ASSEMBLE

ASSOCIATION OF EUROPEAN MARINE BIOLOGICAL LABORATORIES EXPANDED

Acronym: ASSEMBLE Plus

Title: Association of European Marine Biological Laboratories Expanded

Grant Agreement: 730984

Deliverable D4.5

Virtual access hits to ASSEMBLE Plus data resources

January 2022

Lead parties for Deliverable: VLIZ
Due date of deliverable: M 45
Actual submission date: M 52

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	D4.5
Deliverable title	Second virtual access hits to ASSEMBLE Plus data resources
Submission due date	30/06/2021
Actual submission date	12/01/2022
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); **Christos Arvanitidis** (NA2 partner, LifeWatch), Hellenic Center of Marine Research (HCMR); **Dan Lear** (NA2 partner, EMBRC WGEI), Marine Biological Association (MBA); **Klaas Deneudt** (NA2 partner, EMBRC WGEI), Flanders Marine Institute (VLIZ); **Ivaylo Kostadinov** (NA2 partner), Max-Planck Institute for Marine Microbiology (MPIMM)

Dissemination Type

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

Dissemination Level

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

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 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 second reporting on the virtual open access entry point to the data resources of ASSEMBLE Plus (Task NA2.3 of WP4), created for the 36-month point in the project. This document contains updated facts and figures, but otherwise is the same as the first reporting, D4.4.

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.

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

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.

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; D4.2), which includes the data being gathered in WP4 Task NA2.4 (Long-term biodiversity and genomics observations).

2. 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 TA 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

2.1 The Transnational Access data resources

The "TA data resources" are the datasets that are created by the users of the TA programme during their TA visit to the ASSEMBLE Plus marine station(s). The management of these data is explained in the Data Management Plan (Task NA2.1; D4.2): the users are requested to archive their data in the Marine Data Archive (MDA) and catalogue them in the Integrated Marine Information System (IMIS), where they are included in the "ASSEMBLE Plus collection". The datasets linked to 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 TA users in the archiving and cataloguing process. Guidelines are also provided on the ASSEMBLE Plus FAIR data management webpages. The FAIR expectations are explained: for these TA 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 datasets is beyond the resources available.

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

Up until Jan 6, 2022, about 22 TNA data records have been added to IMIS. *Uptake of the archiving and cataloguing is very low*, and responses to emails are rare – possibly most TA users "forget" this commitment. (Call 1 TA users should be excused here because the archiving and cataloguing requirements were not made very clear in the application process for this call.) Enforcement of this is difficult.

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.

2.2 The JRA data resources

This data arising from the Join Research Activities will be 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). A large part of the data resources arising from JRA1 will come from Ocean Sampling Day (OSD), currently including OSD2014, 18, and 19. Data from ASSEMBLE Plus contribution to the Automated Reef Monitoring Systems project (ARMS) also forms part of the JRA1 data resources.
- **JRA 2 Cryopreservation of Marine Organisms**. This JRA will address a constraint in the exploitation of marine genetic and biological resources, namely the current paucity of capability to preserve these resources *ex-situ* with a guaranteed genetic, phenotypic and functional stability. The JRA will develop robust, reproducible cryopreservation methodologies for various life-stages of a range of marine macro-organisms and currently cryo-recalcitrant microorganisms. This JRA will collect best current practises and create new protocols from laboratory experiments in the cryo-preservation of marine organisms.
- **JRA 3 Functional Genomics**. This JRA will address the need to establish links between genomic information and phenotypes of marine model species, by developing small-scale functional genomic approaches for several marine models for the generation of Genetically Modified Marine Organisms (GMOs). This JRA will be largely dedicated to transferring established techniques for the generation of genetic resources, and where necessary adapting those techniques, to model organisms for which these techniques have not yet been applied.
- **JRA 4 Development and Standardisation of On-site Instrumentation for Experimental Marine Biology and Ecology**. The aims of this JRA are (i) to produce detailed technical specifications for biological resource centre infrastructure and experimental facilities; (ii) to produce best practise guidelines for future cross-consortium implementation of standardised experimental systems and associated infrastructure. This JRA will collect technical design specifications of experimental systems and associated infrastructure, with the aim of improving the service provision of future instrumentation.
- **JRA 5 Scientific Diving**. The goal of this JRA enable a standardised employment of emerging or breakthrough diving technologies. The aim is to improve diving-based science delivery by improving the use of emerging technologies. The data collected by this JRA will allow the building of a common service and will generate a wider and more diverse user group of this type of data.

