

ASSEMBLE



ASSOCIATION OF EUROPEAN MARINE BIOLOGICAL LABORATORIES EXPANDED

Acronym: ASSEMBLE Plus

Title: Association of European Marine Biological Laboratories Expanded

Grant Agreement: 730984

Deliverable D4.1

ASSEMBLE Plus First Data Management Plan

03/2018

Lead parties for Deliverable: VLIZ

Due date of deliverable: M6

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.



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	D4.1
Deliverable title	ASSEMBLE Plus Data Management Plan
Submission due date	M6
Actual submission date	06/04/2018
WP number & title	WP4, 1 st Data management plan
WP Lead Beneficiary	VLIZ
Participants (names & institutions)	Katrina Exter (NA2 partner), Flanders Marine Institute (VLIZ); Stefanie Dekeyzer (NA2 partner), Flanders Marine Institute (VLIZ); Georgios Kotoulas (WP Leader JRA1), Hellenic Center of Marine Research (HCMR); Estefania Paredes/Antonio De Villanueva (WP Leader JRA2), University of the Basque Country (UPV/EHN); Hector Escriva (WP Leader JRA3), UPMC; Ian Probert , (WP Leader JRA4), University Pierre and Marie CURIE (UPMC); Piotr Balazy (WP Leader JRA4), Institute of Oceanology Polish Academy of Science (IOPAN); Christos Arvanitidis (NA2 partner, LifeWatch), Hellenic Center of Marine Research (HCMR); Mark Hoebeke (NA2 partner, EMBRC WGEI), Station Biologique de Roscoff (SBR); Erwan Corre (NA2 partner, EMBRC WGEI), Station Biologique de Roscoff (SBR); Coppens Frederik (ELIXIR), Flanders Institute for Biotechnology (VIB); Mathias Obst , (NA2 partner), Gothenburg University (UGOT); Simon Claus (EMODNet Biology), Flanders Marine Institute (VLIZ); Francisco Hernandez (LifeWatch), Flanders Marine Institute (VLIZ); Dan Lear (NA2 partner, EMBRC WGEI), Marine Biological Association (MBA); Klaas Deneudt (WP leader NA2, EMBRC WGEI), Flanders Marine Institute (VLIZ)

Dissemination Type

Report	<input checked="" type="checkbox"/>
Websites, patent filling, etc.	<input type="checkbox"/>
Ethics	<input type="checkbox"/>
Open Research Data Pilot (ORDP)	<input type="checkbox"/>
Demonstrator	<input type="checkbox"/>
Other	<input type="checkbox"/>

Dissemination Level

Public	<input checked="" type="checkbox"/>
Confidential, only for members of the consortium (including the Commission Services)	<input type="checkbox"/>



Document properties

Author(s)	Katrina Exter
Editor(s)	S. Dekeyzer, K. Deneudt
Version	1

Abstract

This is the first version of the Data Management Plan for ASSEMBLE Plus, created for the 6-month point in the project. This document outlines how the research data collected or generated within ASSEMBLE Plus will be handled during and after the end of the project. This DMP builds on that of EMBRC and incorporates 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). This DMP defines the methodologies and standards to be applied to ASSEMBLE Plus data, whether and how the data will be shared and/or made open, and how the data will be curated and preserved. This DMP can evolve and gain substance throughout the project: further versions will be created in months 24 and 48.



1.	Introduction.....	5
1.1	The ASSEMBLE Plus project.....	5
1.2	The DMP meeting.....	6
1.3	The ASSEMBLE Plus data collection.....	6
1.4	The ASSEMBLE Plus data origins	8
1.5	Other users of ASSEMBLE Plus data	8
2.	FAIR Data	10
2.1.	Making data findable.....	10
2.1.1	Metadata	10
2.1.2	File naming	11
2.2.	Making data openly accessible.....	12
2.2.1	Open Access	12
2.2.2	Storage.....	13
2.2.3	Data access.....	13
2.3.	Making data interoperable.....	14
2.3.1	File formats.....	14
2.3.2	Data formatting and vocabulary standards.....	14
2.4.	Making data re-useable.....	16
2.4.1	Quality control.....	16
2.4.2	Long-term re-use	17
2.4.3	Data-creation standards.....	18
3.	Allocation of Resources	19
3.1.	Costs for making data FAIR.....	19
3.2.	Responsibilities.....	19
4.	Data Security	21
5.	Ethical Aspects.....	23
	ANNEX 1 Best-practice guidelines	24
	ANNEX 2 Tables	26



1. Introduction

1.1 The ASSEMBLE Plus project

The focus of ASSEMBLE Plus is on marine science and marine data, and the focus of this data management plan is on making these data Findable, Accessible, Interoperable, and Re-useable (FAIR). By enabling access to the data and knowledge of its members, the lifetime and usefulness of their scientific research will be greatly enhanced. This will be of benefit to marine scientists, and also to the blue economy, policy, and education. ASSEMBLE Plus will provide scientists from academia, industry, and policy with Transnational and Virtual Access to its marine biological research facilities: to their historical observation data, their research data and knowledge, their laboratories, and to advanced training opportunities. The long-term sustainability of the marine stations will be improved by stimulating fundamental and applied research excellence in marine biology and ecology in Europe.

ASSEMBLE Plus brings together 32 marine stations and institutes with modern research infrastructures and track-records of unique service provision, from 14 European and two associated countries, under the leadership of the European Marine Biological Resource Centre (EMBRC), an ESFRI consortium developed from the previous ASSEMBLE (FP7) partnership. Five Joint Research Activities (JRAs) are included in the ASSEMBLE Plus consortium:

- **JRA 1 Genomics Observatories.** The motivation for this JRA is to foster the application of genomics technologies at Long-Term Ecological Research Network (LTER) sites. The project encompasses: populating and verifying databases of taxonomic reference barcodes; harmonising meta-barcoding standard operating procedures (SOPs) across the consortium; and inter-calibration of classical biodiversity data and genomics data. The final objective is the establishment of a distributed Genomics Observatory across the partnership and beyond, of which the data will be available for virtual access (VA).
- **JRA 2 Cryopreservation of Marine Organisms.** This JRA will address a constraint in the exploitation of marine genetic and biological resources, namely the current paucity of capability to preserve these resources *ex-situ* with a guaranteed genetic, phenotypic and functional stability. The JRA will develop robust, reproducible cryopreservation methodologies for various life-stages of a range of marine macro-organisms and currently cryo-recalcitrant microorganisms. This JRA will collect best current practises and create new protocols from laboratory experiments in the cryo-preservation of marine organisms.
- **JRA 3 Functional Genomics.** This JRA will address the need to establish links between genomic information and phenotypes of marine model species, by developing small-scale functional genomic approaches for several marine models for the generation of Genetically Modified Marine Organisms (GMOs). This JRA will be largely dedicated to transferring established techniques for the generation of genetic resources, and where necessary adapting those techniques, to model organisms for which these techniques have not yet been applied.
- **JRA 4 Development and Standardisation of On-site Instrumentation for Experimental Marine Biology and Ecology.** The aims of this JRA are (i) to produce detailed technical specifications for biological resource centre infrastructure and experimental facilities; (ii) to produce best practise guidelines for future cross-consortium implementation of standardised experimental systems and associated infrastructure. This JRA will collect technical design specifications of



experimental systems and associated infrastructure, with the aim of improving the service provision of future instrumentation.

