WP4 - NA2 Improving virtual access to marine biological stations data, information and knowledge

AssemblePlus General Assemblee
10-12th October 2018
Objectives

- Improve virtual access to marine biological stations data, information and knowledge by:
  - Shaping and Assemble Plus community data management plan
  - Improving interoperability with related e-infrastructures
  - Installing an on-line system to provide access to marine biological stations information
  - Improving virtual access to data resources from Assemble Plus stations, with a specific focus on genomic and long term biodiversity data series
  - Offering a virtual analysis platform as a service for genomic and biodiversity data analysis
<table>
<thead>
<tr>
<th>Work package number</th>
<th>NA2</th>
<th>Lead beneficiary</th>
<th>VLIZ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work package title</td>
<td>Improving virtual access to marine biological stations data, information and knowledge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Participant number</td>
<td>17</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Short name of participant</td>
<td>VLIZ</td>
<td>HCMR</td>
<td>MBA</td>
</tr>
<tr>
<td>Person/months per participant</td>
<td>34.5</td>
<td>18</td>
<td>8</td>
</tr>
<tr>
<td>Participant number</td>
<td>14</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>Short name of participant</td>
<td>CCMAR</td>
<td>UPV/EHU</td>
<td>UH</td>
</tr>
<tr>
<td>Person/months per participant</td>
<td>3</td>
<td>2</td>
<td>1.5</td>
</tr>
<tr>
<td>Participant number</td>
<td>23</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Short name of participant</td>
<td>MSS</td>
<td>NIB</td>
<td>IOPAN</td>
</tr>
</tbody>
</table>

NA2.3: Multidisciplinary datasets inventory and archival: UPMC(x3), NIB, NIOZ, UH, IOPAN, UG, NUIG, UGOT, UVOG, EHU, HCMR, HUJI, SZN(x2), UIB, CCMAR (x2), AWI, MPIMM, VLIZ, SAMS, USTAN, MBA, NERC, MSS

NA2.4: LT biodiversity data series & genomic data: UPMC, UH, EHU, HCMR, HUJI, SZN, CCMAR, AWI, VLIZ, SAMS, MBA, NERC, MSS

Objectives
- Improve virtual access to marine biological stations data:
  - Shaping and implementing an on-line system to provide access to marine biological stations information;
  - Improving interoperability with related e-infrastructures;
  - Installing an on-line system to provide access to marine biological stations information;
  - Improving virtual access to data resources from ASSEMBLE Plus stations, with a specific focus on genomic and long term biodiversity data series;
  - Offering a virtual analysis platform as a service for genomic and biodiversity data analysis.
Task NA2.1. Design Data Management Plan and interoperability with related e-infrastructures

Task Leader: VLIZ; participants HCMR, MBA, MPI.

Deliverables: D NA2.1a,b, c: Data Management Plan (M6, M24, M48)

A data management plan (DMP) will be drafted as a reference document for the ASSEMBLE Plus community. This document will outline how the research data collected or generated within the JRAs and as part of the TA will be handled during the ASSEMBLE Plus project and will remain in force after the end of the project.

The DMP will build on current best practices at the national, EU and international levels, and recommendations and guidance from the ICSU World Data System (ICSU-WDS), Research Data Alliance (RDA) and the International Oceanographic Data Exchange (IODE). The DMP will define the methodology and standards to be applied, whether and how these data will be shared and/or made open, and how they will be curated and preserved.
What is a DMP

- **What is a data management plan?**
  - Document defining what data will be generated, how data will be curated and preserved and whether and how data will be shared or made available.

- **Why do we need it?**
  - H2020 requirements
  - Essential part of planning research activities
What is a DMP

• Elements of a Data Management Plan
  • Framework and scope
  • Data Collection
  • Documentation and Metadata
  • Storage and Backup
  • Selection and Preservation
  • Data Access, Sharing and Rights
  • Responsibilities and Resources
Workshop: The ASSEMBLE Plus Data Management Plan, 1-2 March 2018, Oostende, Belgium by Klaas Deneudt

On 1 and 2 March 2018, the ASSEMBLE Plus Data Management Plan Workshop was held at the Flanders Marine Institute (VLIZ), Oostende. The aim of this workshop was to discuss the Data Management Plan (DMP), the first version of which was completed by 30 March.

Participants included partners of Work Package 4 – Networking Activity 2 (NA2), “Improving virtual access to marine biological stations data, information, and knowledge”, and the Work Package leaders of the Joint Research Activities (JRA).

