ASSEMBLE

ASSOCIATION OF EUROPEAN MARINE BIOLOGICAL LABORATORIES EXPANDED

Acronym: ASSEMBLE Plus

Title: Association of European Marine Biological Laboratories Expanded

Grant Agreement: 730984

Deliverable [D4.3]

[ASSEMBLE Plus Data Management Plan]

[10][2022]

Lead parties for Deliverable: [VLIZ]
Due date of deliverable: M 24 [30/09/2022]
Actual submission date: M 24 [31/10/2012]

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

GENERAL DATA

Acronym: **ASSEMBLE Plus**

Contract N°: **730984**

Start Date: **1st October 2017**

Duration: **60 months**

Deliverable number	D4.3
Deliverable title	ASSEMBLE Plus Data Management Plan
Submission due date	30/09/20122
Actual submission date	31/10/2022
WP number & title	WP4, final Data management plan
WP Lead Beneficiary	VLIZ
Participants (names & institutions)	**Katrina Exter** (WP leader NA2), Flanders Marine Institute (VLIZ); **Georgios Kotoulas** (WP Leader JRA1), Hellenic Center of Marine Research (HCMR); **Estefania Paredes** (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 Bałazy** (WP Leader JRA4), Institute of Oceanology Polish Academy of Science (IOPAN); **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); **Francisco Hernandez** (LifeWatch), Flanders Marine Institute (VLIZ); **Dan Lear** (NA2 partner, EMBRC WGEI), Marine Biological Association (MBA

Dissemination Type

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

Dissemination Level

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

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

Document properties

Author(s)	Katrina Exter
Editor(s)	…
Version	1

Abstract

This is the final version of the Data Management Plan for ASSEMBLE Plus, created for the final point in the project. This document outlines how the research data collected or generated within ASSEMBLE Plus is expected to be handled during and after the end of the project by the JRA and TA programmes. 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 final version of the DMP adds on to the previous versions by containing a summary of the data management activities that were performed by the JRAs and by the TA users.

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

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

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

1. Introduction

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 35 marine stations from 24 partners, 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

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

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 its **Transnational Access (TA)** programme. The goal of TA is to provide marine scientists with high-quality services at its member marine stations, through a single access point. 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, e-services, expert advice, and supporting facilities. Via TA services, scientists can 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). Between Jan. 2018 and Sept 2019 there were four calls for applications, and seven calls are expected by the end of the project. A total of 169 applications have been successful in the first three calls.

1.2 The DMP kickoff 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 WGEI). 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 spreadsheets 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. In addition, assistance from VLIZ's IMIS (catalogue) team is standard for those using IMIS to publish their data (IMIS being one of the catalogues recommended in this DMP). Based on our experience with the data collected by the TA projects that have run so far, most datasets are provided as spreadsheets and images, with sequences provided via public archives such as ENA and NBCI.

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

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

- 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 generated in the context of ASSEMBLE Plus, but are anyway explicitly listed by the partners in the IMIS datasets and publications catalogues.

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	• 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	• 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	• Comma-separated values (CSV) file (.csv) • Tab-delimited file (.tab) • Including delimited text of given character set with SQL data definition statements where appropriate	• 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	• ESRI Shapefile (essential - .shp, .shx, .dbf, optional - .prj, .sbx, .sbn) • geo-referenced TIFF (.tif, .tfw) • CAD data (.dwg) • tabular GIS attribute data	• 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

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

Qualitative data. Textual	• 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)	• 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	• TIFF version 6 uncompressed (.tif)	• 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	• Free Lossless Audio Codec (FLAC) (.flac)	• 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	• MPEG-4 (.mp4) • motion JPEG 2000 (.mj2)	
Documentation and scripts	• Rich Text Format (.rtf) PDF/A or PDF (.pdf) HTML (.htm) OpenDocument Text (.odt)	• 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. XHMTL 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 data are from the previous Ocean Sampling Days (2014), and additional data will be generated by the project itself (2018/19/..). 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 origin of the TA data will clearly be the scientific project run by the visiting researcher, and the range of topics covered will be broad. The datasets are mostly limited in scope because they are gathered during relative short (< one month) experiments with a single focus. Nonetheless, these data will be treated in a similar manner to the JRA data by ASSEMBLE Plus.

