ASSEMBLE

ASSOCIATION OF EUROPEAN MARINE BIOLOGICAL LABORATORIES EXPANDED

Acronym: ASSEMBLE Plus

Title: Association of European Marine Biological Laboratories Expanded

Grant Agreement: 730984

Deliverable D32.7

3rd year assessment report on VA users by international review panel

September 2020

Lead parties for Deliverable: VLIZ
Due date of deliverable: M 36 [30/09/2020]
Actual submission date: M 36 [19/10/2020]

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730984. 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	D32.7
Deliverable title	3rd year assessment report on VA users by international review panel
Submission due date	30/09/2020
Actual submission date	19/10/2020
WP number & title	WP32 VA Virtual Access
WP Lead Beneficiary	VLIZ
Participants (names & institutions)	Davide Di Cioccio (ASSEMBLE Plus WP 2 participant), EMBRC ERIC; Erwan Corre (EMBRC partner), Station Biologique de Roscoff (SBR); Cymon J. Cox (ASSEMBLE Plus partner), Centro de Ciências do Mar (CCMAR); Juan Miguel González-Aranda (LifeWatch ERIC), LifeWatch HQ, Seville, Spain; Mark Hoebeke (EMBRC partner), Station Biologique de Roscoff (SBR); Dan Lear (ASSEMBLE Plus partner), Marine Biological Association (MBA)

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 730984. 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)	Erwan Corre, Cymon J. Cox, Juan Miguel González-Aranda, Mark Hoebeke, Dan Lear, Davide Di Cioccio
Editor(s)	Davide Di Cioccio
Version	1

Abstract

This is the first assessment of the international panel on the Virtual Access provision of ASSEMBLE Plus (from the beginning of the project to the end of the 3rd year). The deliverable D32.3 (submitted on September 30, 2020) forms the main input to this panel.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730984. 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 730984. 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 an assessment of the Virtual Access that is provided by ASSEMBLE Plus via its website www.assembleplus.eu. Deliverable D32.3 (submitted on Sept 30, 2020) is the main input to the international panel that is conducting this assessment. Both of these reports cover the timeframe from the start of ASSEMBLE Plus until Sept 30, 2020.

The panel consist of the following:

- From ASSEMBLE Plus, as a work-package participant or from one of the partners:
  - Cymon J. Cox (ASSEMBLE Plus partner), Centro de Ciências do Mar (CCMAR)
  - Dan Lear (ASSEMBLE Plus partner), Marine Biological Association (MBA)
  - Davide Di Cioccio (ASSEMBLE Plus WP2 participant, EMBRC-ERIC)
- From outside of ASSEMBLE Plus:
  - Erwan Corre (EMBRC partner), Station Biologique de Roscoff (SBR)
  - Juan Miguel González-Aranda (LifeWatch ERIC), LifeWatch HQ, Seville, Spain
  - Mark Hoebeke (EMBRC partner), Station Biologique de Roscoff (SBR)

2. Evaluation of Virtual Access to ASSEMBLE Plus data

2.1. Scope of collected data

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.

2.2. The (meta)data in the ASSEMBLE Plus portal

The report also details how to search for data using the dedicated search page. This page allows to use a single keyword either to search through the whole ASSEMBLE Plus collection, or in one of two sub-collections (Long-Term or Long-Term Biological). Search results are lists of meta-data summaries which give access to more complete information.

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

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730984. 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.

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.

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).

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 search and retrieval possibilities. The report does not mention whether this has been considered and planned or discarded or postponed.

2.3. Recommendations related to data FAIRification

The recommendations very accurately pinpoint the main issues still outstanding regarding widespread adoption of best practices in data FAIRIfication. The actual recommendations addressing these issues are also very sound and both tackle the technical aspects, as well as the more institutional ones, and are very complementary (i.e. trying to make FAIRification a bottom-up process and at the same time, making FAIRification a requirement at an institutional level). Overall, the recommendations being to some extent, the outcome of experience gathered throughout ASSEMBLE Plus regarding data management, they may be considered as real-world based and as such practical guidelines, both for the remainder of the ASSEMBLE Plus project, and for other related projects.

3. Evaluation of Virtual Access to scientific data tools, aka Virtual Research Environments (VREs)

Panellist: Erwan CORRE (EMBRC partner, Station Biologique de Roscoff (SBR))

3.1 General description

The report describes a list of VRE developed in the framework of the ASSEMBLE Plus project and allowing access to the datasets made available in the project. It also describes access to analysis environments developed by project partners (inside and outside the ASSEMBLE Plus) that can be used by ASSEMBLE partners to process part of their data.

The search interface of the tools is very functional and intuitive. The thumbnails and texts are very readable. It is possible to filter the searches using keywords.

Proposal: It could be relevant to order these keywords either alphabetically or by classes.

3.2 Access portals

Twenty four access portals to data sets are available. They can correspond to very general data (WORMS, MarineRegions, OBIS, GBIF, etc...) but also to very specific data (polychaetes features catalogues, data from the 2014/2015 OSD campaign).

Proposal: 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

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730984. 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.

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.).

3.3 Virtual research analysis environments

Accessible under the "Analysis tab", nineteen environments are available. Most of them correspond to online analysis tools (RvLab, imngs, BioVeL workflows, LifeWatch ERIC analysis portals and elabs) and are not specifically linked to marine data. There are also some R-shiny applications (LifeWatch ERIC Data Explorer, MarineSPEED) for marine data visualization and mining.