Experts from related initiatives and partnerships were also invited: LifeWatch, ELIXIR, European Marine Observation and Data Network Biology (EMODNet Biology), the European Marine Biological Resource Centre Working Group on E-Infrastructure (EMBRC WGEI), Digital Curation Centre (DCC), International Oceanographic Data Exchange network (IODE) and Ocean Biogeographic Information System (OBIS).

The meeting commenced with presentations framing the ASSEMBLE Plus project and NA2. Specific focus was placed on the design of the DMP and its interoperability with related e-infrastructure (NA2.1). A presentation followed on what Horizon 2020 “Open Research Pilot” data means, and what its Findable, Accessible, Interoperable and Reusable (FAIR) data policy encompasses. The DMP of EMBRC was presented as a side-by-side comparison to the information required for the ASSEMBLE Plus DMP.

A discussion on the most critical points of the DMP was carried out, including: the types of data expected, data file sizes, where data will be stored, open access (or not), and clarifications on the EU template, along with four JRA presentations. The management of TA project data was discussed, as was the format of the spreadsheet that the JRAs need to fill in with information that will allow an ASSEMBLE Plus data inventory to be created.

On the second day the discussion points continued, and JRA representatives filled in a template form to provide specific information that is required for the DMP. Overall, the meeting was successful due to the gathering of information for the DMP.

Important future action points will be the management of very large data files, the Open Access and the FAIR policy in relation to TA project data, whether more harmonisation (file types, databases used, naming conventions etc.) is possible, and the storage of ASSEMBLE Plus data products in the first Flemish marine data bank (IMIS) and the Marine Data Archive.
Deliverable D4.1

ASSEMBLE
ASSOCIATION OF EUROPEAN MARINE BIOLOGICAL LABORATORIES EXPANDED

Acronym: ASSEMBLE Plus
Title: Association of European Marine Biological Laboratories Expanded
Grant Agreement: 730084

Deliverable D4.1
ASSEMBLE Plus Data Management Plan
03/2018

Lead parties for Deliverable: VUZ
Due date of deliverable: M5
Actual submission date: 06/04/2018

All rights reserved
This document may not be copied, reproduced or modified in whole or in part for any purpose without the written permission from the ASSEMBLE Plus Consortium. In addition to such written permission to copy, reproduce or modify this document in whole or part, an acknowledgement of the authors of the document and all applicable portions of the copyright must be clearly referenced.
Data Management Plan for the Transnational Access programme

DMP’s for TNA’s (suggested H2020 template)

1. Data Summary
   • Explanation of the data source and format

2. How will you make your data FAIR (findable, accessible, inter-operable, re-useable)?
   • Discoverability, meta data, keywords, versioning

3. What Access (Open, Partial, Closed) will your data have?
   • If some of the data is not OA, explain why and how access could be granted
   • What software is required to access them?

4. How will you make your data inter-operable?
   • Vocabularies used for the meta data

5. How will the data be made re-usable?
   • License, embargo periods, QA processes, how useable is the data by other parties?

6. Allocation of costs and Data security
   • Mostly a formality

7. Ethics
   • Could be taken from their TA application

8. Publications
   • In an OA journal?

Approach

• From Call 2 onwards:
  • A short DMP section is part of the application form they fill in
  • A fuller DMP will be sent to each accepted applicant after they have signed their UAC

• Call 1: retrospective request after the Activity Report (13 sent so far, 3 returned)
Data Management Plan for the Joint Research Activities

DMP’s for JRA’s

JRA1 Genomics observatories
JRA2 Cryobanking of Marine Organisms
JRA3 Functional genomics
JRA4 Development of instrumentation
JRA5 Scientific diving

- What is the JRA about?
- What are the planned activities that will generate data?
- What types and formats of data will be generated/collected?
- Where are the data being generated and stored?
- To whom might the data be useful?
- Are there any reasons why open by default principles do not apply?
Task NA2.2 Creation of the Knowledge Transfer Platform

**Task Leader: VLIZ.**

An online Knowledge Transfer Platform will be implemented as part of the ASSEMBLE Plus web portal. This online system will become a central tool for provision with relevant stand exchange of information of researchers and research groups, their expertise and scientific output in terms of publications keholder groups. The online system will include two modules:

**Knowledge Outputs module** will provide the infrastructure to provide users with access to an unrestricted, user friendly and searchable database of Knowledge Outputs which may be relevant to them, as provided for by NA3. Outputs from TA and from research developed at each marine station will be submitted through a customized form which includes