There will be some re-use of existing data in ASSEMBLE Plus. JRA1 will use the data from the previous 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.

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

1.5 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.

1.6 The TA DMP

As part of the application process from TA Call 2 onwards, some basic data management questions are asked. The main aim of these questions is to alert the applicants to the requirement that data produced during the course of TA projects should be archived and catalogued, should be made Open Access within at least two years of data collection, and that any eventual refereed publications should be Open Access publications. After a successful TA visit is completed, the user is asked to create a data management plan using a template provided on the TA webpages. This template is based on Horizon2020 DMPs, addressing the Findable, Accessible, Interoperable, and Re-useable points, but is a light version with suggested answers (many TA users have little experience with data management). A guide to FAIR data management for TA users is also provided on the ASSEMBLE Plus website. Uptake of the DMP was slow in the beginning, however from Call 3 onwards the mandatory post-access documentation were combined into a single package, and the uptake since then has been more complete. TA users are requested to archive their data in the MDA (the Marine Data Archive) and catalogue them in IMIS (the Integrated Marine Information System); as these are both VLIZ data systems, assistance in these steps is provided by VLIZ.

1.7 Changes between DMP D4.2 and D4.3

In this final version of the ASSEMBLE Plus DMP we have added an analysis of the actually performed data management of the JRAs, and their data FAIRification activities and Open Access provisions. See Sec. 6.

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

2. FAIR Data

An explanation of FAIR data management, and our expectations for Open Access data and publications, can be found on the FAIR data management pages of the ASSEMBLE Plus site. These explanations and instructions are focussed on the creators of Type 1 data: the JRAs and the TA users, and of Type 2 data: the marine stations, with a particular focus on their long-term and omics datasets.

2.1. Making data findable

Primary responsibility for storage and findability of the data lies with the data creator. Our aim is that 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.

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. 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 approach for storing, finding, and accessing the data generated within the TA projects is the same as that for the JRA data: the datasets should be archived in the MDA, they should be catalogued in IMIS, and they should be made open access, at least within two years of data collection.

Final responsibility to comply with these regulations will be the responsibility of the TA users and providers, however, rather than with ASSEMBLE Plus.

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:

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

- **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.

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

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 for the JRAs. These recommendations will be based on what is currently used within the JRAs and the community standards. For TA users, general recommendations are given in the instructions they are provided with.

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 theJRAs 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. The following opt-outs are allowed:

1. Data will be subject to intellectual property rights
2. Data deal with sensitive information (e.g. rare species or environments, in which case a partial data release may be possible)
3. Data are embargoed until scientific publication (with a maximum wait time of two years)
4. Releasing data early will jeopardise the project's goals (with a requirement that this is explained)

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

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 also 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, however it is the responsibility of the TA researcher to conform with these requirements, ASSEMBLE Plus cannot act as an enforcer.

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, the aim is that 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 once JRA data are being archived (this is not yet the case). An internal listing of the expected and delivered datasets from the JRA is maintained as part of the JRA data management process.

TA users are encouraged to follow the data storage and access policy of ASSEMBLE Plus (archive in the MDA; catalogue in IMIS; make open access after at least two years; and make eventual publications open access), and are assisted if choosing to do so. However, final responsibility lies with the TA user and provider. The full process is documented in the ASSEMBLE Plus website.

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.

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

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 additional, focussed instructions for working with the MDA and IMIS are on the ASSEMBLE Plus webpages.

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 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.

The workshop on FAIR data management, to which all the marine stations with data in the ASSEMBLE Plus collection in IMIS were invited (Sec. 1.7), addressed issues of interoperability. It was agreed that this would be a focus for making FAIR the long-term and omics datasets from our marine stations and which were to be included in the WP4 task *NA2.5: Set up a virtual platform for data analysis*.

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.

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

2.3.2 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

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

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, GTSPP, 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, 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.

