

EMBRC HQ Update - N.07

15th July – 15th October 2017

Prepared by: *Arnaud Delimoges and Ilaria Nardello.*

Background: *The EIB requested the ED to regularly prepare an update on the work progress, with a monthly-bimonthly frequency. The present report exceptionally covers a 3 months period (15th July-15th October), due to the Summer break.*

Scope: *to Update the EIB on progress of main items of the work activity at the Headquarters.*

Towards the ERIC

- The Commission informed us that the EMBRC-ERIC would be published on the Official Journal of the **European Union** (OJ), in Dec. 2017.
- The deadline for submission of **Support Letters** by Founding Members is the 30th of October. Countries who still need to submit a letter are: Spain, Belgium and Norway.
- The date of the **first General Assembly of EMBRC-ERIC** will be decided over the 17th EIB meeting, on the 17th of October.
- A **Launch Event** shall be organized in January 2018, in Paris.

Towards the operational phase

- **EMBRC-ERIC RoOs** have been circulated and subject to a first informal discussion, excluding policies. A new draft will be discussed at EIB 17th, on the 17 of October.
- **EMBRC-ERIC Policies** are still being shaped and may incur some important modifications over the next few weeks. Some of the policies may require a longer time for preparation, i.e. IPR policy; and: Data, Access and Dissemination Policy. Ethics and HR policies are quite final.
- **EMBRC-ERIC transfer of assets:** The UPMC transfer of assets to EMBRC-ERIC is under preparation with the different services at UPMC:

- **Host Premium In-Kind Contributions:** A hosting agreement is under preparation for the rent of the EMBRC-ERIC office spaces and related services (phone, cleaning, heating, mail, internet, network support, parking space, access to meeting rooms). A Service Agreement will describe and regulate this service provision. Terms and conditions for the HR in-kind contributions will also be defined in the Service Agreement.
- **Host Premium Monetary Contribution:** it will be transferred annually by a subsidy agreement. The budget for Headquarter personnels who are now with a UPMC employment contract will have to be assessed depending on the date of EMBRC-ERIC creation, and the status of personnels (new EMBRC contracts or continuation of the UPMC contracts until their end)
- **Member's Contributions:** all members' contributions - but France's, already received by UPMC for 2016-2017, or not yet received but with a signed financial agreement with the member, will be fully transferred to EMBRC-ERIC with a specific transfer agreement once the legal entity is created.
- The **French Member Contribution** cannot be transferred as is (Plan d'Investissement d'Avenir) but this account can remain and be used by EMBRC personnel for EMBRC expenses.
- **Projects:** The project budgets administered by UPMC on behalf of EMBRC-ERIC shall be transferred to EMBRC, once the ERIC is established and capable of participating to the project as a new beneficiary. This will be the case for Assemble Plus. The remaining overheads from the **PP2** project will be transferred through a transfer agreement. **EMBRIC** was not discussed.

EMBRC development

- **“E-infrastructure Strategy Report”:** The general aim of the Working Group on the EMBRC E-Infrastructure has been to define a strategy regarding the adoption of an e-infrastructure by EMBRC.

The “E-infrastructure Strategy Report” (Appendix 1) establishes a guiding principle for the development of the EMBRC e-infrastructure: EMBRC shall avoid building all the necessary components and, instead, adopt existing services and components that can be provided on a national level or by related European e-infrastructures, such as LifeWatch, ELIXIR, EGI, INDIGO-DataCloud, EMODnet, EurOBIS, WoRMS and Euro-BioImaging (*See Chapter 4.B.3.: Consultation with related e-infrastructures*).

A continuation of the WGEI as platform for e_infrastructure-related discussions and positioning is essential, and must be adequately resourced (*See Chapter 6.A. Short term implementation*). In particular, the costs of the various options are still underdeveloped and will need further time to be developed.

The WGEI will meet on the 18th of October in Paris to exchange on the update of the WGEI report (March 2017), with a focus on the cost modelling.

