ASSEMBLE
ASSOCIATION OF EUROPEAN MARINE BIOLOGICAL LABORATORIES EXPANDED

ASSEMBLE
ASSOCIATION OF EUROPEAN MARINE BIOLOGICAL LABORATORIES EXPANDED

Acronym: ASSEMBLE Plus

Title: Association of European Marine Biological Laboratories Expanded

Grant Agreement: 730984

Deliverable [D4.6]

[3rd virtual access hits to ASSEMBLE Plus data resources]

[10][2022]

Lead parties for Deliverable: [VLIZ]
Due date of deliverable: M 60 [30/09/2022]
Actual submission date: M 60 [31/10/2022]

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.

GENERAL DATA

Acronym: **ASSEMBLE Plus**

Contract N°: **730984**

Start Date: **1st October 2017**

Duration: **60 months**

Deliverable number	D4.6
Deliverable title	3rd virtual access hits to ASSEMBLE Plus data resources
Submission due date	30/09/2022
Actual submission date	31/10/2022
WP number & title	WP4, Improving virtual access to marine biological stations data, information, and knowledge
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); **Dan Lear** (NA2 partner, EMBRC WGEI), Marine Biological Association (MBA); **Ivaylo Kostadinov** (NA2 partner), Max-Planck Institute for Marine Microbiology (MPIMM)

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)	Katrina Exter
Version	1

Abstract

This is the final reporting on the virtual open access entry point to the data resources of ASSEMBLE Plus (Task NA2.3 of WP4), created for the 60-month point in the project (bearing in mind that the project was extended by a year). This document contains updated facts and figures, but otherwise is the same as the first reporting.

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.

Introduction

This deliverable is to report on the *virtual open access entry point to the ASSEMBLE Plus data resources*: what the data resources being accessed are, the platform via which they are accessed, the future plans, and a reporting on the hits to the site. The bulk of the text on those subjects is a copy of the previous reports, with updates to the numbers.

Setting up this access point is the work of WP4 Task NA2.3, and the data resources being accessed are largely the Type 1 and 2 data discussed in the Data Management Plan (WP4 Task NA2.1; D4.3), which includes the data being gathered in WP4 Task NA2.4 (Long-term biodiversity and genomics observations).

The data resources being accessed

The ASSEMBLE Plus data resources that are being accessed via the open access entry point are:

- Data produced by the users of the Transnational Access programme during their TNA visits
- Data and data products produced by the Joint Research Activities
- Data records and datasets gathered from our ASSEMBLE Plus marine stations and partners in Task NA2.4: Long-term marine biodiversity and genomics observations

Note that the first two are data that are expected to participate in the Open Data Pilot, i.e. to be published and made open access; while data from the latter are not required to be open access.

The Open Data pilot

Because ASSEMBLE Plus participated in the Open Data pilot, it was clearly stated at the beginning of the project that the JRAs will create metadata records for their scientific outputs in the IMIS datasets and publications catalogues. The dataset records should describe the datasets and also allow access to those data via a direct download or via a request sent by email. These datasets could be archived in the MDA or in any other accessible online archive/portal, in both cases the web link to the archived dataset would be provided via the IMIS dataset record. This is described in the ASSEMBLE Plus DMP. For publications, the DOI of the publication should be provided and the information to add that publication to the IMIS catalogue can be obtained therefrom. Where those articles are not published in open access journals, the authors were expected to make them available via the ASSEMBLE Plus Open Repository (typically a pre-publication version of the article is placed there, and access to that version is allowed only after the user has accepted the conditions of not passing that article on to others).

The same conditions are placed on the TNA users, and this is made clear as part of their user access contract.

The Transnational Access data resources

The "TNA data resources" are the datasets that are created by the users of the TNA programme during their TNA visit to the ASSEMBLE Plus marine stations. The management of these data is explained in the Data Management Plan (Task NA2.1; D4.3): the users are requested to archive their TNA data in the Marine Data Archive (MDA) or in another appropriate online archive, and to create metadata records in the Integrated Marine Information System (IMIS), where they can be included in 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.

"ASSEMBLE Plus collection". The datasets linked in the records are required to be open access at least two years after data collection.