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

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

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

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

3. 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 for the JRA publications, but not for TA publications. 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 is made clear during the TA application process and a DMP is part of the submission material for all calls (excepting Call 1).

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.

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

- 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.

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

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)

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

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)

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

5. 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.

6. The state of the FAIR data management at the end of the project

Here we give a summary of the data management activities that were carried out by the JRAs and by the TA users, and a summary of the success of the FAIR and Open Access focus of ASSEMBLE Plus.

Statistics

Up until mid-October, 2022, 30 TNA metadata records have been added to the IMIS datasets catalogue.

- All except three records use CC BY or “unrestricted after moratorium period” for their licence.
- Nine records actually provide a download link to the data, with these data provided via the MDA or as an attachment (usually spreadsheet) to the record. For the rest, the contact details are provided so anyone wishing to do so can contact the data creator with their requests.
- About a dozen TNA scientists used the MDA to store their data but they did not publish them in IMIS. It is possible that the MDA was used here in its capacity as a place for referees to access data linked to the publications they are reviewing, and subsequently the TNA scientists forgot to make a metadata record for those data.

Up until mid-October, 2022, all JRA1 data have been published and a subset of the JRA2 data also.

In addition, at least 60 of the publications arise from TNA projects and the majority of those (85%) are open access. At least 41 publications are from the JRAs (JRAs 1,2 and 3; the additional numbers for JRAs 4 and 5 uncertain due to a lack of provision of the necessary information) and the majority of those (75%) are also open access.

JRA1: Genomics Observatories

The most successful data management was carried out for the two projects of JRA1: the Ocean Sampling Day programme (OSD) and the ARMS-MBON programme.

Ocean Sampling Day

Ocean Sampling Day was a programme for which the sampling activities occurred once a year, over a very wide and diverse geographical range of sampling sites. All the physical (samples) and digital (logsheets) data collected by the participants of OSD were handled by HCMR, with subsequent data

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

management performed by VLIZ. The data management activities – including a description of what digital data were produced, how they were handled, quality control steps, how and what semantic annotations were added, and what data formats were used (for human and for machine access) – are outlined in the Data Management Plan of OSD, which can be found in the OSD GitHub repository.

The data produced by OSD are published (made *Findable* and *Accessible*) as metadata records in IMIS (whereby its human-focussed and webservice can be used to search, view, and download), with the data archived in the MDA, ENA (for the sequences) and in the OSD GitHub pages.

- IMIS records: OSD2018 record, OSD2019 record, and for completeness we have added a copy of the OSD2014 record which is published in PANGAEA and was managed by the MicroB3 project.
- MDA: the data and documentation archived in the MDA for OSD2018 and 2019 can be accessed from those IMIS records.
- OSD GitHub: a wider range of formats for the same OSD data (sampling and environmental data, sequences, protocols, etc), are provided via these GitHub pages. This includes data provided in machine-accessible formats with semantic annotations taken from controlled vocabularies.
- ENA: the sequences from OSD are all archived in ENA, with Project accession numbers 2018: PRJEB40757 (16S), PRJEB55999 (18S), PRJEB40760 (shotgun metagenomes); 2019: PRJEB40762 (16S), PRJEB56005 (18S), PRJEB40764 (shotgun metagenomes).

Semantic annotation with controlled vocabularies (BODC, Darwin Core, Schema.org, etc.) were added to the logsheet data obtained from the sampling scientists, and these data and metadata are provided in standard file formats (CSV, turtle, and packaged as Ro-Crates), making the data *Interoperable*. Quality checks were carried out on the logsheet data obtained from the sampling scientists (on the geographical coordinates, on the environmental measurements, and including a conversion of non-standard units to the required units for these environmental measurements). The genomics data are provided in FASTQ format in ENA and provided with an ENA-standard set of metadata there also (making these data also *Findable*, *Accessible*, and *Interoperable*).

All data are open access (CC BY), making them *Reusable*. We note here that additional work on the provenance aspects of our *Reusable* data is being addressed in the management of the OSD data, but as this depends on deliveries from another Horizon project, it will not be completed until after the ASSEMBLE Plus project has ended.