**Publications module** will be integrated as a publication catalogue and archive in the web portal. This module will serve as an open archive for publications resulting from the ASSEMBLE Plus project. The use of the Knowledge Transfer Platform will be registered and reported using in-built tools (D NA2.4).
Knowledge Transfer Platform

Explore the knowledge outputs of ASSEMBLE Plus

The knowledge outputs of ASSEMBLE Plus can be accessed via the Knowledge Transfer Platform. These knowledge outputs arise from all ASSEMBLE Plus activities – the research of our JRA participants, research results arising from ASSEMBLE Plus collaborations, publications arising from the Access programme, or from associated activities, and so on. These knowledge outputs are organised by the sectors listed below, and within each sector they are categorised as:

- publications: books, scientific publications, technical documentation, protocols...
- data: datasets created by research conducted within ASSEMBLE Plus
- activities: training activities and material, learning modules, ASSEMBLE Plus science meeting outputs...

Click on the links below to access these knowledge outputs:

Aquaculture and Fisheries

- Aquaculture
- Fisheries

Marine Technology and Energy

- Marine technology
- Marine energy

Technology and instrumentation

- Scientific instrumentation
- Scientific diving
- Cryobanking

Marine Sciences

- Genomic observatories
- Functional genomics

Climate Change and Environmental Management

- Climate change
- Environmental management
Marine Genomics

Genomics observatories are research facilities that produce genomic-level biodiversity observations that are contextualised, localised geographically, with activities performed in compliance with international data acquisition standards, and for which genomic data acquisition is a long-term activity, monitoring the stream of genetic variations in marine ecosystems. JRA1 will work on the establishment of Genomic observatories across the AssemblePlus partnership.

The Lund declaration. Europe must focus on the grand challenges of our time

Published works found:


Unexpected high genetic diversity at the extreme northern geographic limit of Taurus bubalis (Euphrasen, 1786)


In: PloS One. Public Library of Science: San Francisco; ISSN 1932-6203

Author(s)
Vlaams Instituut voor de Zee: Open access 306923 [download pdf]

Keywords
Biodiversity, Marine Genomics; Marine Sciences; Scientific Community, Scientific Publication, Marine

Project
- Association of European marine biological laboratories

Authors
- Almada, V.C.
- Almada, F.
- Francisco, S.M.
- Castilho, R.
- Robalo, J.J.
EMBRC Open Access Publications Archive

Article 29.2 of the Model Grant Agreement sets out detailed legal requirements on open access to scientific publications: under Horizon 2020, each beneficiary must ensure open access to all peer-reviewed scientific publications relating to its results.

1. Self-archiving / 'green' OA: – the author, or a representative, archives (deposits) the published article or the final peer-reviewed manuscript in an online repository before, at the same time as, or after publication. Some publishers request that open access be granted only after an embargo period has elapsed.

2. Open access publishing / 'gold' OA’s - an article is immediately published in open access mode. In this model, the payment of publication costs is shifted away from subscribing readers.

Purpose
- A central location for all publications produced under the ASSEMBLE Plus and ASSEMBLE Marine programmes
- Will contain all OA publications and “self-archived”

Expected contributors
- TA users
- JRA scientists
- Potentially from other scientific activities related to A+
Task NA2.3 Set up virtual open access entry point to data resources (1/3)

**Task Leader:** VLIZ; participants: HCMR, MBA, MPI (and all other partners).

**Deliverables:** D NA2.2a, b, c: Virtual Access hits to ASSEMBLE Plus data resources (M24, M36, M48)

The central entry point that allows users to access common ASSEMBLE Plus data resources (from JRAs, NA and TA) will be created based on existing components.

The datasets module of the Integrated Marine Information System will be plugged into the ASSEMBLE Plus website and will allow users to query and select data resources based on a set of relevant search criteria. In order to create a digital inventory of the data resources available at the ASSEMBLE Plus stations, each of the partners will describe their datasets in a digital catalogue, using an online form. The scope, characteristics, state and accessibility of the data will be described following common standardized formats compliant to ISO19115 metadata standard. In case the data is already accessible through local online databases a web link to the existing interfaces will be included in the dataset description.
Marine biological stations dataset inventory

- In IMIS information system
  - 384 datasets from VLIZ
  - 201 datasets from UGENT
  - 106 datasets from HCMR
  - 78 datasets from AWI
  - 48 datasets from NERC
  - 25 datasets from IMAR
  - 22 datasets from MBA
  - 15 datasets from IOPAN
  - 13 datasets from NIB
  - 13 datasets from UG
  - 11 datasets from SBR
  - 8 datasets from CCMAR
  - 7 datasets from SAMS
  - 7 datasets from SZN
  - 6 datasets from IIMC
  - 5 datasets from OOV
  - 3 datasets from CIIMAR
  - 3 datasets from HBS
  - 3 datasets from OOB
  - 2 datasets from SLC
  - 2 datasets from UIB
  - Total 963 datasets