- **JRA 5 Scientific Diving.** The goal of this JRA enable a standardised employment of emerging or breakthrough diving technologies. The aim is to improve diving-based science delivery by improving the use of emerging technologies. The data collected by this JRA will allow the building of a common service and will generate a wider and more diverse user group of this type of data.

Another important part of ASSEMBLE Plus is **Transnational Access (TA)**. The goal of TA is to provide marine scientists with high-quality services, through a single access point, to its member marine stations. These marine stations provide access to a high diversity of marine environments: from the high Arctic and Antarctic to the tropics and the mid-Atlantic ridge. Within mainland Europe, access is provided to the Mediterranean, the Atlantic and the Baltic Sea. There are eight categories of services provided: biological resources, ecosystem access, experimental facilities, technology platforms, training and education, e-infrastructure, data and services; as well as supporting facilities and expert advice. Via TA services, scientists will be able to conduct research at the marine stations (hence accessing both the physical services and the knowledge of the staff and researchers), or they can request remote access (e.g. request that samples are sent to their own labs). The first call was closed in Jan. 2018 and the successful projects will be announced in March. Further calls will be issued twice-yearly for the duration of the ASSEMBLE Plus project. A User Selection panel, made of internal and external experts, will select among the proposed projects: it is envisaged that many dozens of projects will be successful in each call.

1.2 The DMP meeting

This DMP takes much input from that of EMBRC, however information unique to ASSEMBLE Plus, in particular the JRAs and TA, was required. A workshop to gather and discuss these necessary information took place on March 1-2 2018 at VLIZ. Participants included the representatives of the JRAs, and experts from related initiatives and partners: Lifewatch, ELIXIR, European Marine Observation and Data Network Biology (EMODNet Biology), EMBRC Working Group on E-Infrastructure (EMBRC WG EI). Also invited were experts from the Digital Curation Centre (DCC), International Oceanographic Data Exchange network (IODE) and Ocean Biogeographic Information System (OBIS).

1.3 The ASSEMBLE Plus data collection

The data collected by ASSEMBLE Plus will be as diverse as the topics covered by the five JRAs, ranging from straightforward databases to 3D imagery and models. Where appropriate, these data will include raw data, curated data (partially-processed data products), and final reports, publications, and models.

The data that will be collected by the TA projects cannot be known *a-priori*, however since the projects will be carried out at the ASSEMBLE Plus member marine stations, in-house expertise will be provided to the visiting researchers so that they can conform to the norms of the ASSEMBLE Plus data collection.

Two types of data will be created during the ASSEMBLE Plus lifetime:

- Type 1: data generated by a public research project carried out at a member stations and originating from ASSEMBLE Plus user access provision and where the project covers the operating costs.



- Type 2: ASSEMBLE Plus partners’ institutional data that are not necessarily generated in the context of ASSEMBLE Plus, but are explicitly listed by the partners as part of the service offer.

Type 1 data are “foreground” data (i.e. data which is generated by ASSEMBLE Plus activities); Type 2 data are background data (e.g. long-term data that was gathered by the institute or station before the start of the ASSEMBLE Plus project).

The data formats used and data types produced depend on how the data were gathered and the software used to analyse them. The ASSEMBLE Plus consortium is diverse in research topics and so necessarily is the list of data formats and types. However, consolidation of these lists is a priority of the data management activities of ASSEMBLE Plus, particularly within thematic data types. Standardisation at the data level will be performed by applying community-based standards as proposed by international initiatives, such as the International Oceanographic Data Exchange programme (IODE), the Ocean Biogeographic Information system (OBIS), the Genomics Standards Consortium (GSC), and the Open Geospatial Consortium (OGC).

Several data types require different file types. Table 1 gives an overview of possible file types for sharing, re-use and preservation (based on the UK Data Archive). ASSEMBLE Plus recommends avoiding the use of propriety software formats where possible. If data files are in a different format, it is recommended to convert the data to a more common data format. More details on the specific file formats used for the thematic data types produced in ASSEMBLE Plus is given in Sec. 2.3.1.

Table 1 - Overview of possible file types for sharing, re-use, and preservation (UK Data Archive)

Type of data	Acceptable formats for sharing, re-use and preservation	Other acceptable formats for data preservation
Quantitative tabular data with extensive metadata. A dataset with variable labels, code labels, and defined missing values, in addition to the matrix of data	<ul style="list-style-type: none"> • SPSS portable format (.por) • Delimited text and command ('setup') file (SPSS, Stata, SAS, etc.) containing metadata information • Some structured text or mark-up file containing metadata information, e.g. DDI XML file 	<ul style="list-style-type: none"> • Proprietary formats of statistical packages e.g. SPSS (.sav), Stata (.dta) • MS Access (.mdb/.accdb)
Quantitative tabular data with minimal metadata. A matrix of data with or without column headings or variable names, but no other metadata or labelling	<ul style="list-style-type: none"> • Comma-separated values (CSV) file (.csv) • Tab-delimited file (.tab) • Including delimited text of given character set with SQL data definition statements where appropriate 	<ul style="list-style-type: none"> • Delimited text of given character set - only characters not present in the data should be used as delimiters (.txt) • Widely-used formats, e.g. MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb), dBase (.dbf) and OpenDocument Spreadsheet (.ods)
Geospatial data. Vector and raster data	<ul style="list-style-type: none"> • ESRI Shapefile (essential - .shp, .shx, .dbf, optional - .prj, .sbx, .sbn) • geo-referenced TIFF (.tif, .tiff) • CAD data (.dwg) • tabular GIS attribute data 	<ul style="list-style-type: none"> • ESRI Geodatabase format (.mdb) • MapInfo Interchange Format (.mif) for vector data • Keyhole Mark-up Language (KML) (.kml) • Adobe Illustrator (.ai), CAD data (.dxf or .svg) • binary formats of GIS and CAD packages
Qualitative data. Textual	<ul style="list-style-type: none"> • eXtensible Mark-up Language (XML) text according to an appropriate Document Type Definition (DTD) or schema (.xml) • Rich Text Format (.rtf) • plain text data, ASCII (.txt) 	<ul style="list-style-type: none"> • Hypertext Mark-up Language (HTML) (.html) • widely-used proprietary formats, e.g. MS Word (.doc/.docx) • some proprietary/software-specific formats, e.g. NUD*IST, NVivo and ATLAS.ti



Digital image data	<ul style="list-style-type: none"> TIFF version 6 uncompressed (.tif) 	<ul style="list-style-type: none"> JPEG (.jpeg, .jpg) but only if created in this format TIFF (other versions) (.tif, .tiff) Adobe Portable Document Format (PDF/A, PDF) (.pdf) standard RAW image format (.raw) Photoshop files (.psd)
Digital audio data	<ul style="list-style-type: none"> Free Lossless Audio Codec (FLAC) (.flac) 	<ul style="list-style-type: none"> MPEG-1 Audio Layer 3 (.mp3) but only if created in this format Audio Interchange File Format (AIFF) (.aif) Waveform Audio Format (WAV) (.wav)
Digital video data	<ul style="list-style-type: none"> MPEG-4 (.mp4) motion JPEG 2000 (.mj2) 	
Documentation and scripts	<ul style="list-style-type: none"> Rich Text Format (.rtf) PDF/A or PDF (.pdf) HTML (.htm) OpenDocument Text (.odt) 	<ul style="list-style-type: none"> plain text (.txt) some widely-used proprietary formats, e.g. MS Word (.doc/.docx) or MS Excel (.xls/.xlsx) XML marked-up text (.xml) according to an appropriate DTD or schema, e.g. XHTML 1.0

For most of the projects included in ASSEMBLE Plus, the size of the data collections will not be large, of the order of MB. For two of the JRAs the data will be heavier: JRA1 expects to produce datasets of TB per individual project (especially those combining FlowCam assays with genomic screening, and MicroCT scan images); JRA 5 will produce photogrammetry datasets that could also be TB in size.