- **EU Projects Working Group:** As the number of projects in which EMBRC is involved grows, there is a need for more fully and widely appreciating the projects’ outputs, to facilitate a synergistic approach and maximize projects’ impact. For the purpose, a new work group has been established: the “Projects Working Group (PWG)”. It gathers the Scientific Projects Managers of the projects connected to EMBRC.

The PWG first met on the 13th of October in Paris and will meet every two months.

List of Members:

- Stecy Jombert and Florence Guillot, Assemble Plus
- Tim Deprez, IMBRSea
- Nicolas Pade, ENVRIPlus
- Mery Pina, EMBRIC
- Cristina Secadas, EBB
- Ilaria Nardello, Assemble plus, CORBEL & BMS RIs Strategic Group

ESFRI Report

- We have been requested to provide some further info in relation to the **ESFRI monitoring of EMBRC**:
 - Have there been any new documents released since April 2017 in relation to collaboration/interactions with industry, ERA-NETs and JPIs, socio-economic impact studies or reports, and e-infrastructure and e-needs? If yes, please provide them.
 - Have any new countries or partners joined the Infrastructure since April 2017?
 - What is the current status towards establishing the ERIC? Which countries have signed the step 2 ERIC-application?
 - Do you have metrics for all Key Performance Indicators? If yes, please provide them.

We shall organise this response by the 5th of November. ESFRI will then analyse the additional documents by the end of November and will invite us to a hearing. The hearing will take place in Brussels between the 29th of January and the 2nd of February 2018.

- A draft version of the Work Programme 2018-2020 for Research Infrastructures has just been published: <https://ec.europa.eu/programmes/horizon2020/en/european-research-infrastructures-including-e-infrastructures---horizon-2020-work-programme-2018>
- EU Staff Working Document on the long-term sustainability of Research Infrastructures: <http://www.esfri.eu/ri-world-news/staff-working-document-long-term-sustainability-research-infrastructures>

State of the Union

- **Finland**: No further news.

- **Ireland:** Unwilling to invest unless the research community (partly) supports the effort.

HR Recruitments

- **Scientific Communication and Access officer:** Florence Guillot was recruited on the 1st of September as EMBRC Scientific Communication Officer (0.50FTE) & Access Officer (0.50FTE).
- **Legal Officer:** Melindy Matthews has joined the EMBRC Secretariat as EMBRC legal officer, since the 1st of October, as part of UPMC in-kind contribution (0.30FTE). Melindy is already reviewing the EMBRC-ERIC Rules of Operations and related policies. Oskar Ozun contribution, instead, has now concluded.

Communication

- **Communication materials**

Several high-impact communication materials were recently developed in collaboration with a communication agency:

- Business Plan Update 2017;
- High-Level brochure;
- Service Brochure;
- Pull-up banner;
- PPT presentations.

The electronic and print-ready versions of the documents are available, including printing specs, on the shared Dropbox folder “Comm_Material”: *-this will be communicated over email.*

Past Events

- **Oceans meeting:** « The ocean and human health », 7-8 September, Lisbon. Pitch in the business session; and booth next to CCMAR. Lucie Salvaudon presented.
- **EMBRIC General Assembly**, 11-14 September, Faro.
- **OECD Workshop:** 11 October, SZN, Naples "Innovation for a sustainable economy", highlights on the role of research-based innovation against unsustainable blue growth. Ilaria Nardello invited to attend.
- **Seminar “Les rendez-vous de Concarneau”:** 20-21st September, Station Marine de Concarneau, Biomaterials on Biotechnologies. Florence Guillot in attendance.
- **Workshop ‘Food System Microbiome’**, 29 September 2017 in Brussels. The aim of this workshop was to better understand the needs of microbiome R&I. The workshop gathered a limited number of stakeholders from various areas of the food systems microbiome (soil and plant, animal, human and marine respectively). Ilaria Nardello invited to attend.

