear who can give permission to do so, or because the data "have gone missing". Trying to get hold of the necessary information is so time-consuming, that many institutes, labs, marine stations, etc. do not have the resources to pursue this.

While there are many FAIR projects running in Europe at present, often the focus is on current and future datasets. This leaves an enormous amount of existing data unattended. We believe that

1. *To overcome the barrier of "too little time, too few people", there should be resources provided specifically for FAIRifying existing datasets and data records. Incentives should be found to "reward" the time spent by individual scientists and institutes doing this 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.

2. *Institutes should have clear polices about who owns the data, about documenting its provenance, and what is to be done with the data after the data creator moves on.*
3. *When doing this work at a practical level, it is best to concentrate on those datasets that have the most re-use potential (for example, long-term data collections that can be used to study climate change) and for which FAIRification is an achievable possibility.* Within ASSEMBLE Plus, we will concentrate on the long-term collections, where the data owners can provide the necessary metadata and can assign the data with a open access licence.

Providing FAIR access to all the data and publications produced under ASSEMBLE Plus (point 2 above). We require that all data and knowledge produced under ASSEMBLE Plus are made FAIR and (immediately or eventually) open access. At a practical level this requires a pro-active data management that begins when the project itself begins: the data should be completely described (all processing steps should be documented for full provenance tracing to be possible), 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. While it is common to consider finalised data and data products to be public, raw data is often forgotten.

Doing pro-active data management still not a common practice in many institutes, and 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 scientists (who are largely PhD students and early post-docs), has been very uneven. In fact, by requiring the TNA scientists to do some FAIR data management, we are in fact teaching them how to do FAIR data management!

Another problem we have had is that while we can require that the TNA scientists archive their data in the MDA, catalogue them in IMIS, and provide us with their open access publications afterwards, we cannot enforce that. Of the ~150 TNA projects that have run so far, we have only ~20 IMIS records and ~20 publications, and these are largely not the same 20. This, despite numerous emails, reminders, and detailed instructions on the website.

We recommend that

1. *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 may consider whether an additional online workshop would be suitable for the final year.
2. *Institutes invest in systems that make data management an integral and instinctive part of the scientific process, e.g. via user-friendly and useful e-labs. 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.*
3. *Systems to help remind scientists to make their data open access once they have published there results, and to link those publications to the metadata record of the related datasets, would also be useful.*
4. *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 best 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.*

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.

5. *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, not just at the researcher level.*

3.2 User access and tracking

Tracking of the page views or hits to our any single website is relatively easy. However, tracking user trails across websites is impossible with the normal tracking tools that are available (and while honouring GDPR). As our ASSEMBLE Plus VREs are a collection of existing tools, provided by our partners and hosted on their own websites, but catalogued on a LifeWatch Belgium website, it is impossible to track users from the potential beginning (the ASSEMBLE Plus site) to the potential end (a dataset loaded into a VRE).

Tracking the of use of your data resources and provenance tracing of your data as they are accessed, analysed, and re-uploaded to portals or archives, is a topic of discussion among the EOSC community. Hopefully, enactable recommendations will be an outcome that we can adopt in the future.

3.3 Creating VREs for the ASSEMBLE Plus data

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 data points to access. In ASSEMBLE Plus WP4, Task NA2.5, we state that we will "set up a virtual platform for data analysis, where the standardised data from long-term biodiversity and genomics observatories will be connected to a virtual analysis portal (based on R webservices and Taverna workflow systems) that enables used to selected from available datasets and perform predefined processing and analysis workflows."

It has become clear that to create VREs for the long-term biodiversity data in the ASSEMBLE Plus data collection is not going to be possible. The FAIRness of these data – in particular their interoperability – is still being worked on (for reasons explained in Sec. 3.1). Many of these datasets are also still not open access. It is true that many of our partners do publish data via EurOBIS, and the data formats (DwC-A) *are* then standardised and the data are open access, but there are already many existing data exploration and analysis tools that are provided by EMODNet and by the projects under which these data were gathered (e.g. LifeWatch). It was therefore decided to concentrate on the genomics observatory data produced by JRA1 via the OSD and ARMS projects for Task NA2.5.

As mentioned in Sec. 2.3.2, we are working with LifeWatch to create workflows to do scientific analysis on data from NIS projects. The data of ARMS form the scientific input to one of these workflows. The workflow environment will incorporate numerous tools that our scientists use on ARMS data – PEMA, RvLab and a DwC explorer as mentioned before, an image analysis tool, and more. A sophisticated user interface will allow users to load data, explore the data, select tasks to link together in simultaneous "runs", inspect the output, and much more. It will fully track what is done to the data, so that provenance tracking is automatically taken care of.

ASSEMBLE Plus has provided extensive input to the design and development of this workflow: over the course of several months we have discussed the scientific requirements, the tools to use, the type of data to read in, produce, and read out, the requirements for the UI. The LifeWatch ICT team has been working since 2019 to develop the workflow environment and the UI, has turned some of our tools into webservices, has created a blockchain to follow the data into, through, and out of the workflow, and much more. *All of this is to say that when the goal is to develop a new workflow for the analysis*

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.

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 background to do this. But working together with LifeWatch, we can combine the strength of ASSEMBLE Plus (and here we should include EMBRC, our parent ERIC) – with its marine stations, its genomics observatory scientists, labs, expertise – with the strength of LifeWatch – with its extensive expertise in e-science – 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*.

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.