ARMS-MBON

ARMS-MBON are hard-bottom sampling devices, which are placed in a fixed position for months at a time. Over 20 partners participate in ARMS-MBON, and data collection started in 2018. The sampling and sample processing were carried out by each observatory, and the samples shipped to HCMR for sequence processing. A combination of a data management platform, PlutoF, and google sheets was used for the data (images, ASVs) and metadata (sequence IDs, event metadata) capture at all stages from sampling to sequencing. Bringing these (meta)data together, adding semantics, performing basic quality control, and turning the (meta)data into human- and machine-interoperable formats was done by VLIZ on the ARMS-MBON GitHub repository, providing in one place all the (meta)data, data links (to the photographic images, to the sequences in ENA, to the ASVs), semantics, and descriptions, and the

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

data are packaged and described there in Ro-Crates. For more information, see the Data Management Plan of ARMS-MBON.

The (meta)data from all fully complete ARMS-MBON events, up to Sept, 2022, are published (made *Findable* and *Accessible*) as metadata records in IMIS (whereby its human-focussed and webservice can be used to search, view, and download), with some data archived in the MDA and the rest in the ARMS-MBON GitHub repository.

- IMIS dataset record in the ASSEMBLE Plus collection: at present all sampling events for which sequence data are available and open access in ENA are published here; to be added by then end of 2022 are download links for the (semantically annotated) image data
- MDA: the data and documentation that are archived in the MDA can be accessed from those IMIS records
- ARMS-MBON GitHub repository: for a wider range of formats for the ARMS data (in particular machine-accessible formats), for documentation, and for the scripts for downloading and performing quality control on the metadata, see these GitHub pages.
- ENA: the sequences from ARMS-MBON are all archived in ENA. The data for each station can be found under a unique project accession number: these are provided in the ARMS datasets (and can be found in ENA using the search term "Autonomous Reef Monitoring Structures (ARMS)")

In the ARMS GitHub respository, the data downloaded from PlutoF and the googlesheets will be semantically annotated with controlled vocabularies (BODC, Darwin Core, Schema.org, etc.; work in progress) and provided in standard file formats (CSV, turtle, and packaged as Ro-Crates), making the data *Interoperable*. The genomics data are provided in FASTQ format in ENA and provided with a standard set of metadata there also (making these data also *Findable*, *Accessible*, and *Interoperable*).

The ARMS-MBON project decided to allow for an embargo period of 6 months from the date of adding the sequences in ENA, with an extension to a year for the Covid-19 affected years. After that, all data (sequences and images) are open access (CC BY), making them *Reusable*. (We note here that additional work on the provenance aspects of Reusable data is being addressed in the management of the OSD data, but as this depends on deliveries from another Horizon project, this will not be completed until after the ASSEMBLE Plus project has ended. For more information, see the DMP of ARMS-MBON).

JRA2: Cryobanking

The main outputs of JRA2 was a set of protocols for the cryobanking of marine organisms, and this Toolbox can be found on the A+ website, as a deliverable.

A set of publications related to these protocols have been archived in the ASSEMBLE Plus publications collection, and four datasets have been added to the IMIS datasets catalogue, made open access (CC BY) and provided as a direct download. Much work on standardising the spreadsheets to make them machine-interoperable was carried out by JRA2 with the help of VLIZ.

Unfortunately, not all of the protocols have such linked data records and datasets. Despite numerous requests and the provision of template spreadsheets in which such data could be described (or even just asking for a raw data dump), the remaining JRA scientists did not provide their data for publishing. From discussions with these partners, it is clear that this is because it is not common practise to archive

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

such data: instead, scientific publications and the protocols themselves are considered to be sufficient; the observed/raw data are not considered to be something that should be published separately to the purely "scientific" outputs. If the scientific journals do not require a *full and proper* publication of data, it is also not considered to be important. This is a huge hurdle that still need to be overcome for those of us doing data management.