Management of the JRA data is described in the DMP (Task NA2.1; D4.2): they are to be archived in the Marine Data Archive (MDA) and catalogued in the Integrated Marine Information System (IMIS), and to be included in the ASSEMBLE Plus collection in IMIS. 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; **A**ccessible is the responsibility of VLIZ, as the owners of the MDA and IMIS; and **I**nteroperable and **R**e-useable are the responsibility of the JRAs who have to ensure they add

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.

the necessary metadata and conform to the (meta)data standards and formats as described in the DMP.

Most of these data will be open access:

- **JRA 1**. All OSD and ARMS metadata will be open and freely available, and the data will also be open access from the day that the sequences are published in ENA for OSD, and after 6 months to a year for ARMS. Sensitive data (e.g. endangered species/habitats) may have access restricted to members of the consortium and the EC. As of Jan 06 2022, the OSD data from 2018 and 2019 have been added to the ASSEMBLE Plus datasets collection, and the 2018 data from the ARMS project. In 2022 it is planned to add the OSD 2014 data (which is anyway already published in PANGEA), and the full set of ARMS data (2018 to 2021).
- **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. As of Jan 06, 2022, there are 4 datasets from JRA2 which are recorded in IMIS.
- **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

2.3 The long-term and genomics 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 in bulk to the ASSEMBLE Plus collection. These records are for data that were collected as part of international and national projects, monitoring observations, data for specific projects (e.g. theses), and so on. The bulk of the datasets were taken from the 1980s onwards, with the oldest being from 1570.

There are currently 515 data records in the ASSEMBLE Plus collection from our marine stations: of these, 173 are defined as long-term data series (defined as having more than 2 years of data collecting activity), of which 154 are specifically 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. The scientific scope of these datasets is clearly wide.

The data are owned by the marine stations themselves, not by ASSEMBLE Plus, and so there is no requirement that they are open access. About half the data records (and 2/3 of the LTEDS) state that they are open access (e.g. CC BY or “unrestricted”), however a great deal (~70% of LTEDS) do not actually have a direct download link in the IMIS record. Curation of these data records is therefore an important part of Task NA2.3. A “FAIR checker” application has been developed by the VMDC, and this, together with an improved management template for the IMIS catalogue, will be used to go over

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.

the records in the collection to identify improvements necessary to the metadata, and these recommendations will be send to the data owners to enact (if they so chose).