The MDA and IMIS are two VLIZ data systems, and as part of the remit of the VLIZ Data Centre (VMDC), assistance is given to the TNA users in the archiving and cataloguing process. Guidelines for TNA users are also provided on the ASSEMBLE Plus FAIR data management webpages. The FAIR expectations are explained: for these TNA data the emphasis is on the **F**indable (creating a suitable metadata record in IMIS), and **A**ccessible and **R**e-useable (open access at least after two years from data collection, and obtainable via a direct download link). Advice about creating **I**nteroperable, i.e. standardised, datasets is given on the ASSEMBLE Plus webpages, however a full curation of the interoperability of these TNA datasets out of scope of ASSEMBLE Plus.

The individual TNA datasets are limited in scope as they are research projects that are usually smaller parts of a larger whole, and are run for at most a month. The scope of the *topics* covered by the TNA part of the data collection is, however, wide.

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.

The rate of submissions has increased in the final two years of the project, and it is probable that more (meta)data will come in from the TNA programme throughout 2023. *Nonetheless, it is clear that the uptake of archiving and cataloguing among TNA users is very low, despite the accessible documentation and frequent reminders by the Access Officer.* (Call 1 TNA users should be excused here because the archiving and cataloguing requirements were not made very clear in the application process for this call.)

The JRA data resources

This data arising from the Join Research Activities are varied in size and scope.

- **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). Two programmes have been running under JRA1: Ocean Sampling Day (OSD) and the Automated Reef Monitoring Systems project (ARMS-MBON). These programmes collect event, sampling, and environmental data together with sequence data and image data, from dozens of sampling sites each year.
- **JRA 2 Cryopreservation of Marine Organisms**. This JRA addresses a constraint in the exploitation of marine genetic and biological resources, namely the current paucity 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.

capability to preserve these resources *ex-situ* with a guaranteed genetic, phenotypic and functional stability. The JRA has developed robust, reproducible cryopreservation methodologies for various life-stages of a range of marine macro-organisms and currently cryo-recalcitrant microorganisms. This JRA has collected best current practises and created new protocols from laboratory experiments in the cryo-preservation of marine organisms. The data from JRA2 are mostly in the form of spreadsheets containing the experimental measurements made during the protocol investigations.

- **JRA 3 Functional Genomics**. This JRA addresses 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 is 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. The data from JRA3 include sequences, images, and spreadsheets with experimental measurements.
- **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 has collected technical design specifications of experimental systems and associated infrastructure, with the aim of improving the service provision of future instrumentation.
- **JRA 5 Scientific Diving**. The goal of this JRA is to 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. The data consist of spreadsheets of experimental measurements and photogrammetry datasets.

The FAIR and open access management of the JRA data is described in the ASSEMBLE Plus DMP (Task NA2.1; D4.3), and links there to the JRA1 OSD and ARMS DMPs can be found. As stated above, the publications are expected to be made open access, and the datasets resulting from the JRA work are expected to be published as metadata records in IMIS with the data themselves being archived in the MDA or any other suitable public archive. Making data **F**indable is the responsibility of the JRAs who must initiate the creation of their metadata records, and VLIZ who will assist in the process and curate the results where the MDA and IMIS are used; **A**ccessible is the responsibility of VLIZ (for the MDA and IMIS); and **I**nteroperable and **R**e-useable are the responsibility of the JRAs who have to ensure they add the necessary metadata and conform to the (meta)data standards and formats as described in the DMP. VLIZ has offered assistance on achieving interoperability, especially for JRAs1 and 2.

Some of the JRAs have indeed created datasets records in IMIS, from where their experimental data can be accessed.

- **JRA 1**. All OSD and ARMS-MBON metadata are open and freely available, and the data are also open access from the day that the sequences are published in ENA for OSD, and after 6 months to a year (during COVID) for ARMS-MBON. It was agreed that sensitive data (e.g. endangered species/habitats) may have access restricted to members of the consortium and the EC, but in

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.

fact no data fell into this definition. As of Sept 2022, metadata records for OSD 2018 and 2019 have been added to the ASSEMBLE Plus datasets collection in IMIS, and a copy of the earlier OSD2014 data has also been added for completeness (the original dataset record is published in PANGAEA); and the 2018-20 data from the ARMS project have also been published as a metadata record. For OSD and ARMS data both, these datasets contain all event, environmental, and omics (meta)data. These JRA1 data have had particular attention paid to their interoperability and they can be considered to be very FAIR. In addition, these data are also accessible from the respective Github pages for OSD (or also here) and ARMS, which additionally provides machine-accessible versions of the same data, and Ro-crate data packaging and a more detailed provision of provenance information. The intention is to continue with the data management after the project, publishing these data on EurOBIS/OBIS/GBif (the raw data and the species determinations therefrom), and providing the data in LOD (linked open data) formats suitable for ingesting into various workflows (some are already accessing these data – see D4.9 and D32.4), others will be in the future).