1.4 The ASSEMBLE Plus data origins

The origin of the data is the marine stations that are included in the ASSEMBLE Plus consortium. For JRA1 the current data are from the previous Ocean Sampling Days, and additional data will be generated by the project itself. For JRA2 and JRA3 most of the data will be created in the laboratory. For JRA4 the data will be taken from existing design specifications. For JRA5 the data will originate from future photogrammetry dives and from data-loggers situated on sub-tidal buoys.

TA projects are research activities offered to, and carried out by, visiting researchers operating under the umbrella of ASSEMBLE Plus. The projects carried out by the TA programme are not yet known: the first call has just closed and the projects will be selected in March; up to 8 calls will be issued over the lifetime of ASSEMBLE Plus. The origin of the TA data will clearly be the scientific project run by the visiting researcher.

There will be some re-use of existing data in ASSEMBLE Plus. JRA1 will use the data from the previous three Ocean Sampling Days. JRA3 will collect genomic samples taken at long-term monitoring sites, including therefore data taken in the past. Reference gene databases (SILVA, BOLD) will be used together with the newly-generated data. For JRA4, much of the data collected will originate from existing technical manuals of elements of the equipment/infrastructure. JRA 2 will use some data from an existing culture collection database.

1.5 Other users of ASSEMBLE Plus data

The ASSEMBLE Plus consortium covers a wide panel of scientific fields: instrumentation, cryobanking, genomics, photogrammetry. The data will clearly be of use to the wider marine scientific community.



Possible applications for the ASSEMBLE Plus data are in ecological, fundamental and applied research; policy support; conservation; blue technology; pharmaceutical research, etc.

Application sectors could include gene and cell engineering (molecular farming, cell factories), bio-refineries, biostatistics, software development, sub-sea industries, nutrition, medicine and health care, aquaculture, environmental remediation, and bioenergy and biomaterials. Other potential users of the data include Marine Protected Areas agencies and Marine Scientific boards.

The ASSEMBLE Plus public web-pages will include communications to the general public, based on the work carried out within the consortium, and this will also have pedagogical uses.



2. FAIR Data

2.1. Making data findable

The primary responsibility for storage and findability of the data lies with the data creator. However, all data – raw and curated data, final data products – created within the JRAs (i.e. in their marine stations and laboratories) will be stored in one central archive for long-term preservation: The Marine Data Archive (MDA), which is hosted by the Marine Data Centre of VLIZ. This does not preclude additional use of European and international repositories. The aim of the MDA is data *preservation*; data *access* (open or not) is a separate issue: more information on this and on the databases used for the ASSEMBLE Plus data collection can be found in Sec. 2.2.2.

All datasets will include metadata: discovery and technical metadata that defines the what, where, when, why, and how data of the data. Where appropriate, the data creators will assign Digital Object Identifiers (DOIs) or a Persistent URL (PURL) to their datasets. For data stored in the MDA, these metadata will be held in IMIS, the Integrated Marine Information System, which is also located at the Data Centre of VLIZ. IMIS is a digital catalogue, with dedicated modules for metadata, publications, datasets, institutes, people, projects, events and maps. As well as all ASSEMBLE Plus metadata, all ASSEMBLE Plus documentation (reports, publications, etc.) will be stored there. IMIS is linked to the MDA: users can search via keywords in IMIS, and from there be directed to the dataset in the MDA. *Therefore, via these linked systems – the MDA and IMIS – data will be findable and accessible.*

The approaches for storing, finding, and accessing the data generated within the TA projects is still under discussion. The expectations on FAIR data will be made clear to all that use and provide the TA services. The available tools and approaches above will be offered to TA users and providers. Final responsibility to comply with EU regulations will be the responsibility of the TA users and providers. The full process will be drawn up once a full TA cycle has been completed (at the time of writing, no TA projects had been initiated).

2.1.1 Metadata

The use of common data and metadata standards and formats are a key aspect for technological and semantic data operability, allowing the data to be discoverable and hence promoting international and interdisciplinary access to, and use of, research data. Standard dictionaries of metadata will be used for all data types produced by the data creators. The use of LIMS and FIMS – which are used to store and organise data obtained in the laboratory or in the field – will allow not only good file management but also for the creation of standard metadata *in situ*. In case the data is already accessible through local online databases, a web link to these local systems will be included in the dataset description. Any unique or uncommon metadata keywords will be mapped to a more common standard during the import to the MDA and IMIS. A technical quality control will ensure that all required information is included. Training in metadata creation and its storage in IMIS for the ASSEMBLE Plus users is planned.

Different communities and disciplines develop and adopt various metadata standards and/or practices for the management of their research data and materials. Some of the more commonly used metadata standards are:

- **ISO:** The International Organization for Standardization (ISO) creates documents that provide requirements, specifications, guidelines or characteristics that can be used consistently to ensure that materials, products, processes and services are fit for purpose. So far, ISO has



published 21,578 International Standards. A commonly-used ISO standard for geographic data is the ISO 19115:2003 Geographic information – Metadata standard, which defines how to describe geographical information and associated services, including contents, spatial-temporal purchases, data quality, access and rights to use. It is maintained by the ISO/TC 211 committee.

- **GCMD:** NASA’s Global Change Master Directory (GCMD) holds more than 34,000 Earth science data sets and service descriptions, which cover subject areas within the Earth and environmental sciences. The project mission is to assist researchers, policy makers, and the public in the discovery of and access to data, related services, and ancillary information (which includes descriptions of instruments and platforms relevant to global change and Earth science research). Within this mission, the directory also offers online authoring tools to providers of data and services, facilitating the capability to make their products available to the Earth science community. In addition, citation information to properly credit dataset contributions is offered, along with direct links to data and services. As an integral part of the project, keyword vocabularies have been developed and are constantly being refined and expanded. These vocabularies are also used in other applications within the broader scientific community. <https://gcmd.nasa.gov>
- **EML:** Ecological Metadata Language (EML) is a metadata standard developed by and for the ecology discipline. It is based on prior work done by the Ecological Society of America and others. EML is implemented as a series of XML document types that can be used in a modular and extensible manner to document ecological data. Each EML module is designed to describe one logical part of the total metadata that should be included with any ecological dataset.
- **OpenAIRE (OAI-PMH v2.0):** The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) provides an application-independent interoperability framework based on metadata harvesting. <http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm>
- **GeOMe (Genomic Observatories Metadatabase):** is a new open access repository for geographic and ecological metadata associated with biosamples and genetic data. GeOMe will use the metadata standards of the Genomic Standards Consortium. It will allow the metadata required for ecological and evolutionary analyses to be assigned to biosamples and genetic data. <http://www.geome-db.org>

2.1.2 File naming

A structured data storage is essential for proper and secure storage of data files and records. For any file-based storage this includes clear and unambiguous file naming, the use of proper versioning, clear and intuitive folder structure. The use of a standard file naming convention per data type, and a clear versioning will be encouraged by ASSEMBLE Plus, and in some cases is ensured by use of community services such as ELIXIR (the European Life-sciences Infrastructure for Biological Information) and OBIS. Moreover, the use of relational database system to store those data where there is repeated need for entry, storage, querying and analysis will be encouraged. Specific systems are available to support the information management of certain lab and field workflows. These LIMS (Laboratory Information Management Systems) and FIMS (Field Information Management Systems) have predefined data models and interfaces that can facilitate a lot of the repeating processes in the research environment.