Upcoming Events

- **CORBEL General Assembly**, 25-26 October, Amsterdam, NL
- **BMS Strategic Group meeting**, 26 October, Amsterdam, NL
- **ERIC Network**, 16-17 November, Gratz, AT

EU Projects

- **Assemble+ KoM**, 19th-20th October 2017, Paris. About 50 people will be attending.
- **EBB KoM**, 22-23 November., Vigo, ES

Committee of the Nodes Meetings (CoN)

On the 20-21th of September, the CoN VII meeting was held at SZN, Naples. Discussions took place on Assemble Plus, the upcoming funding opportunities and Communication.

On the 20th of October, a “mini-CoN” will take place after Assemble Plus meeting, focusing on future funding opportunities for the EMBRC-ERIC Community.

Next EIB/GA Meetings

- **October 17th, 2017:** 17th EIB Meeting, in Paris
- **December 12th-13th, 2017:** 18th EIB Meeting, venue to be established

Annex 1: Working Group e-Infrastructure Report



Strategy Report for the EMBRC e-Infrastructure

May 2017

Contributors:

Stefanie Dekeyzer, stefanie.dekeyzer@vliz.be
Klaas Deneudt, klaas.deneudt@vliz.be
Mark Hoebeke, mark.hoebeke@sb-roscoff.fr
Erwan Corre, corre@sb-roscoff.fr
Dan Lear, dble@MBA.ac.uk
Lennert Tyberghein, lennert.tyberghein@vliz.be
Ilaria Nardello, ilaria.nardello@upmc.fr
Arnaud Delimoges, arnaud.delimoges@upmc.fr
Marco Borra, marco.borra@szn.it
Georgios Kotoulas, kotoulas@hcmr.gr
Claire Gachon, Claire.Gachon@sams.ac.uk

Revision	Date	Modification	Author
V1.0	May 2017	First draft	EMBRC WGEI

Acknowledgements:

EMBRC WGEI would like to thank the following invited experts:

- **Simon Claus** (Flanders Marine Institute, VLIZ) – European Marine Data and Observation Network Biology (EMODnet Biology);
- **Petra ten Hoopen** (European Bioinformatics Institute; EMBL-EBI) – ELIXIR;
- **Christos Arvanitidis** (Institute of Marine Biology, Biotechnology and Aquaculture (HCMR-IMBBC) – LifeWatch;
- **Jesus Marco de Lucas** (Institute of Physics of Cantabria, IFCA) – European Grid Infrastructure (EGI), LifeWatch Competence Center; INDIGO-DataCloud;
- **Frank Oliver Glockner/Renzo Kottmann** (Max Plank Institute for Marine Microbiology, MPIMM);
- **Francisco Hernandez** (Flanders Marine Institute, VLIZ) – World Register of Marine Species (WoRMS) and European Ocean Biogeographic Information System (EurOBIS);
- **Wiro Niessen** (Technical University Eindhoven) – Euro-BiolImaging.

Table of content:

1. Executive summary
 2. Introduction
 3. Methodology and approach
 4. E-infrastructure requirements of EMBRC
 - A. View from preparatory phase
 - B. Use cases and required components
 - B.1 Selection and analysis of use cases*
 - B.2 Identification of general and specific requirements*
 - B.3 Consultation with related e-infrastructures*
 - B.4 Prioritization*
 5. E-infrastructure architecture
 - A. Recommended EMBRC developments
 - B. Architecture model
 6. Short term implementation plan and long term vision
 - A. Short term implementation
 - B. Long term vision
-
- ANNEX 1: Specific e-infrastructure requirements for each use case
- ANNEX 2: Specific e-infrastructure requirements for the planned developments within ASSEMBLE Plus
- ANNEX 3: List of required e-infrastructure components (“EMBRC shopping list”)

1. Executive summary

This report collates the work of the EMBRC Working Group on E-Infrastructures (WGEI) during a series of meetings held in 2016-2017. The general objective of the WGEI has been to define a strategy regarding the use and provision of an e-infrastructure by EMBRC.