- **JRA 2**. The protocols have been published as deliverables (and hence are open access) and in various other publishing sites (journals, conference proceedings, protocols.io, etc.). Four open access and immediately downloadable dataset records have been published in IMIS in the ASSEMBLE Plus collection (here, here, here, and here), and these are linked to the related protocols and the scientific publications that are based on those data. We note that JRA2 also took significant efforts to make their spreadsheet data interoperable following templates prepared by VLIZ, with the CSV data being both human- and machine-friendly.
- **JRA 3**: The protocols have been published as deliverables (and hence are open access) and in various other publishing sites (journals, conference proceedings, etc.). No dataset records for the associated experimental data have been published, although datasets are accessible via the various journal publications.
- **JRA 4**. The data gathered by this JRA are published via an instrument catalogue. Basic information on equipment/infrastructure are open access, but it still to be decided whether the more technical data will be open access or limited to EMBRC partners. As the output from this JRA is itself a catalogue, no other dataset publications in IMIS are required.
- **JRA 5**. The final data products are curated images, 3D models, environmental data from sub-tidal buoys. 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. The outputs of JRA5 have been published as scientific articles, but no dataset records in IMIS have been created.

It is noted that no dataset records have been published for JRA3 and 5, although the relevant data *are* published via their scientific articles published in journals. This is because of a misunderstanding that it was expected to publish the data *separately* to publishing the science, and lack of knowledge in how to publish such data among the scientists. Training on FAIR data management was given as part of NA2 (see below), but it is clear there is a much greater need of training than ASSEMBLE Plus could provide.

The partner data resources

At the time of writing, the bulk of the ASSEMBLE Plus data collection consists of data records from the ASSEMBLE Plus marine stations/partners and which were already catalogued in IMIS: these were added to the ASSEMBLE Plus collection at the start of the project. These records are for data that were

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.

collected as part of international and national projects, monitoring observations, data for specific projects, and so on. The bulk of the datasets were taken from the 1980s onwards, with the oldest being from 1570.

There are currently 552 data records in the ASSEMBLE Plus collection: of these, 207 re defined as long-term data series (defined as having more than two years of data collecting activity), and 185 are defined as long-term ecological data series (aka LTEDS). The figure below shows the range of topics that are covered by these datasets, as measured by the keywords included in the data records. For the records concerning biological datasets, the most common topics are: fish, plankton, benthos, ecology and biodiversity, invertebrates, macroalgae. For the records concerning non-biological datasets the most common topics are: water composition, fisheries, physical records (water currents, etc), pollution, dredging, coastal studies.

Around 500 of the records describe partner resources, i.e. these data were not created by the JRA or TNA activities but rather the partners themselves, as part of their own work or projects they are engaged in. As these data are owned by the marine stations themselves, not by ASSEMBLE Plus, there is no requirement that they are open access or FAIR. Nonetheless, as records in IMIS, the IMIS team is keen to ensure an appropriate level of FAIRness, and improving these records was part of the work of NA2 (Task 2.3). After a year into ASSEMBLE Plus, a survey of the FAIRness of these metadata records was carried out, with spreadsheets created for each ASSEMBLE Plus partner in which the metadata from the most important fields were listed along with instructions as to whether these were FAIR or not. The important fields considered were: licence, contact details, keywords, title, description, download links. For some fields the metadata provided by the partners are good but there are some fields that are often poorly provided: for example, about half the data records (and 2/3 of the LTEDS) state that they are open access (e.g. CC BY or “unrestricted”) although a great deal of those (e.g. ~70% of LTEDS) do not actually have a direct download link in the IMIS record. Gathering this information was performed by a “FAIR checker” application that was developed at VLIZ. These spreadsheets were sent to each partner in ASSEMBLE Plus prior to and again following a FAIR data mananagement workshop that was held in June 2019 at VLIZ. There were 20 representatives from 12 marine stations and the workshop, and a hands-on session there was devoted to helping individual partners with these spreadsheets and the work to do. All partners did attempt to improve their records, but the actual number of improved records was rather limited. For example, about 10% of the titles and contact details were corrected, and 25% of the descriptions, but very few improvements to the keywords, licences, or other fields were made. A full 25% of the records still have no or insufficient licence

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.

information. Reasons given for the slow progress included a lack of local resources to do the work and being unable to find the necessary information for records that could be several years old.