Some recommendations for good file names:

- Create meaningful names (file names can contain e.g. project acronyms, researchers' initials, file type information, a version number, file status information and date);
- Use file names to classify broad types of files;
- Avoid using spaces and special characters;
- Avoid very long file names;

(From the UK Data Archive)

Within ASSEMBLE Plus a set of recommendations for file naming for the various thematic data types will be drawn up. These recommendations will be based on what is currently used within the JRAs and the community standards.

2.2. Making data openly accessible

The ASSEMBLE Plus data policy is primarily based on the EMBRC data policy. The EMBRC-ERIC Data Policy builds on the general principles described by the EMBRC-ERIC Statutes and the EMBRC Technical and Scientific Description for research data.

The ASSEMBLE Plus Data Policy

1. Covers data acquired, assembled or created through research, survey and monitoring activities by or involving ASSEMBLE Plus projects that are either fully or partially funded. The ASSEMBLE Plus Data Policy also applies to data managed by ASSEMBLE Plus.
2. Will promote e-infrastructure interoperability and standardisation to optimise dealing with large volumes of generated data, for which ASSEMBLE Plus has data-handling protocols, tools and expertise in place.
3. Acknowledges the importance of long-term preservation of research data in trustworthy accredited repositories: this included the data repositories used by the individual marine stations and organised by the fix JRAs leads, and the MDA repository that will store a copy of all data.

2.2.1 Open Access

Following the guidelines of the H2020 regulations defined for the Open Research Data pilot, all data collected or produced within the ASSEMBLE Plus consortium will by default be open; providing a motivation for non-OA will be a requirement. Some data may be shared only under restrictions (e.g. within the consortium only). The expectation is that data becomes open at publication or after two years of the data collection.

- **JRA 1.** All metadata will be open and freely available from the day of creation. Most of the remaining data will be open access at publication. Sensitive data (e.g. endangered species/habitats) may have access restricted to members of the consortium and the EC.
- **JRA 2,3.** Once the protocols have been developed, these (as refereed publications or ASSEMBLE Plus reports) and the laboratory data leading to them (JRA 2) will be provided with open access.



- **JRA 4.** Basic information on equipment/infrastructure will be open access. Data embargo for part of the data due to the potential creation of a consultancy service within the EMBRC infrastructure is being considered.
- **JRA 5.** The final data products (curated images, 3D models, environmental data from sub-tidal buoys) will be provided with open access. Provision of the raw image files with open access is still under discussion because of their large number and file sizes, and their limited re-use potential compared to the curated products.
- **TA.** For TA projects, the researchers will only need to provide a motivation if their data is not to be provided with open access once their research has been published or after two years of data collection. Each request will be considered with due diligence.

2.2.2 Storage

The storage of data created by the research activities carried out in the ASSEMBLE Plus marine stations – JRA and TA activities – will primarily be responsibility of the data creators. However, all ASSEMBLE Plus data will also be stored at the MDA, which ASSEMBLE Plus will use as an archival storage facility, for data preservation; this therefore does not exclude contributions to other recognised data initiatives. To stimulate discovery and potential re-use of the data, ASSEMBLE Plus will also maximally exploit existing data-flow pathways to share data through European e-infrastructures such as EMODnet, LifeWatch and ELIXIR, and global initiatives such as the Global Biodiversity Information Facility (GBIF) and the Global Earth Observation System of Systems (GEOSS), especially its biodiversity component (GEOBON). Table 2 (Annex 2) lists some potential (European and non-European) data repositories for redistribution for the thematic data types generated within ASSEMBLE Plus, for the JRA and TA data. This list of repositories will be narrowed down for the second release of the DMP.

TA users will be encouraged to follow the data storage and access policy of ASSEMBLE Plus, and will be assisted if choosing to do so. However, final responsibility lies with the TA user and provider. The full process will be drawn up once a full TA cycle (proposal—acceptance—experiment—data creation) has been completed.

2.2.3 Data access

As described in the grant agreement the MDA and IMIS systems will be used for storage of ASSEMBLE Plus data and metadata. Various thematic repositories are listed in Table 2 that can be used by any researcher to store or obtain data, and these will be used within ASSEMBLE Plus. In addition to these, general repositories such as Zenodo (<https://zenodo.org/>) and national repositories (e.g. <https://ecds.se/>: JRA 1) may be used. For data stored in the MDA and other data repositories, access is via standard web database interfaces. The documentation describing how to find, get, and access the data are described in the data repository or software documentation, and no additional documentation is required from ASSEMBLE Plus.

For the cases where access to the data needs to be restricted, this can be identified at the time the data is submitted to the data repository – this is true for the MDA and many of the archives listed in Table 2. For users to obtain data from the MDA, registration is necessary and access is then via password. For restricted JRA1 data (because the data concerns endangered species or habitats) there will be a contact point for requests for data access.

For much of the data produced by the ASSEMBLE Plus consortium members JRA 1, 2, 3, and 4, the data are of formats that can be handled by commonly-used commercial or open software (e.g. they are



spreadsheets, images). Well-established routines are available for genomics data (e.g. ELIXIR). Data that are produced by community-standard software systems will be accessible to those using the same, community-standard, software (JRA 1). The 3D models produced by JRA 5 will be made available via online platforms such as Sketchfab that can handle and display these formats, and the model data themselves can also be provided in a neutral file format.

2.3. Making data interoperable

ASSEMBLE Plus realises that common data and metadata standards and formats are a key aspect for technological and semantic data operability. Standardisation makes the data discoverable and this way promotes international and interdisciplinary access to, and use of, research data. To ensure correct and proper use and interpretation of the ASSEMBLE Plus data by the owners and re-users, the use of standardised vocabularies and ontologies is also necessary.

2.3.1 File formats

The format and software in which research data are created depend on how researchers choose to collect and analyse data, often determined by discipline-specific standards and customs. Ensuring long-term usability of the ASSEMBLE Plus data requires consideration of the most appropriate software and file formats. Table 3 (Annex 2) lists the file formats, broken down by thematic data type, that will be used in the ASSEMBLE Plus consortium.

Given the broad spectrum of data types, there is a large offer of domain-specific formats for data sharing and thematic systems that can serve redistribution of those data. An important challenge for ASSEMBLE Plus will be to assess the different standards and workflows that will be proposed by the five JRAs under which umbrella data will be produced. The challenge to decide on commonality, both within ASSEMBLE Plus and with the broader marine research world, and the aim of ASSEMBLE Plus is to reduce the variety of file types used as much as is feasible. The EMBRC Scientific Strategy report refers to the need for close collaboration with ESFRI initiatives, such as ELIXIR and LifeWatch, and this will play also a crucial role in determining how data is managed, shared and processed within ASSEMBLE Plus.