More specifically WGEI has organized meetings, discussions and consultations with the following objectives in mind:

- To review and revise the e-infrastructure strategy regarding the information management system underpinning the functioning of EMBRC, their data custody and data use;
- To define e-infrastructure architecture model options based on the current developments and landscape of existing e-infrastructures, internal and external to EMBRC;
- To assess feasibility and cost implications of the e-infrastructure strategy implementation, with the application of different implementation scenarios.

To achieve these objectives, a **stepwise approach** was followed: (1) Review of available literature and inventory of existing capabilities; (2) Identification of use cases; (3) Identification of general and specific requirements; (4) Consultation related e-infrastructures; (5) Processing available information and discussion; (6) Defining e-infrastructure architecture; (7) Prioritization; (8) Implementation scenarios; (9) Development and cost implications; (10) Condensation into report. See further Chapter 3: Methodology and approach.

Important groundwork on the strategy to institute a suitable e-infrastructure for EMBRC has been laid out during the preparatory phase of EMBRC. This exercise builds on the output of the preparatory phase and expands this work towards a practical implementation plan. The **main goals** of the EMBRC e-infrastructure remain as previously envisaged: To ensure that operation of EMBRC is supported by an adequate e-infrastructure. To ensure that EMBRC partners and external users have the access to each of those components to allow marine researchers to undertake research projects and then process, store, analyze and make publicly available the large volumes of data that are generated at EMBRC marine stations and laboratories. The previously identified needs in terms of network access, storage capacity, computational resources, software, and human resources are largely confirmed. However, where the pp1EMBRC reports define the EMBRC data collection as primarily molecular data and identify the need for a central or distributed EMBRC bioinformatics facility, WGEI recognizes a broader and more diverse

data collection and hence data management needs. In addition, the landscape of e-infrastructures at European level is rapidly evolving. For a large share of the identified e-infrastructure resources required at the local and central level, the EMBRC strategy should be to make optimal use of European e-infrastructure developments like the E-infrastructure Commons and the European Open Science Cloud.

As part of the followed procedure to identify the e-infrastructure requirements, a number of scientific and managerial **use cases** were selected from a variety of sources: pp1EMBRC, pp2EMBRC, EMBRC HQ, EMBRIC, ASSEMBLE Plus JRA's, etc. Based on the analysis of requirements for these use cases, a list of needed **e-infrastructure components** was created. Components were identified for following categories: Administrative tools; Registers and catalogues; Knowledge output module; Repositories; Integrated thematic databases; Analysis tools; Human resources; Local databases; Data storage and computing; Networking and connectivity; and Training. (See Chapters 4.B.1. *Selection and analysis of use cases* and 4.B.2. *Identification of general and specific requirements*).

A guiding principle for the establishment of the EMBRC e-infrastructure is that EMBRC will not develop or supply all components of the required e-infrastructure itself. Instead EMBRC will make beneficial use of existing services and components that can be provided on a national level or by related European e-infrastructures, such as LifeWatch, ELIXIR, EGI, INDIGO-DataCloud, EMODnet, EurOBIS, WoRMS and Euro-Biolmaging. A number of **experts were consulted** in order to map the EMBRC e-infrastructure requirements with the current landscape of e-infrastructures in Europe and describe the relevant services they can provide. (See Chapter 4.B.3. *Consultation with related e-infrastructures*).

For each of the components, there was an evaluation of the **priority of development**, whether or not these components should be developed by EMBRC or could be co-developed or simply used from existing developments outside of EMBRC. (See Chapter 4.B.4. *Prioritization*).

Based on the combination and interaction of the identified e-infrastructure components, a **general architecture model** was drawn up and put forward. This model is based on the guiding principle – already put forward in pp1EMBRC – that the general strategy approach for EMBRC regarding e-infrastructure is to work towards interoperability with existing thematic initiatives, and to pursue co-development regarding lacking components. The proposed architecture recommends the implementation of a number of components at the central level that can support a more integrative approach regarding the processes at the local level and a better interaction between the local operators and thematic initiatives. The e-infrastructure architecture model suggests a two-way interaction between EMBRC and related e-infrastructures such as LifeWatch, ELIXIR,

EMODnet and Pangaea. The EMBRC nodes receive the necessary training to comply with standards and make use of European e-infrastructures and their components. These e-infrastructures should take into account the specific requirements of the EMBRC community and provide visibility to EMBRC output. (See Chapter 5. *E-infrastructure architecture*).