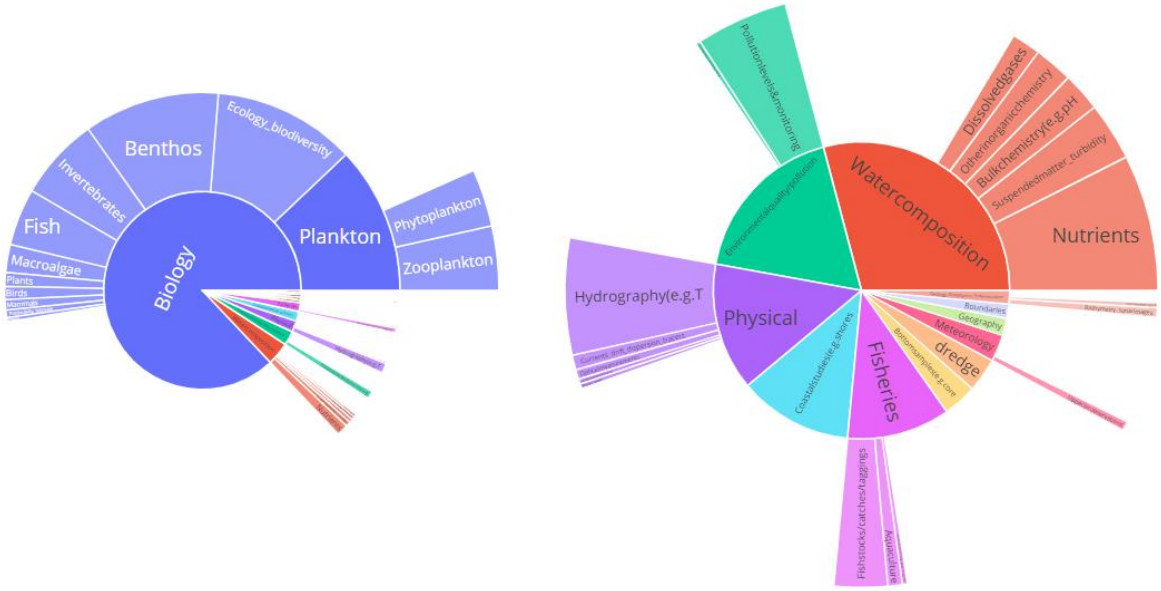

Figure 1 *Graphic displaying the topics covered by the ASSEMBLE Plus data collection. Left are the areas covered by all datasets, right is for the non-biological datasets. This interactive figure is provided as part of the virtual open access point (see Sec. 3): clicking on any part of the pie will zoom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

A FAIR Data Management workshop took place at VLIZ in June 2019. All stations with records in the LTEDS part of the ASSEMBLE Plus collection were invited, as were partners from JRA1, to discuss curation of the OSD and ARMS data. Eventually, 20 representatives from 12 marine stations attended. Resulting from the workshop, the processes to jointly curate these data records were agreed: VLIZ will guide the process, station by station, while the inputs will come from the stations themselves. Emphasis will first be on the **F**indability (the completeness of the metadata records in IMIS), then on **A**ccessibility and **R**e-usability (providing the data download link as part of the record and ideally making most open access). **I**nteroperability will only be addressed for the data that will be included in WP4 Task NA2.5 (*Set up virtual platform for data analysis*). For this final step, assistance from EurOBIS and EMODnet (based on VLIZ) will be enlisted.

2.4 The publications

Publications that are linked to ASSEMBLE Plus are included in the "data resources" that we provide access to via our website. These publications are collected in the IMIS publications catalogue: depositors are required to sent the citation, DOI, and sometimes PDF of the publication to add them to this collection. There are currently 388 ASSEMBLE Plus 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. About 40 and 20 of the publications arise from TNA and JRA work, respectively.

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.

An interactive 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 TA and JRA programmes that all refereed publications are open access. JRA publications can call on ASSEMBLE Plus resources to pay for this (where absolutely necessary), TA users cannot. It is made clear to TA users that a condition of accepting ASSEMBLE Plus funding is that any refereed publication that is based on their TA data must be open access. In order to broaden the range of journals that TA 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. In reality, uptake of adding publications to our collection and publishing as open access by TA users is slow, and enforcement of these point is difficult.

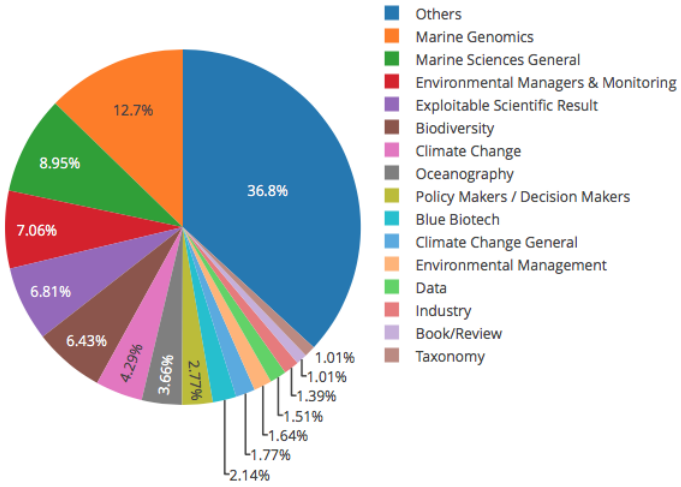

Figure 2 *Topics covered by the ASSEMBLE Plus publications collection*

3. The virtual open access entry points

The "virtual access entry points" for the ASSEMBLE Plus collections are gathered and introduced 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 TA users, the JRAs, and the marine stations
- Internal reports

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

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.