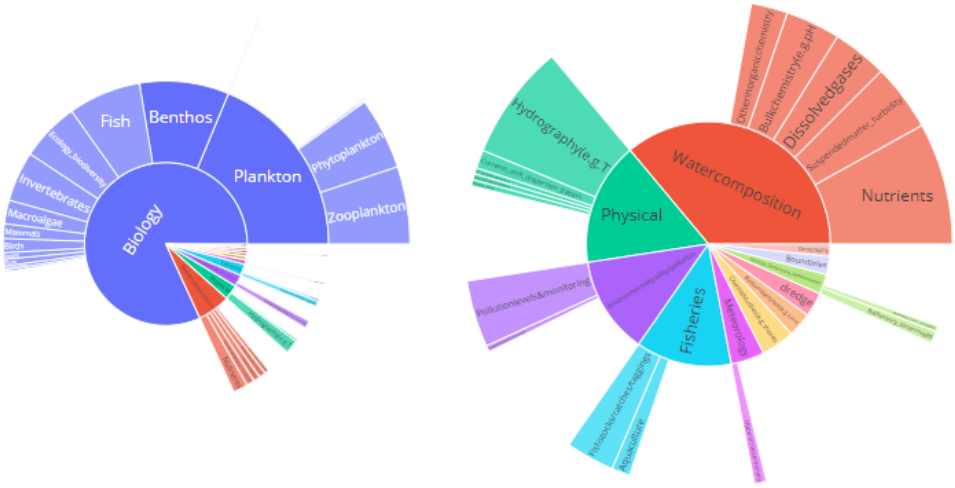

Figure 1 *Graphic displaying the topics covered by the ASSEMBLE Plus data collection. Left are the areas covered by all datasets, right is for the non-biological datasets. This interactive figure is provided as part of the virtual open access point (see Sec. 3): clicking on any part of the pie will zoom in on its associated sub-topics (it is otherwise difficult to read the text). The active figure can be found here.*

The JRA and TNA publications

Publications that are linked to ASSEMBLE Plus partners and JRA/TNA activities are organised in the ASSEMBLE Plus collection in the IMIS publications catalogue. Depositors are required to send the citation, DOI, and sometimes PDF of the publication to add them to this collection. There are currently 454 publications in the collection, of which 234 are from Assemble Marine (grant agreement nr. 227799), which was a precursor project to ASSEMBLE Plus, and also operated under EMBRC. By looking for an acknowledgement to ASSEMBLE Plus in publications, we determined that at least 60 of the publications arise from TNA projects. An additional 41 publications are from the JRAs (JRAs 1,2 and 3) with the numbers for JRAs 4 and 5 uncertain due to a lack of provision of the necessary information.

A graphic with an overview of the topics included in this collection at present is shown in Fig. 2: marine genomics, environmental impact, biodiversity, climate change, oceanography are the most common topics.

It is a requirement of the TNA and JRA programmes that all refereed publications are open access. JRA publications can call on ASSEMBLE Plus resources to pay for this, but TNA users cannot. It is made clear to TNA users that a condition of accepting ASSEMBLE Plus funding is that any refereed publication that is based on their TNA data must be open access. In order to broaden the range of journals that TNA and JRA researchers can publish in, the ASSEMBLE Plus Open Repository was created: PDFs of the pre-print can be deposited by the authors publishing in so-called “green” access journals, and anyone accessing the publication via our collection can download the PDF to read, but not to distribute. Therefore, the aim is that all ASSEMBLE Plus-related publications accessed via our publications collection can be downloaded directly from the collection. Of the 60 known TNA publications, only 8

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.

are not open access or available via the ASSEMBLE Plus Open Repository, and we can consider this to be a success.