An assessment was made of the feasibility and cost implications of the e-infrastructure **implementation**. Some of the top priority developments at the central level have already started, for example the EMBRC portal (www.embrc.eu), the service discovery and access request system and the European Marine Training Portal (www.marinettraining.eu). Other priority developments such as a Literature register, Expert register, Datasets register, Knowledge output module, Dataset and raw data file repository and Virtual analysis platform are planned in the framework of the ASSEMBLE Plus Horizon 2020 project. The integration of an Event management system is developed as part of the EMBRC website upgrade. (See Chapter 6.A. *Short term implementation*).

The remaining components that were considered high priority but currently not planned are a Service allocation system, an User registry, Local and shared data storage and computing capacity, and Local and central networking and connectivity. Other high priority components include IT staff, Liaison officers and Trainers. (See Chapter 6.A. *Short term implementation*).

Cost assessment of the remaining components was considered extremely difficult. Especially since the functional requirements of these components have not been studied in detail. E.g. for the administrative tools, factors such as edition (server-based versus cloud-based), payment plan (free tools, monthly versus yearly costs or one-time payment), and the amount of intended users need to be taken into account, as these can severely alter the pricing. Also additional costs for training, deployment and maintenance should be considered. A very rough cost estimate for the administrative tools ranges between very low costs to several tens and even hundreds of thousands euros. More follow-up discussions are needed to assess the additional components to be developed at the central level regarding registers and catalogues, integrated thematic databases and analysis tools. Networking and storage are considered a more local issue until the central services are fully developed. (See Chapter 6.A. *Short term implementation*).

Important efforts and actions are planned or ongoing to advance the EMBRC e-infrastructure in the short term. Positive dialogues have been initiated with related e-infrastructures. Mutual opportunities should be identified and formalized in formal agreements. The European e-infrastructure landscape is rapidly evolving and EMBRC should investigate how to take advantage of all arising opportunities that can fulfil the EMBRC community requirements and can support the EMBRC e-infrastructure offer. A

continuation of the WGEI as platform for e-infrastructure related discussions and positioning is seen as essential, however must be adequately resourced. (*See Chapter 6.A. Short term implementation*).

2. Introduction

The European Marine Biological Resource Centre (EMBRC) is a distributed research infrastructure that aims to provide a strategic delivery mechanism for excellent and large-scale marine science in Europe. With its services, EMBRC will support both fundamental and applied research based on marine bio-resources and marine ecosystems. In particular, EMBRC aims to drive forward the development of blue biotechnologies. EMBRC will provide the suitable research environment for users from academia, industry, technology and additional sectors.

EMBRC was created to provide value-added access to European marine resources (ecosystems, organisms, and research platforms) for private and academic scientists. The background and goals are extensively described in the EMBRC Scientific Strategy Report. In summary:

- Offer users from academia and the private sector access to a portfolio of research platforms, biological resources, analytical services and data (TSD, 2.1, 2.2 and 3.2);
- Develop integrated workflows of high quality services for access to biological, analytical and data resources by deploying common underpinning technologies and practices (TSD 3.3);
- Strengthen the connection of science with industry through a coordinated knowledge and technology transfer service (TSD 3.4);
- Offer training facilities and courses for researchers and technical personnel (TSD 3.5);
- Collaborate and engage European maritime regions in the development and integration of EMBRC and contribute to de-fragment their RDI policies (TSD 3.6).

Fulfilment of these goals will require the support of an adequate system (the EMBRC e-infrastructure), available to all partners, including elements such as network access, storage capacity, computational resources, software, and related human resources. In this framework, the EMBRC Working Group on E-Infrastructures (WGEI) was formed.