3.1 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 of the landing page is shown in Fig. 3. An overview of the collection is given, including the interactive pie-chart that is described in Fig. 1. Clicking on the orange bar at the top of the page leads to the catalogue browse page, which is shown in Fig. 4.

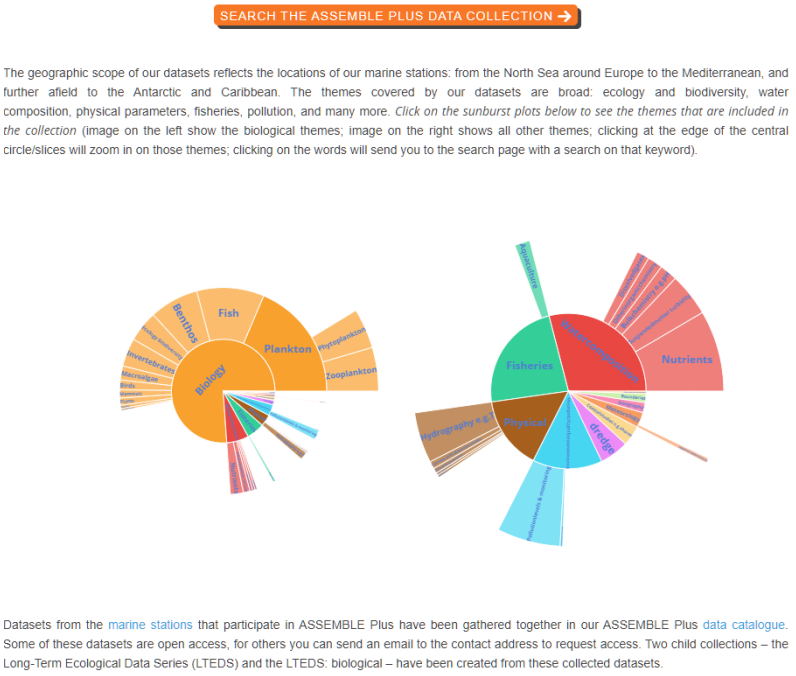

Figure 3 *Datasets landing page*

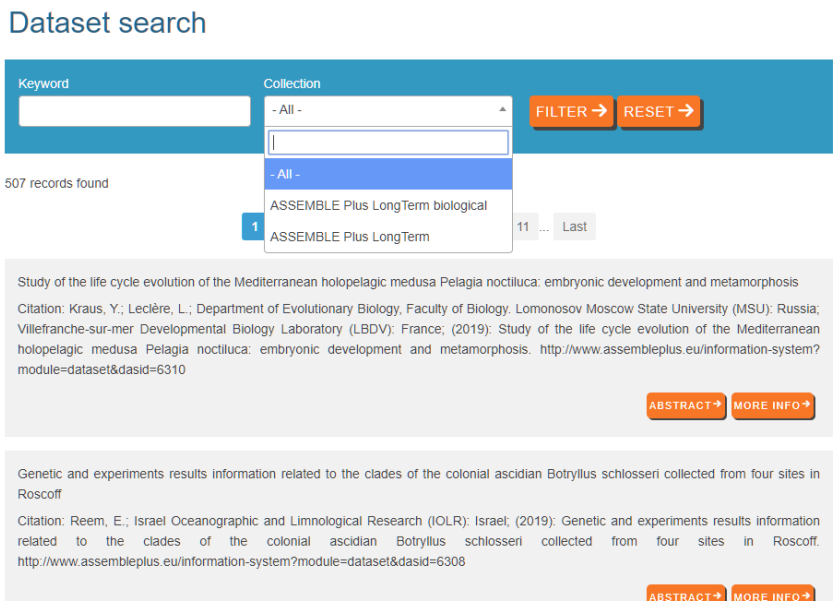

Figure 4 *Datasets collection browse page*

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.