2.3.2 Data formatting and vocabulary standards

Standardisation at the data level will be performed by applying community-based standards as proposed by international initiatives such as the International Oceanographic Data Exchange programme (IODE), the Ocean Biogeographic Information System (OBIS), the Genomics Standards Consortium (GSC) and the Open Geospatial Consortium (OGC). Standards dictionaries and the measurement norms used by the JRAs are of the following themes:

- **JRA 1, 3:** genomic and taxonomical, environmental, physico-chemical, imaging
- **JRA 2:** biological and cryobiological, references to prior-acquired data such as culture collections
- **JRA 4:** technical
- **JRA 5:** environmental, physico-chemical, biological



Standards in file formats, data formatting, controlled vocabulary, metadata dictionary, and measurement that will be used by ASSEMBLE Plus will follow the recommendations of the following initiatives:

- **Darwin Core Archive (DwC-A):** The Darwin Core is designed to facilitate the exchange of information about the geographic occurrence of organisms and the physical existence of biotic specimens in collections. Extensions to the Darwin Core provide a mechanism for sharing additional information, which may be discipline specific or beyond the commonly-agreed upon scope of the Darwin Core itself. The Darwin Core and its extensions are minimally restrictive of information content by design, since doing so would render the standard useless for the implementation of data quality tools. <http://rs.tdwg.org/dwc/>
- **OBIS-ENV-DATA:** The OBIS-ENV-DATA format is a new data standard for combined marine biological and environmental datasets. It is based on Darwin Core Archive standard and recommended by the Ocean Biogeographic Information System part of the International Oceanographic Data Exchange network. <https://bdj.pensoft.net/articles.php?id=10989>
- **M2B3 Reporting Standard:** The M2BR standard was developed during the MB3 project to ensure that the collected data could be correctly directed to, and stored in, their respective domain-specific data archives, which were the ENA for molecular data and PANGAEA for environmental data and morphology-based biodiversity data. <https://standardsingenomics.biomedcentral.com/articles/10.1186/s40793-015-0001-5>.
- **Data formats for read data (genomic data):** Read data can be submitted in several standard (CRAM, BAM, Fastq) and platform specific formats (SFF, PacBio, Oxford Nanopore, Complete Genomics). ENA recommends that read data is either submitted in BAM or CRAM format. http://ena-docs.readthedocs.io/en/latest/format_01.html
- **Data formats for assembled and annotated sequences (genomic data):** The main format for assembled and annotated sequences is the flat file format, which is defined in full detail in the ENA Assembled Sequence User Manual. Assembled and annotated sequences are available in flat file and other formats, namely FASTA and XML, through the ENA Browser (<https://www.ebi.ac.uk/ena/submit/sequence-format>). Furthermore, Multiple Sequence Alignment (MSA) formats and General Feature Formats (GGF2, GGF3) for annotation exist.
- **NetCDF (oceanographic data):** NetCDF (Network Common Data Form) is a set of interfaces for array-oriented data access and a freely distributed collection of data-access libraries for C, Fortran, C++, Java, and other languages. The NetCDF libraries support a machine-independent format for representing scientific data. Together, the interfaces, libraries, and format support the creation, access, and sharing of scientific data. <https://www.unidata.ucar.edu/software/netcdf/docs/faq.html>
- **Ocean Data View (ODV) format (oceanographic data):** The ODV data format allows dense storage and very fast data access. Large data collections with millions of stations can easily be maintained and explored on inexpensive desktop and notebook computers. Data from ARGO, GTSP, CCHDO, World Ocean Database, World Ocean Atlas, World Ocean Circulation Experiment (WOCE), SeaDataNet, and Medar/Medatlas can be directly imported into ODV (<https://odv.awi.de/>). The British Oceanographic Data Centre distributes a SeaDataNet version of the general ODV format to carry additional information required by SeaDataNet. https://www.bodc.ac.uk/resources/delivery_formats/odv_format/



Where unique or uncommon ontologies or vocabularies are found (the main potential here being with the TA projects), a mapping of these to common vocabularies forms part of the process of submitting the data to the MDA and its metadata to IMIS.

2.4. Making data re-useable

2.4.1 Quality control

For the ASSEMBLE Plus research data collection, the quality control of the data can happen at various stages during the quality assurance process. An initial quality control is needed at the local level and early in the collection process. Additional controls will take place at a later stage of the data life-cycle. A final quality control of metadata takes place during its input into IMIS.

The initial quality control of the data, during data collection, is the primary responsibility of the ASSEMBLE Plus data creator/owner, who must ensure that the recorded data reflect the actual facts, responses, observations and events. The quality of the data collection methods used strongly influences data quality, and documenting in detail how data are collected provides evidence of such quality. Errors can also occur during data entry. Data are digitised, transcribed, entered in a database or spreadsheet, or coded. Here, quality is ensured by standardised and consistent procedures for data entry with clear instructions.

Some suggestions for quality control measures during data collection or data entry:

- Calibration of instruments to check the precision, bias and/or scale of measurement;
- Taking multiple measurements, observations or samples;
- Checking the truth of the record with an expert;
- Using standardized methods and protocols for capturing observations, alongside recording forms with clear instructions
- Setting up validation rules or input masks in data entry software;
- Using data entry screens;
- Using controlled vocabularies, code lists and choice lists to minimize manual data entry;
- Detailed labelling of variable and record names to avoid confusion;
- Designing a purpose-built database structure to organize data and data files;
- Accompanying notes and documentation about the data.

(From the UK Data Archive)

An additional quality control, which is highly recommended for the ASSEMBLE Plus operators, can be performed when checking of the data when the data are edited, cleaned, verified, cross-checked and validated. Checking typically involves both automated and manual procedures (based on the UK Data Archive):



Some suggestions for additional quality control at data level:

- Double-checking coding of observations or responses and out-of-range values;
- Checking data completeness;
- Adding variable and value labels where appropriate;
- Verifying random samples of the digital data against the original data
- Double entry of data;
- Statistical analyses such as frequencies, means, ranges or clustering to detect errors and anomalous values;
- Correcting errors made during transcription;
- Peer review.

(From the UK Data Archive)

Ecological data gathered by JRA1 will use the QC protocols and tools that are part of OBIS. ASSEMBLE Plus will organise training in quality assurance processes for their operators, and this for the different types of data.

2.4.2 Long-term re-use

ASSEMBLE Plus supports the concept of FAIR data, and will work towards making ASSEMBLE Plus research data FAIR. The decision about long-term provision will be taken as the data are stored: open access data will be made FAIR for as long as possible, and a CC-BY licence will be the expectation for ASSEMBLE Plus OA data. Any necessary discussion for particular datasets (for example, particularly large datasets, IPR-sensitive datasets) will involve the JRA representatives and the ASSEMBLE Plus data management and the Project Implementation Committee. Most of the research data of ASSEMBLE Plus will be open access as soon as the protocols have been developed or the research been completed and published, other data collections will be made available immediately following cleaning and pre-processing (e.g. Ocean Sampling Day datasets). Embargo or opt-outs from open access are outlined in Sec. 2.2.1. The long-term reuse of most data is provided for in terms of access rights.

There are no plans to end provision of the ASSEMBLE Plus data, and they will therefore be available for re-use for as long as the archives exist. The potential for re-use is not the same for all projects in ASSEMBLE Plus: the laboratory data of JRA 2 and 3 are focussed on the protocols that they are aimed at developing and the potential for re-use is small. For the remaining three JRAs, however, the potential for re-use is large:

- **JRA 5:** methodology purposes, long-term monitoring
- **JRA 1:** scientific, policy, or commercial uses (commercial use could demand negotiation and alignment with the Nagoya Protocol on Access and Benefit-sharing)
- **JRA 4:** laboratory infrastructure design, commercial uses

The data produced under ASSEMBLE Plus will be re-usable for as long as the information they contain are relevant. The aim of certain projects is to develop protocols (JRA 2 and 3), and as such the data are relevant primarily to these end products (which will also be made available by ASSEMBLE Plus). The instrumentation data of JRA 1 will be re-usable until advances in technology make them obsolete.