3. Methodology and approach

In order to come to a practical implementation plan, the EMBRC e-Infrastructure Working Group (WGEI) built on the groundwork that has been laid out during the preparatory phase of EMBRC. The WGEI passed several steps considered essential to reach the outlined objectives:

STEP 1: Review of available literature and inventory of existing capabilities

Relevant documents were consulted and discussed:

- D2.5 - EMBRC Scientific Strategy Report;
- D3.1 - Report on e-infrastructure requirements and e-workflow scenarios;
- D3.2 - Detailed evaluation of potential e-workflow scenarios;
- D3.3 - Plan for the requirements of the EMBRC e-infrastructure;
- D11.10 - Plans for digital library;
- pp2EMBRC Inventory of Services
- Proposal Association of European Marine Biological Laboratories Expanded
- EMBRC e-infrastructure Working Group Terms of Reference.

STEP 2: Selection of use cases

A number of scientific and managerial use cases were selected from a variety of sources: pp1EMBRC, pp2EMBRC, EMBRC HQ, EMBRIC, ASSEMBLE Plus JRA's, etc.

STEP 3: Identification of general and specific requirements

Based on the analysis of requirements the identified use cases, a list of 35 unique required e-infrastructure components was created. This list is referred to as the "EMBRC e-infrastructure shopping list".

STEP 4: Consultation related e-infrastructures

A guiding principle for the establishment of the EMBRC e-infrastructure is that EMBRC will not develop or supply all components of the required e-infrastructure itself. Instead EMBRC will make beneficial use of existing services and components that can be provided on a national level or by related European e-infrastructures.

A number of experts were consulted in order to map the EMBRC e-infrastructure requirements with the current landscape of e-infrastructures in Europe and describe the relevant services they can provide.

WGEI organized five 2-day workshops and one WebEx conference to prepare this work: Meetings I (July 4th-5th 2016), II (September 7th-8th 2016), III (October 26th-27th 2016), IV (December 14th-15th 2016), V (February 6th 2017, through WebEx), and VI (February 23rd-24th 2017). The experts were invited to present and illustrate their viewpoints during meetings I and IV.

The following experts were called:

- *Simon Claus* (Flanders Marine Institute, VLIZ) – European Marine Data and Observation Network Biology (EMODnet Biology);
- *Petra ten Hoopen* (European Bioinformatics Institute, EMBL-EBI) – ELIXIR;
- *Christos Arvanitidis* (Institute of Marine Biology, Biotechnology and Aquaculture, HCMR-IMBBC) – LifeWatch;
- *Jesus Marco de Lucas* (Institute of Physics of Cantabria, IFCA) – European Grid Infrastructure (EGI), LifeWatch competence center; INDIGO-DataCloud;
- *Frank Oliver Glockner/Renzo Kottmann* (Max Plank Institute for Marine Microbiology, MPIMM);
- *Francisco Hernandez* (Flanders Marine Institute, VLIZ) – World Register of Marine Species (WoRMS), European Ocean Biogeographic Information System (EurOBIS) (unable to attend);
- *Wiro Niessen* (Technical University Eindhoven – Euro-BioImaging (unable to attend).

STEP 5: Processing available information and discussion

Based on the input of the previous steps, the WGEI formulated a number of options for the architecture model of the EMBRC e-infrastructure. This model names the different building blocks of the e-infrastructure and describes the relations between these components.

Pros and cons of the different options were discussed during Meetings IV (December 14th-15th 2016), V (February 6th 2017, through WebEx), and VI (February 23rd-24th 2017) of WGEI. The discussion included considerations regarding competition and complementarity with related RI's.

STEP 6: Defining e-infrastructure architecture

Based on the combination and interaction of the identified e-infrastructure components a general architecture model was drawn up and put forward. This model is based on the guiding principle –already put forward in pp1EMBRC, that the general strategy approach for EMBRC regarding e-infrastructure in general is to work towards interoperability with existing thematic initiatives and to pursue co-development regarding lacking components. The proposed architecture recommends the implementation of a number of components at the central level that can support a more integrative approach regarding the processes at the local level and the interaction between the local operators and thematic initiatives.