JRA3,4,5

The output from the other JRAs can be found Public Deliverables section of the ASSEMBLE Plus website

- JRA3 Functional Genomics: Protocols for GMO production in novel emerging metazoan, macroalgal and microalgal model organisms available via ASSEMBLE Plus web portal. This deliverable is one set of the protocols developed by this JRA. (By "ASSEMBLE Plus web portal", they mean this document as provided on the ASSEMBLE Plus website).
- JRA3 Functional Genomics: Protocols for genetic transformation of novel emerging metazoan, macroalgal and microalgal model organisms available via ASSEMBLE Plus web portal
- JRA5 Scientific Diving: Standard operating procedure guidelines for photogrammetry
- A set of publications related to these JRAs have been archived in the ASSEMBLE Plus publications collection.

No separate dataset publications (either archived data or data records) have been made for JRA 3 and 5. This is not the same as saying that no data are published – JRA3 publications, for example, includes figures of the used raw images and include tables of measurements. However, the experimental data leading to the scientific outputs (measurements, images) were not archived separately (in the MDA or anywhere else) or published (in IMIS) because it was not realised by the JRAs that this was necessary: as with some of the JRA2 scientists, the realisation that data need to be published separately to the scientific results and in data archives rather than a journal, was not made. As the journals did not require making all these data accessible via data publications, it was not considered important. A notable exception here are the DNA data produced, for which the journals do require reference to archived data and therefore those data *are* published.

We note that as the output for JRA4 is, itself, a catalogue, there was no need to additionally add any JRA4 outputs to the ASSEMBLE Plus datasets catalogue.

For future reference, we recommend that if FAIR and Open Access data arising from scientific investigations is considered to be an important part of a project, that *(1) the FAIRification and archiving process is required to be carried out as soon as data exist, and as ongoing work during the project (with associated milestones and deliverables), (2) a WP is not considered to be complete until the data management is also, (3) that more resources are given to these scientific data producers specifically in order to do this data management (which is often a new area of work for them), and (4) tailored data management plans and templates are created at the very beginning of the project for each WP/topic/partner to help them do this data management, and that this is done as a collaborative work rather than top-down. The scientists need to, and need to want to and be able to, take ownership of publishing their data.*

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

TA programme

The TA programme usually created very specific datasets which were very often part of a larger study, and as such any datasets archived from the TA programme will be incomplete – it will only include what was carried out at the ASSEMBLE Plus station, rather than all data upon which any eventual scientific results were based. For this reason, making TA data FAIR (in particular, making them Findable by archiving them, and interoperable by formatting them), was not the primary focus of ASSMBLE Plus (that being reserved for the more extended JRA programmes).

Given that the TA programmes were mostly carried out by pre-PhD or young post-doc scientists, who (apparently) receive no training in FAIR data or data management, there was quite some success in the FAIRification of TA datasets and in making TA publications Open. Most of the publications from the TNA programme were published in Open Access journals, and while comparatively few datasets were published in the IMIS datasets catalogue (~30), we can report that a good number those who did publish their data made exceptional efforts to understand how to create FAIR metadata and FAIR datasets. *Our recommendation is that data management and how to make data (rather than science) FAIR, should be much better taught at university and lab level. All data-creating organisations should have a data manager who can advise on, or even implement, FAIR data management practises and templates that the data-creating scientists can use.*

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

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
    - Crypopreservation and revival of marine fish cell lines
    - Detection of alkaline phosphatase (ALP) activity in marine fish cell lines

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

- Development of primary cell cultures for *Ciona intenstinalis* 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 reversive 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 tricornutum* cryopreservation
  - *Phaeodactylum tricornutum* transformation by means of the PDS-1000/He Microprojectile Accelerator

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

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

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

		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

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

			Protein Circular Dichroism Data Bank (PCDDB)	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

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

	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		

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

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	• 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	• 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	• 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		• 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		• 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

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

	Experimental	data resulting from lab experiments		• 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	• 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	• 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,...		• 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		• 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

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

Technical	**Instrumentation**	experimental instruments, laboratory set-ups	images, technical drawings, documentation, spreadsheets	• 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 727610. This output reflects the views only of the author(s), and the European Union cannot be held responsible for any use which may be made of the information contained therein.