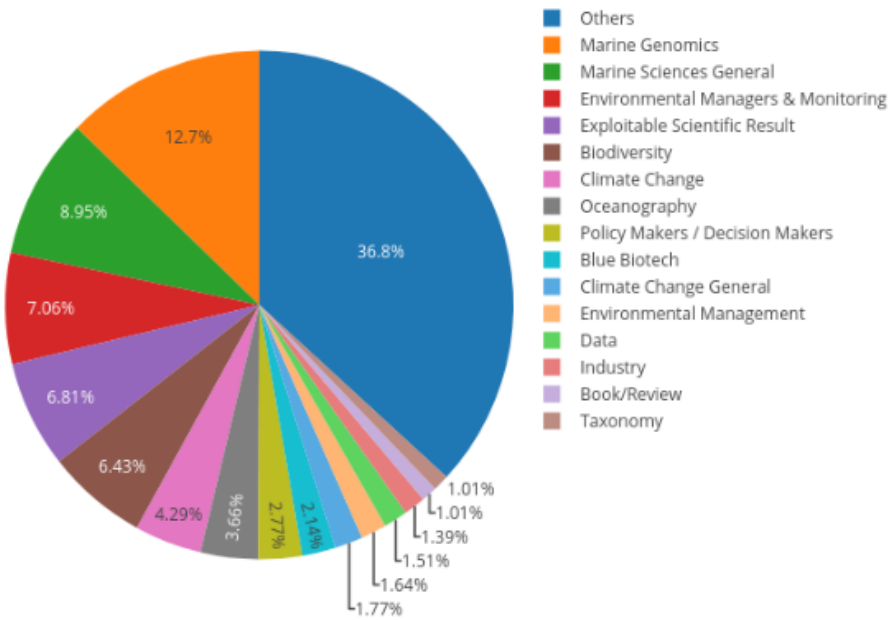

Figure 2 *Topics covered by the ASSEMBLE Plus publications collection*

The virtual open access entry points

The "virtual access entry points" for the ASSEMBLE Plus collections are made accessible on a page on the ASSEMBLE Plus site. The data resources that are linked to on this page are:

- The ASSEMBLE Plus datasets collection (Task NA2.3, 2.4)
- The virtual research environments (Task NA2.5)
- The ASSEMBLE Plus publications collection (Task NA2.2)
- Guidelines for FAIR data management in ASSEMBLE Plus for the TNA users, the JRAs, and the marine stations
- Internal reports/deliverables

Of interest to this report are the access point to the *datasets collection* and the *publications collection*.

The datasets access point

The datasets access point consists of a landing page where the ASSEMBLE Plus data collection is summarised, and a link to the page where the data collection can be browsed and read. A screenshot

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

of the landing page is shown in Fig. 3. An overview of the collection is given, and a link at the top of the page leads to the catalogue browse page, which is shown in Fig. 3.

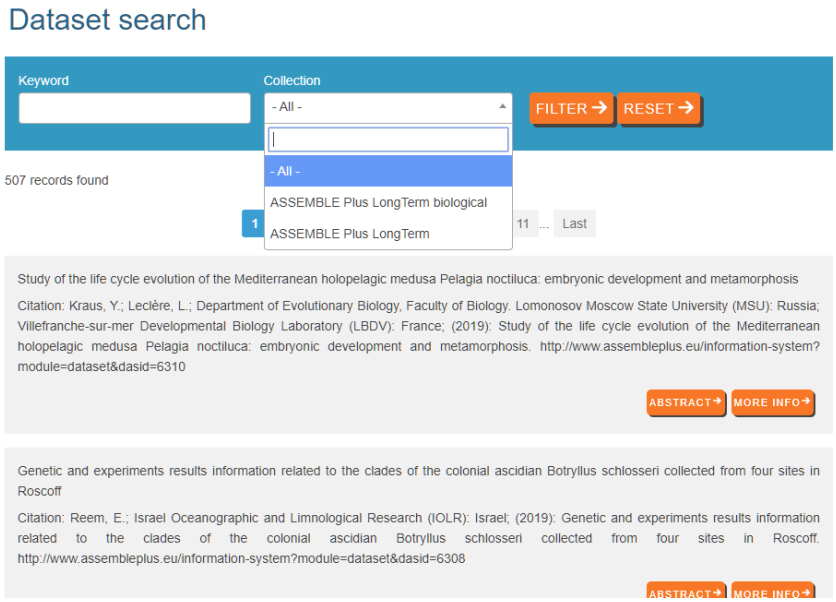

***Figure 3** Datasets collection browser (a filter on KO Type has since been added)*

Browsing the ASSEMBLE Plus data collection uses an IMIS API (Application Programing Interface): one can filter on all ASSEMBLE Plus records and its two child collections (Long-term, Long-term ecological), on a set of Knowledge Output types (being the five Joint Research Activities and the Transnational Access records), and free-text keywords that can be entered. A list of results is returned, with the title displayed. Clicking allows one to read the abstract of the record, or to open the complete record. This metadata record includes the following information:

- Title, data creator contact details, citation, access rights (including licence)
- Abstract and longer description
- Keywords that describe the scope of the collection, these being entered by the record creator via a drop-down ASFA listing or as free text
- The geographic, temporal, and taxonomic coverage
- Parameters of the data
- Information about the contributing agency, and any other links the data creator provided
- Completion status and information about any related datasets

For a more advanced search of the datasets catalogued in IMIS, the IMIS advanced search page can be used (https://www.vliz.be/en/imis?module=dataset&extfrm=1&spcol=990).