- After additional enquiry
  - 18 datasets from SZN
  - 15 datasets from ARI
  - 14 datasets from PIE-UPV
  - 12 datasets from IUI
  - 6 datasets from SAMS
  - 5 datasets from NIOZ-CNSI
  - 5 datasets from SLC
  - 4 datasets from TZS
  - 3 datasets from HCMR
  - 3 datasets from MSS
  - 2 datasets from SOI
  - 1 dataset from ECIMAT
  - Total 88 datasets

- 12 stations have responded as requested
- 6 stations have responded not to have the intention to describe additional datasets
- 6 stations have not answered to the request
- 5 marine stations need more time
### Dataset search

<table>
<thead>
<tr>
<th>Sector</th>
<th>End User</th>
<th>KO Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>- All -</td>
<td>- All -</td>
<td>- All -</td>
</tr>
</tbody>
</table>

**Keyword**

<table>
<thead>
<tr>
<th>APPLY</th>
<th>RESET</th>
</tr>
</thead>
</table>

895 records found

1 2 3 4 5 6 7 8 9 10 11 ... Last

---


---

**Report on the Danish Oceanographical Expeditions 1808-1910 to the Mediterranean and adjacent seas - Introduction**


---


Task NA2.3 Set up virtual open access entry point to data resources (2/3)

Task Leader: VLIZ; participants: HCMR, MBA, MPI (and all other partners).
Deliverables: D NA2.2a, b, c: Virtual Access hits to ASSEMBLE Plus data resources (M24, M36, M48)

Data will be deposited into a central ASSEMBLE Plus data repository for archival purposes.

Data infrastructure components that are specific to the genomics observatories data and that were developed during the FP7 MicroB3 project will be revitalized and connected to the entry point to meet the GOs data requirements. This includes links to other publicly available virtual environments, services, standards registries, and data repositories (e.g. Moorea Biocode, the biocode commons, the GSC, workflows developed during MicroB3, ENA-EBI data repository, etc.). Data gathered in JRA1 (Genomics Observatories) will be developed into standardized and user friendly formats and incorporated in a “one stop shop” virtual access entry point to GO data and services developed by MicroB3. In addition to providing access to, the portal will provide information on community data policies and standards, experimental protocols/logs, and the outcomes of benchmarking exercises.
EMBRC - European Marine Biological Resource Centre

EMBRC Catalogue  EMBC repository  How to Archive  How to Publish

EDIT FILE METADATA

<table>
<thead>
<tr>
<th>Type</th>
<th>Datatype</th>
<th>Dataprovder</th>
<th>Archive date</th>
</tr>
</thead>
<tbody>
<tr>
<td>pdf</td>
<td>Documents</td>
<td>Marine Biological Association</td>
<td>2018-02-27</td>
</tr>
</tbody>
</table>
Task NA2.3 Set up virtual open access entry point to data resources (3/3)

Task Leader: VLIZ; participants: HCMR, MBA, MPI (and all other partners).

Deliverables: D NA2.2a, b, c: Virtual Access hits to ASSEMBLE Plus data resources (M24, M36, M48)

Data that can be brought into the public domain will be prepared for data publication. A technical quality control will verify that all required information is included. Each of the datasets proposed for publication will be made citable and will be labeled with a digital object identifier (DOI). All gathered information will be checked, improved and validated during a common data workshop to which all providers will be invited.

The use of the virtual open access platform will be registered and monitored using built-in tools. Data availability and “hits” will be assessed by the Advisory Board and reported as a yearly deliverable to the EC (D-NA2.2a, b, c at M24, 36, 48).
Data archiving & publication service

IMIS
Integrated Marine Information System

do
DataCite

MDA
Marine Data Archive

creative commons
Task NA2.4 Improve data access and standardization of genomic and long term marine biodiversity observation

Task leader: VLIZ with contributions from HCMR, MBA, MPI, all genomic & biodiversity observatory stations

Most marine biological stations collect and manage historical times series of biodiversity data. However, many valuable, historic datasets still remain inaccessible to the larger scientific community. The stations that manage these biodiversity data series and the genomic observatory stations will be invited to dedicated sessions during the common data workshop during which they will receive training in providing access to these types of data.