Nevertheless, the policy of ASSEMBLE Plus is to provide data so that future users can access and use that data as and when they wish.

2.4.3 Data-creation standards

ASSEMBLE Plus acknowledges the fact that there are as many standards and methodologies in the creation of data as there are instruments and data types. Harmonisation of procedures will be undertaken where it makes sense in the ASSEMBLE Plus projects. A “best practice” guideline to help achieve high and compatible standards in experimental methods, culture and husbandry, data collection and analysis is being developed. The guidelines currently in place are presented in Annex 1. This list aggregates a number of standard operating procedures, protocols, and guidelines on research activities such as holding, breeding, and culturing of marine organisms, working with genetic and genomic resources, etc. This toolbox will be expanded on during the ASSEMBLE Plus lifetime, and it will become part of the ASSEMBLE Plus Knowledge Transfer Platform.



3. Allocation of Resources

3.1. Costs for making data FAIR

Resources for long-term preservation of datasets will be ensured by its storage in the repositories outlined in Table 2, and in the MDA. The costs for storing data at the location stations will be born locally. The MDA and IMIS are a service provided by VLIZ and its costs are covered by that organisation and its EMBRC involvement.

The costs for making publications open access can vary from 500 to 5000 euro, depending on the journal. This will be covered by ASSEMBLE Plus. The costs of transforming historical data to machine-readable forms will also be covered by ASSEMBLE Plus.

There may be costs associated with providing processing space for data- and memory-heavy ASSEMBLE Plus projects to be carried out, and with the storage of large raw datasets. These points are still under discussion in the consortium.

3.2. Responsibilities

Implementation of the DMP is necessary at both the central and local level. The coordination of the ASSEMBLE Plus DMP lies with the ASSEMBLE Plus implementation board. The individual operators will ensure the DMP is implemented at the local level, and VLIZ will oversee compliance. VLIZ will also support the operators by organising the necessary training in the framework of work package NA2.

The data collected in the course of TA projects will be subject to the same data management requirements as all ASSEMBLE Plus activities. This requirement will be made clear during the TA application process and a (brief) DMP will be part of the submission material for all future calls. For the now-closed call, it will be requested only from the successful applicants.

Throughout this DMP, the responsibilities for several steps in the data management process have been described. These are summarised here.

- ASSEMBLE Plus will advise and train its members to organise a structured data storage.
- Quality control of the data, during data collection, is the primary responsibility of the ASSEMBLE Plus operators (as data providers). ASSEMBLE Plus will organise training in quality assurance processes for its operators, for the various types of data.
- Each ASSEMBLE Plus data creator will describe their generated datasets in the digital Metadata Catalogue IMIS, using online forms, and this catalogue is open to any user. These metadata will describe the scientific and the technical aspects of the data.
- In general, the primary responsibility for ensuring adequate storage capacity lies with the ASSEMBLE Plus operators. ASSEMBLE Plus will engage in mapping the needs of the operators with the aim of using European and international repositories. In addition to this, all ASSEMBLE Plus data will also be stored at the MDA repository.
- The primary responsibility for back-up and recovery of the data also lies with the ASSEMBLE Plus operators, and for the data stored in the MDA that responsibility lies with VLIZ/MDA.
- ASSEMBLE Plus will undertake all required efforts needed to protect the data, products and services against unauthorised use. The primary responsibility to take necessary measures to ensure data security lies with the ASSEMBLE Plus operators and with VLIZ/MDA.



- ASSEMBLE Plus will undertake all efforts required to provide secure access to data. Where applicable, authentication systems will be used, requesting log-in before providing access to secured data and information.
- Furthermore, ASSEMBLE Plus will take measures to be compliant with the EU regulations regarding the protection of personal data (<http://ec.europa.eu/justice/data-protection/>).
- ASSEMBLE Plus will work towards providing virtual access to the data that is considered part of the service offer.
- ASSEMBLE Plus promotes a culture of openness and sharing of data and will therefore stimulate the exchange of good practices in data access and sharing by liaising with existing European initiatives or relevance for environmental and biological data and bioinformatics.



4. Data Security

ASSEMBLE Plus will undertake all required efforts needed to protect the data, products and services against unauthorised use. The primary responsibility to take necessary measures to ensure data security lies with the its operators, and once stored on a community data repository (such as the MDA) the security provisions come from there.

Physical security, network security and security of computer systems and files all need to be considered to ensure security of data and prevent unauthorised access, changes to data, disclosure or destruction of data.

Some suggestions for **physical data security**:

- Controlling access to rooms and buildings where data, computers or media are held;
- Logging the removal of, and access to, media or hardcopy material in store rooms;
- Transporting sensitive data only under exceptional circumstances, even for repair purposes, e.g. giving a failed hard drive containing sensitive data to a computer manufacturer may cause a breach of security.

(From the UK Data Archive)

Some suggestions for **network security**:

- Not storing confidential data such as those containing personal information on servers or computers connected to an external network, particularly servers that host internet services;
- Firewall protection and security-related upgrades and patches to operating systems to avoid viruses and malicious code;
- Install anti-virus packages and schedule regular scans.

(From the UK Data Archive)



Some suggestions for **security of computer systems and files**:

- Locking computer systems with a password and installing a firewall system;
- Protecting servers by power surge protection systems through line-interactive uninterruptible power supply (UPS) systems;
- Implementing password protection of, and controlled access to, data files, e.g. no access, read only, read and write or administrator-only permission;
- Controlling access to restricted materials with encryption;
- Imposing non-disclosure agreements for managers or users of confidential data;
- Not sending personal or confidential data via email or other file transfer means without first encrypting them;
- Destroying data in a consistent manner when needed;
- Remember that file sharing services such as Google Docs or Dropbox may not be that secure.

(From the UK Data Archive)



5. Ethical Aspects

ASSEMBLE Plus will be compliant with the EMBRC ethics review guidelines, and will follow the guidelines of ENVRI+ (an H2020 programme: Environmental and Earth System Research Infrastructures: <http://www.envriplus.eu/>). Furthermore, ASSEMBLE Plus will take measures to be compliant with the EU regulations regarding the protection of personal data (<http://ec.europa.eu/justice/data-protection/>). An ASSEMBLE Plus ethics checklist, which is also a required submission for the project at this 6-month point, can be consulted for further information.