The publications access point

The publications access point consists of a landing page where the ASSEMBLE Plus publications collection is summarised, and a link to the page where the collection can be browsed, and publications can be downloaded. An interactive graphic with an overview of the topics included in this collection is given on the landing page, as was shown in Fig. 2. Clicking on a link on the landing page leads to the catalogue browse page, which is shown in Fig. 4.

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.

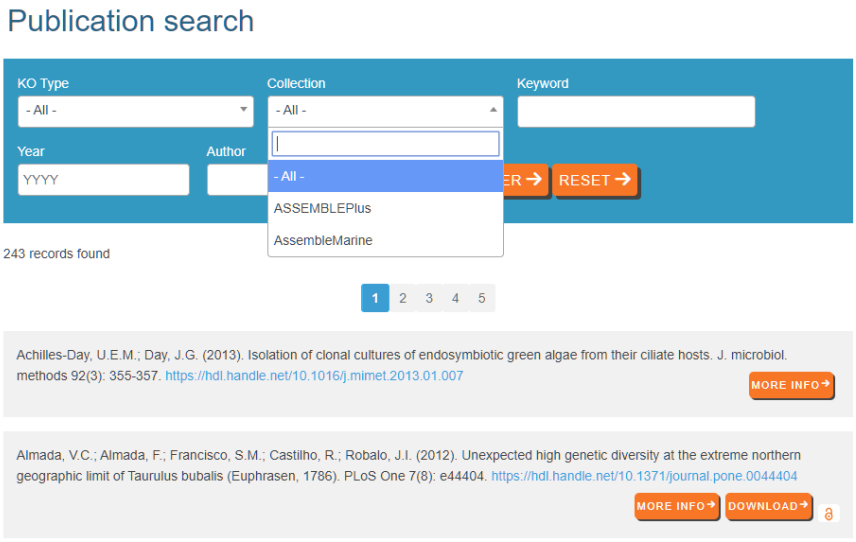

***Figure 4** Publications collection browse page*

Browsing the ASSEMBLE Plus publications collection uses an IMIS API: one can filter on the collection (All, ASSEMBLE Plus, Assemble Marine), KO type (publication, book, case study, modelling/software, prototype, exploitable result, services, JRA[1,2,3,4,5], and TNA) and free-text keyword. A list of results is returned, with the title displayed. Clicking allows to read the IMIS record or (for open access publications) directly download the PDF. The IMIS record includes:

- Title, author, DOI
- Access constraints, link
- Author-entered keyword
- Author list

For publications accessible only via the ASSEMBLE Plus Open Respository, the provided draft (usually a pre-final version of the paper) can be downloaded after clicking a button accepting to "not distribute that draft".

Hits to the virtual open access entry points

The datasets and publications access points have been page available since March 2019 (M18). The number of unique visits in the years 2019, 2020, 2021, and 2022 (to date) to the datasets search page has been 58, 158, 76, 69, and to the publications search page have been 27, 94, 88, 68. Almost 65% of the visits are from Europe, with just over 20% and just under 10% from North America and Asia.

The number of hits to the ASSEMBLE Plus collection since the beginning of the project are as follows:

- Dataset records: from 2018 until 2022, the number of hits are 6, 442, 596, 523, and 1030. The most popular dataset (190 hits) is the ARMS 2018 dataset (JRA1), with the Global tide Variables (143; a publication of the University of Gothenburg, one of the ASSEMBLE Plus partner institutes) being second. The three OSD records – having been created at the end of 2021 – have 81 hits. There are 23 hits to the four JRA2 datasets, but these were added in April and October 2021 and again late in 2022.
- Publication records: from 2018 until 2022, the number of hits are 23, 127, 251, 216, and 651. The deliverable from JRA2 on the Cryomar Protocol Toolbox has been uniquely downloaded 96 times since it was deposited on the ASSEMBLE Plus site.

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