This includes documenting, annotating, archiving, quality controlling and dealing with the IPR issues through existing licensing approaches (e.g. use of moratorium periods and accreditation by producing data citations and data papers). By M18 the described data series will be made accessible by the stations to the task team for final validation and incorporation in the VA platform. A contact point of the task team will be ready to assist the stations with any data related technical issues.
Marine biological stations Long Term Biodiversity data

- Enquiry by Sidonie Gras for EMBRC: 14 responses from 7 countries

| Station Name                                      | T | S | DO | pH | Fluor | Turbidity | NO3 | PO4 | NO2 | SiO4 | NH4 | Chl a | Phytoplankton | Zooplankton | Presence | Diversity | Molecular | DN15 | DC13 |
|--------------------------------------------------|---|---|----|----|-------|-----------|-----|-----|-----|------|-----|-------|--------------|-------------|----------|-----------|-----------|----------|------|------|
| Institut de la Mer de Villefranche               |   |   |    |    |       |            |     |     |     |      |     |       |              |             |          |           |           |          |      |      |
| Station Biologique de Roscoff                    |   |   |    |    |       |            |     |     |     |      |     |       |              |             |          |           |           |          |      |      |
| Observatoire Océanologique de Banyuls             |   |   |    |    |       |            |     |     |     |      |     |       |              |             |          |           |           |          |      |      |
| Royal Belgian Institute of Natural Science        |   |   |    |    |       |            |     |     |     |      |     |       |              |             |          |           |           |          |      |      |
| VLIZ - Flanders Marine Institute                  |   |   |    |    |       |            |     |     |     |      |     |       |              |             |          |           |           |          |      |      |
| ECIMAT - University of Vigo                       |   |   |    |    |       |            |     |     |     |      |     |       |              |             |          |           |           |          |      |      |
| Interuniversity Institute for Marine Sciences (IUI)| | | | | | | | | | | | | | | | | | | | | | | |
| Husö Biological Station                           |   |   |    |    |       |            |     |     |     |      |     |       |              |             |          |           |           |          |      |      |
| Tvärminne Zoological Station, University of Helsinki | | | | | | | | | | | | | | | | | | | | | | | |
| Archipelago Research Institute, University of Turku |   |   |    |    |       |            |     |     |     |      |     |       |              |             |          |           |           |          |      |      |
| Stazione Zoologica Anton Dohrn                    |   |   |    |    |       |            |     |     |     |      |     |       |              |             |          |           |           |          |      |      |
| British Arctic Survey (BAS)                       |   |   |    |    |       |            |     |     |     |      |     |       |              |             |          |           |           |          |      |      |
| Marine Biological Association (MBA)               |   |   |    |    |       |            |     |     |     |      |     |       |              |             |          |           |           |          |      |      |
| Marine Scotland Science (MSS)                     |   |   |    |    |       |            |     |     |     |      |     |       |              |             |          |           |           |          |      |      |

**KEY**
- 19% Data collected and available for download
- 26% Data collected and available upon request
- 11% Data collected but not available
- 44% Data not collected

Need to adress additional AssemblePlus stations
Genomics Observatory data

- Reinstated the OSD sampling and metadata registration tool
- Needs testing before use for 2019 event
- Discussions to be organised with GFBIO and EBI on:
  - Metadata standards
  - Data infrastructure
  - Data brokering
Task NA2.5 Set up virtual platform for data analysis

Task leader: HCMR with contributions from VLIZ, MPI, and UGOT.

Deliverables: D NA2.3a, b, c: Virtual Access runs to analysis platform (M24, M36, M48)

The standardized data from long-term biodiversity and genomics observatories will be connected to a virtual analysis platform. This platform will be an online portal (based on R web services and Taverna workflow systems) that enables the user to select from available datasets and perform predefined processing and analysis workflows on the selected data. Examples of workflows offered by the platform are existing outlier detection, visualization, statistical analysis and index calculation workflows for biodiversity observation data and bioinformatics pipelines (QIIME based demultiplexing, quality filtering, OTU picking, etc.). Depending on the type of workflow, output will be displayed as maps, reports and/or data tables. An authentication system will allow users to log on to the platform and access the pre-installed workflows.

A workshop (M24) will be organized to allow the data providers to test and validate the analysis platform. During the workshop a webinar will be prepared for teaching and stimulating the further use of the platform.

Use of the virtual analysis platform will be monitored using built-in tools. Number of users and “runs” per user will be assessed by the Advisory Board and reported on a yearly deliverable to the EC (DNA2.3a, b, c at M24, 36, 48).
Thank You!