ANNEX 1 Best-practice guidelines

The ASSEMBLE Plus “virtual tool-box of best practice guidelines” are can be found in <http://www.assemblemarine.org/virtual-tool-box-of-best-practice-guide-lines/>. This toolbox aggregates the following protocols and guidelines:

- On-site access – hosting guest scientists
 - Draft letter of acceptance
 - Best practice guidelines for on-site access
- Remote access – shipping of marine organisms
 - Best practice guidelines for remote access
- Whole, multicellular marine organisms – holding, breeding, culturing, etc.
 - Sampling strategy for *Ciona intestinalis*
 - Fertilisation tests in *Ciona intestinalis*
 - Material and packaging techniques for the shipment of *Ciona intestinalis*
 - Health and maturation of *Ciona intestinalis*
 - Axenic *Ectocarpus* cultures
 - Biolistic delivery to *Ectocarpus* cultures
 - Genetic crosses in *Ectocarpus*
 - Genetic crosses in *Ectocarpus*
 - Genetic crosses in *Ectocarpus*
 - Extraction of total protein from *Ectocarpus*
 - Isolation and regeneration of protoplasts from *Ectocarpus*
 - *Ectocarpus* RNA extraction method
 - Preparation of seawater media for *Ectocarpus* culture
 - How to cultivate *Ectocarpus*
 - Immunostaining of *Ectocarpus* cells
 - Isolation of *Ectocarpus* gametophytes
 - Passage of *Ectocarpus* cultures
 - Spawning of gametes in *Paracentrotus lividus*
 - Dissection of the endostyle in *Ciona intentinalis*
 - Improved maintenance protocols for corals (*Stylophora pistillata*)
 - Culture technique for Hagfish
 - Culture technique for *Fucus vesiculosus* or *F. serratus*
 - Improved maintenance protocols for kelp gametophytes
- Marine protists and cell lines of marine animals – development, transfection, cryopreservation, etc.
 - Primary cultures of functional Atlantic code melanophores
 - Detection of glycosaminoglycans and fibrous collagen in marine fish cell lines
 - Establishing non-axenic monoclonal cultures of *Skeletonema marinoi*
 - Detection of mineral nodules in marine fish mineralogenic cell lines
 - Development of primary cell cultures from calcified tissues of marine fish
 - Nucleofection of marine fish cell lines
 - Development of primary cell cultures from *Ciona intenstinalis* explant
 - Preparation of fish serum for the culture of marine fish cell ines
 - Cryopreservation and revival of marine fish cell lines
 - Detection of alkaline phosphatase (ALP) activity in marine fish cell lines



- Development of primary cell cultures for *Ciona intestinalis* larval stage
- Transfection of marine fish cell lines using polyethylenimine (PEI)
- Semi-quantitative analysis of cell proliferation in *Asterias rubens* cell monolayers
- Phagocytic behavior of *Asterias rubens* blood cells or coelomic epithelia cells in vitro
- Epithelial cell primary cultures isolated from buds of *Botryllus schlosser*
- Preparation of Ascidian hemolymph for the culture of *Ciona* cells
- Preparation of sea urchin gametes and development of sea urchin embryos
- Primary cell culture from pituitary of *Dicentrarchus labrax*
- Primary cell cultures from total *Paracentrotus lividus* coelomocytes
- Cryopreservation of marine microalgae employing a controlled rate cooler
- Cryopreservation of *Ectocarpus*
- Cryopreservation of marine microalgae employing a passive freezer
- Automated extraction of PCR-grade gDNA from *Ciona intestinalis* tissue
- Automated whole-mount in situ hybridization on developmental stages of *Ciona intestinalis* for the identification of reversible mutations with subtle phenotype
- Cryopreservation protocol for *Ciona intestinalis* sperm
- Screening protocol for the identification of spontaneous mutations in *Ciona intestinalis*
- Maintenance of marine amoebae
- Maintenance of marine ciliates
- Isolation of algal endosymbionts from protists
- Maintenance of marine heterotrophic protists
- Use of FlowCam to enumerate and differentiate between marine protists
- Elimination of bacteria from microalgal culture using antibiotics
- Tips for physical isolation of bacteria-free, clonal microalgae from marine environmental samples
- Medium scale culture of micro-algae in polycarbonate carboys
- Medium scale culture of micro-algae in plastic bags
- AFLP analysis as a tool to investigate genetic diversity in microalgae
- Molecular barcoding of protists
- Culturing of *Pseudo-nitzschia* species on agar medium
- Diatom cleaning with nitric/sulfuric acids
- Dinoflagellate isolation from Cnidarian
- Microalgae preparation for scanning electron microscopy; dehydration
- Genetic and genomic resources
 - Extraction of high quality genomic DNA from *Ectocarpus*
 - Preparation of plugs of *Ectocarpus* material for pulse field electrophoresis
 - UV mutagenesis of *Ectocarpus* gametes
 - *Phaeodactylum tricorutum* cryopreservation
 - *Phaeodactylum tricorutum* transformation by means of the PDS-1000/He Microprojectile Accelerator



ANNEX 2 Tables

Table 2 - Potential data repositories for redistribution of the thematic data types within ASSEMBLE Plus

	Thematic data-type category	Research domains	Potential data repositories for redistribution	URLs
Biological	Biodiversity	biogeographic research, species traits, taxonomy	European Ocean Biogeographic Information System (EurOBIS)	www.eurobis.org
			Ocean Biogeographic Information System (OBIS)	www.iobis.org
			Global Biodiversity Information Facility (GBIF)	http://www.gbif.org
			Biology Portal of the European Marine Observation and Data Network (EMODnet Biology)	www.emodnet.eu/biology
			World Register of Marine Species (WoRMS)	www.marinespecies.org
			Morphobank.org	www.morphobank.org
	Genomic	sequencing (DNA, RNA), annotation of features, protein structural information, gene expression profiles, alignment data, chromosomal mapping, phylogenetic trees, Single	European Nucleotide Archive (ENA)	https://www.ebi.ac.uk/ena
			GenBank	https://www.ncbi.nlm.nih.gov/genbank/
			DNA Data Bank of Japan (DDBJ)	https://www.ddbj.nig.ac.jp/
			dbSNP	https://www.ncbi.nlm.nih.gov/snp



	Nucleotide Polymorphisms (SNPs), functional genomics, metabolomics, proteomics, environmental DNA	European Variation Archive (EVA)	https://www.ebi.ac.uk/eva/
		dbVar	https://www.ncbi.nlm.nih.gov/dbvar
		Database of Genomic Variants Archive (DGVA)	https://www.ebi.ac.uk/dgva
		EBI Metagenomics	https://www.ebi.ac.uk/metagenomics/
		NCBI Trace Archive	https://www.ncbi.nlm.nih.gov/Traces/home/
		NCBI Sequence Read Archive (SRA)	https://www.ncbi.nlm.nih.gov/sra
		NCBI Assembly	https://www.ncbi.nlm.nih.gov/assembly
		NCBI Reference Sequence Database (RefSeq)	https://www.ncbi.nlm.nih.gov/refseq/
		Entrez Nucleotide	https://www.ncbi.nlm.nih.gov/nucleotide
		UniGene	https://www.ncbi.nlm.nih.gov/unigene
		HomoloGene	https://www.ncbi.nlm.nih.gov/homologene
		Protein Information Resource (PIR)	https://pir.georgetown.edu/
		Protein Data Bank (PDB)	https://www.wwpdb.org
		UniProt	www.uniprot.org
Entrez Protein	https://www.ncbi.nlm.nih.gov/protein		



			Protein Circular Dichroism Data Bank (PCDDDB)	pcddb.cryst.bbk.ac.uk/home.php
			ArrayExpress	https://www.ebi.ac.uk/arrayexpress/
			Gene Expression Omnibus (GEO)	https://www.ncbi.nlm.nih.gov/geo/
			GenomeRNAi	www.genomernai.org
			dbGAP	https://www.ncbi.nlm.nih.gov/gap
			The European Genome-phenome Archive (EGA)	https://www.ebi.ac.uk/ega/
			Database of Interacting Proteins (DIP)	http://dip.doe-mbi.ucla.edu/dip/Main.cgi
			IntAct	https://www.ebi.ac.uk/intact/
			Japanese Genotype-phenotype Archive (JGA)	https://www.ddbj.nig.ac.jp/jga
			Biological General Repository for Interaction Datasets	https://thebiogrid.org
			NCBI PubChem BioAssay	https://pubchem.ncbi.nlm.nih.gov
			MetaboLights	https://www.ebi.ac.uk/metabolights/
			PeptideAtlas	http://www.peptideatlas.org
			PRIDE	https://www.ebi.ac.uk/pride/archive/
			ProteomeXchange	www.proteomexchange.org