Browsing the ASSEMBLE Plus data collection uses an IMIS API: the collection(s) to browse are identified in the API, and a limited set of search filters are added. Currently one can filter on the collection (All, Long-term, and Long-term ecological) and free-text keyword. A list of results is returned, with the title displayed. Clicking allows one to read the abstract or to open its IMIS metadata record. This metadata record includes the following information:

- Title, data creator contact details, citation, access rights (e.g. 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

As the ASSEMBLE Plus collection is quite wide in scope – scientific, temporal, geographic scope, parameter space, data types – a planned future development is to improve the filtering offered on the search page, and to categorise the search results to allow users to select on a broad range of topics (that will reflect this wide scope). This is a development that is planned within VLIZ for its IMIS data system, and we will therefore benefit from these developments.

3.2 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. 5.

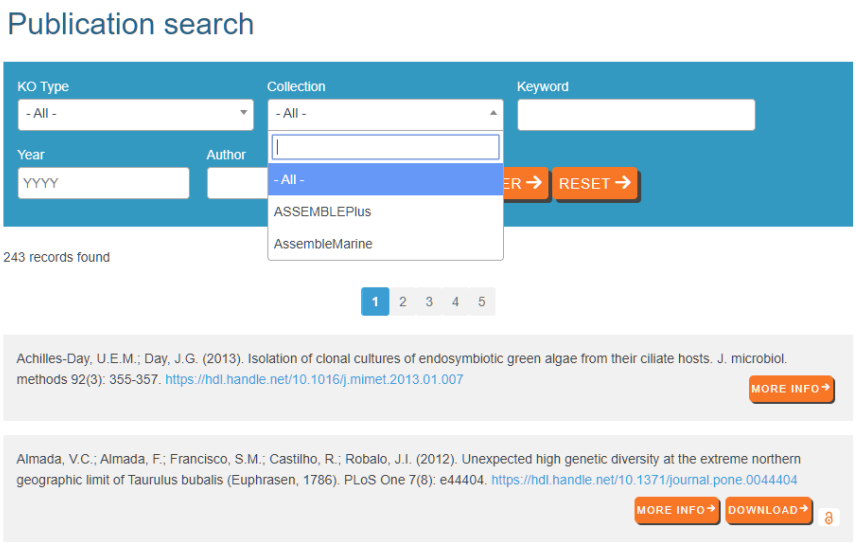

Figure 5 *Browse the catalogue of ASSEMBLE Plus publication*

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.

Browsing the ASSEMBLE Plus publications collection uses an IMIS API: the collection(s) to browse are identified in the API, and a limited set of search filters are added. Currently one can filter on the collection (All, ASSEMBLE Plus, Assemble Marine), KO type (publication, book, case study, modelling/software, prototype, exploitable result, services) 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

3.3 Planned developments

A planned future development is to improve the categorisation of the datasets and publications added to the ASSEMBLE Plus collection, to allow TNA and JRA additions to be uniquely search on, as well as the long-term additions. Some new keywords will be added to all the datasets and publications, following recommendations from experts. Additionally, more search parameters will be added to the datasets search page. Finally, preparation for handing the ASSEMBLE Plus collection for after the project ends will begin, although we emphasize that the VMDC *will* continue to provide support for this collection, cost-free, for some years after the project.

4. Hits to the virtual open access entry points

This datasets and publications access points have been page available since March 2019 (M18). The number of unique visits in the years 2019, 2020, and 2021, to the datasets search page has been 58, 158, and 76 (152), and to the publications search page have been 27, 94, and 88 (209). Almost 70% 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:

- Datasets: 1630, 153 of those being the ARMS 2018 dataset (JRA1) and 130 being the Global tide Variables (a publication of the University of Gothenburg, one of the ASSEMBLE Plus partner institutes)
- Publications: 639, with by far the most popular (109 hits) being one of our TNA projects on Density-dependent microzooplankton grazing.

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.