On the other hand, some of these tools correspond to standalone applications to be executed on a workstation (PlanktonToolbox, PEMA, AMBI, GISMO) or to R packages (ETN R, worms R, Rhea R packages) and could be considered as Virtual Research Environment (https://en.wikipedia.org/wiki/Virtual_research_environment). They should be classified by another category (softwares; R-packages)

Some tools seem old and not maintained (ISS-Phyto (last update 2011). Some recording systems are non-functional (ISS-Phyto, Access to Rvlab (LifeWatch Greece) connection with ORCID doesn't seem to be functional).

Proposal: It would be relevant to enrich this catalogue with additional VRE, particularly with regard to the analysis of genetic data (metaB and metaG).

For the analysis of environmental data many pipelines are available in the European and national Galaxy instances (https://usegalaxy.eu/, https://usegalaxy.fr/,http://galaxy4metab.sb-roscoff.fr/, etc.). An instance specifically dedicated to environmental data (https://github.com/65MO/Galaxy-E) is hosted on the European instance (https://ecology.usegalaxy.eu/). EBI has also done a great deal of work on data structuring to provide a single portal for collecting and analysing metagenomic and metagenomic data (https://www.ebi.ac.uk/metagenomics/). The data collected in MGnify are now referenced in GBIF (DOI: 10.3897/biss.3.37036).

It could be interesting to get closer to the Biodiversity working group currently being structured within ELIXIR infrastructure to promote the interoperability of the catalogues collected by the two groups in order to limit maintenance efforts.

For example, the biotools catalogue https://bio.tools/ offers very exhaustive access to a very large number of tools and general or specific ERVs. (for the sailor, we can cite: Marine Metagenomics Portal (https://mmp.sfb.uit.no/), METdb (http://metdb.sb-roscoff.fr/metdb/), OGA/OBA (http://tara-oceans.mio.osupytheas.fr/ocean-gene-atlas/), genoma (http://bioinfo.szn.it/genoma/) etc. In addition, the Workflow-hub (https://workflowhub.eu/), currently being developed as part of the EOSC life project, aims to eventually list a very broad spectrum of workflows (CWL, snakemake, nextflow, galaxy) for data analysis.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730984. 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.

Panellist: Dan Lear (ASSEMBLE Plus Partner), Marine Biological Association (MBA)

3.4 General description

The VRE list on the website presents a clear, image-driven point of access to the VREs for browsing and further interrogation. There is a "keywording" filter in the left-hand column.

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.

Access portals: 24 access portals are presented using the thumbnail and keyword approach described above.

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

3.5 Virtual research analysis environments

Nineteen analysis environments are presented to the user using the same interface as described above. The detail page for each resource provides a useful summary and overview.

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.

The detail page provides a good overview for the user, but in order to get to this point the user needs to have a deeper understanding of the specific resource as there is insufficient information on the summary/browse page. Additional search/browse tools would help the user experience.

Additionally, there is no indication of when the resources were last accessed or updated. The date of submission, could be augmented with an "updated" or "accessed" date to improve user confidence.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730984. 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.

Panellist: Cymon J. Cox (ASSEMBLE Plus partner), Centro de Ciências do Mar (CCMAR)

3.6 Recommendations for future virtual access

As noted, data that is FAIR is a necessary prerequisite before data can be usefully provided to users. However, currently 180 of 526 IMIS records do not have a licence indicated, thereby severely limiting the usefulness of the data, and potentially reducing the overall quality of the IMIS repository. Moreover, attempts to retroactively FAIRify the data records have proven difficult.

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.

Regrettably, of the ~150 Trans-National Access projects only ~20 have deposited (meta)data with IMIS and public data repositories. Despite project proposals requiring the presentation of a Data Management Plan, few projects fulfil these obligations.

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.

Panellist: Juan Miguel González-Aranda (LifeWatch ERIC), LifeWatch HQ, Seville, Spain.

3.7 General description

In addition, the report makes us aware that new innovative progresses are being implemented in relation to the Workflows & associated e-tools that are being co-designed by and co-deployed on initial versions of LifeWatch ERIC **Tesseract VRE ARMS** (Autonomous Reef Monitoring Structures) workflow. This consists of a data-chaining pipeline that uses data from a network of placed in the vicinity of marine stations and Long-term Ecological Research sites (LTER) to assess the status of, and changes in, hard bottom communities of near coast environments, using genetic methods. The ASSEMBLE Plus scientists have provided a complete step-by-step tutorial with description of the inputs and outputs, the dependencies, and the codes, in each step of the ARMS workflow. Evolutive versions (DevOps) of the corresponding mock-up are already available.

In fact, LifeWatch ERIC e-Biodiversity (both Scientists & ICT teams) are currently working in the final phase of definition of all the wrappers provided by researchers: [1] In review process: PhotoQuad, RVLab, Produce various output formats; [2] Already deployed wrappers on: Web Get Masterfile; Filter & Select; Choose and Parametrise; Web Get Images; Web Get Sequences; PEMA; BOLD; Unify / Extract from different OTU-sources; Taxon Check WoRMS; Invasive Check WRIMS.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730984. 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.