	Imaging	zooscan images, flowcam images, flow cytometry images, VPR videos	EcoTaxa	ecotaxa.obs-vlfr.fr
			FlowRepository	flowrepository.org
	Biogeochemical	biochemical pathways, nutrients	Data Exchange Portal of MPI	https://www.bgc-jena.mpg.de/geodb/projects/Home.php
	Experimental	data resulting from lab experiments		
Oceanographical	Physical	seawater temperature, salinity, ocean currents, waves, geospatial	The Physical Portal of the European Marine Observation and Data Network (EMODnet Physics)	www.emodnet-physics.eu/map/
	Chemical	pollution, heavy metals	The Chemical Portal of the European Marine Observation and Data Network (EMODnet Chemistry)	www.emodnet-chemistry.eu/welcome
Climatological	Meteorological	air temperature, wind speed, ...		
	Modelling	climate models	BioModels Database	https://www.ebi.ac.uk/biomodels-main/
			Kinetic Models of Biological Systems (KiMoSys)	https://kimosys.org
Technical	Instrumental	experimental instruments, laboratory set-ups		



Table 3 - Overview of the thematic data type categories generated within the research domains of ASSEMBLE Plus. For each data-type category, the common data formats and file types are listed.

	Thematic data type category	Research domains	Data formats	Data file types
Biological	Biodiversity	biogeographic research, species traits, taxonomy	Darwin Core Archive (DwC-A), OBIS-ENV-DATA	<ul style="list-style-type: none"> Quantitative tabular data, minimal metadata: .csv; .tab; .xls; .xlsx; .txt; .mdb; .accdb; .dbf; .ods Quantitative tabular data, extensive metadata: .por; SPSS, Stata, Sas, DDI XML, .sav; .dta Qualitative data: .xml; .rtf; .txt; .html; .doc; .docx; proprietary/software-specific formats, e.g. NUD*IST, NVivo and ATLAS.ti
	Genomic	sequencing (DNA, RNA), annotation of features, protein structural information, gene expression profiles, alignment data, chromosomal mapping, phylogenetic trees, Single Nucleotide Polymorphisms (SNPs), functional genomics, Metabolomics, Proteomics, environmental DNA	<ul style="list-style-type: none"> M2B3 Reporting Standard Read data: general (CRAM, BAM, Fastq) and platform specific (SFF, PacBio, Oxford Nanopore, CompleteGenomics) Assembled and annotated sequence data: flat file format (FASTA, XML), Multiple Sequence Alignment (MSA) formats 	<ul style="list-style-type: none"> Quantitative tabular data, minimal metadata: .csv; .tab; .xls; .xlsx; .txt; .mdb; .accdb; .dbf; .ods Quantitative tabular data, extensive metadata: .por; SPSS, Stata, Sas, DDI XML, .sav; .dta Qualitative data: .xml; .rtf; .txt; .html; .doc; .docx; proprietary/software-specific formats, e.g. NUD*IST, NVivo and ATLAS.ti
	Imaging	zooscan images, flowcam images, flow cytometry images, VPR videos, photogrammetry		<ul style="list-style-type: none"> JPEG (.jpeg, .jpg), TIFF (.tif, .tiff), Adobe Portable Document Format (PDF/A, PDF) (.pdf), standard applicable RAW image format (.raw), Photoshop files (.psd), MPEG-4 (.mp4), motion JPEG 2000 (.mj2) Photogrammetry image and video: .bmp; .NEE; .CR2; .ORF; .avi; .wmv; .mp4; .mpg Photogrammetry models: STL; OBJ; FBX; COLLADA; 3DS; VRML/X3D
	Biogeochemical	biochemical pathways, nutrients		<ul style="list-style-type: none"> Quantitative tabular data, minimal metadata: .csv; .tab; .xls; .xlsx; .txt; .mdb; .accdb; .dbf; .ods Quantitative tabular data, extensive metadata: .por; SPSS, Stata, Sas, DDI XML, .sav; .dta Qualitative data: .xml; .rtf; .txt; .html; .doc; .docx; proprietary/software-specific formats, e.g. NUD*IST, NVivo and ATLAS.ti



	Experimental	data resulting from lab experiments		<ul style="list-style-type: none"> Quantitative tabular data, minimal metadata: .csv; .tab; .xls; .xlsx; .txt; .mdb; .accdb; .dbf; .ods Quantitative tabular data, extensive metadata: .por; SPSS, Stata, Sas, DDI XML, .sav; .dta Qualitative data: .xml; .rtf; .txt; .html; .doc; .docx; proprietary/software-specific formats, e.g. NUD*IST, NVivo and ATLAS.ti
Oceanographic	Physical	seawater temperature, salinity, ocean currents, waves, geospatial	NetCDF, Ocean Data View (ODV) format, ASW86	<ul style="list-style-type: none"> Quantitative tabular data, minimal metadata: .csv; .tab; .xls; .xlsx; .txt; .mdb; .accdb; .dbf; .ods Quantitative tabular data, extensive metadata: .por; SPSS, Stata, Sas, DDI XML, .sav; .dta Qualitative data: .xml; .rtf; .txt; .html; .doc; .docx; proprietary/software-specific formats, e.g. NUD*IST, NVivo and ATLAS.ti
	Chemical	pollution, heavy metals	NetCDF, Ocean Data View (ODV) format	<ul style="list-style-type: none"> Quantitative tabular data, minimal metadata: .csv; .tab; .xls; .xlsx; .txt; .mdb; .accdb; .dbf; .ods Quantitative tabular data, extensive metadata: .por; SPSS, Stata, Sas, DDI XML, .sav; .dta Qualitative data: .xml; .rtf; .txt; .html; .doc; .docx; proprietary/software-specific formats, e.g. NUD*IST, NVivo and ATLAS.ti
Climatological	Meteorological	air temperature, wind speed,...		<ul style="list-style-type: none"> Quantitative tabular data, minimal metadata: .csv; .tab; .xls; .xlsx; .txt; .mdb; .accdb; .dbf; .ods Quantitative tabular data, extensive metadata: .por; SPSS, Stata, Sas, DDI XML, .sav; .dta Qualitative data: .xml; .rtf; .txt; .html; .doc; .docx; proprietary/software-specific formats, e.g. NUD*IST, NVivo and ATLAS.ti
	Modelling data	climate models		<ul style="list-style-type: none"> Geospatial data (vector and raster data): .shp; .shx; .dbf; .prj; .sbx; .sbn; .tif; .tfw; .dwg; tabular GIS attribute data; .mdb; .mif; .kml; .ai; .dxf; .svg; binary formats of GIS and CAD packages



Technical	Instrumentation	experimental instruments, laboratory set-ups	images, technical drawings, documentation, spreadsheets	<ul style="list-style-type: none"> • Quantitative and qualitative data: .doc(x); .pdf;.xml; .csv • Images, technical drawings: JPEG; TIFF; .pdf;
------------------	------------------------	---	--	--



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.