STEP 7: Prioritization

For each of the components, there was an evaluation of the priority of development, whether or not these components should be developed by EMBRC or could be co-developed or simply used from existing developments outside of EMBRC.

STEP 8: Implementation scenario's

A short term implementation plan describes the plans in place to set up the highest priority components. In addition a long term vision presents a view on positioning the EMBRC e-infrastructure in the evolving European landscape.

STEP 9: Development and cost implications

Some very approximate cost implications were put forward for part of the e-infrastructure development. More detailed cost definitions need a detailed study of the functional requirements of the components to be developed.

STEP 10: Condensation into report

The output of Steps 1 to 9 is assimilated in this EMBRC Working Group on E-Infrastructures (WGEI) report to be delivered to EMBRC HQ for review by the appropriate governing bodies.

Figure 1 gives an overview of the updated EMBRC WGEI strategy and methodology:

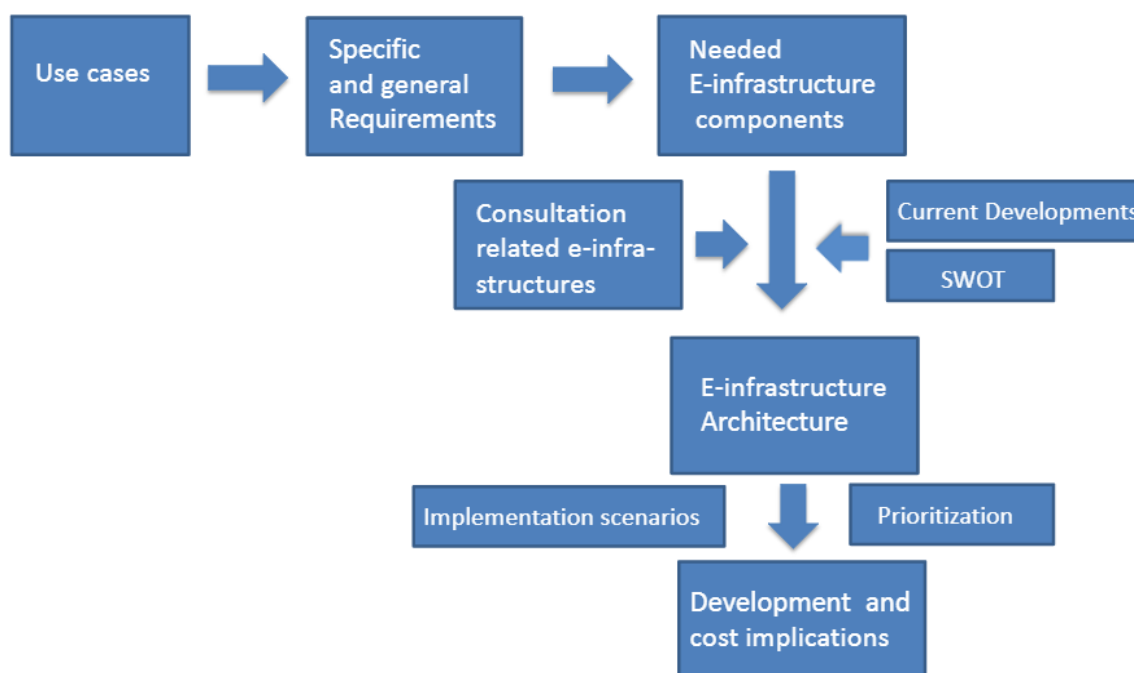


Figure 1 – Strategy of the EMBRC Working Group on E-Infrastructures (WGEI)

4. E-infrastructure requirements of EMBRC

An important task of EMBRC WGEI is to further define the niche of the EMBRC e-infrastructure and the relation and interaction with existing initiatives such as LifeWatch, ELIXIR, ENVRIplus, CORBEL, etc. A better understanding of the e-infrastructure requirements is essential to identify where collaboration with these existing e-infrastructure initiatives can be realized. Significant work was already done during EMBRC preparatory phase 1 (pp1EMBRC) in describing needs and context for the EMBRC e-infrastructure. During WGEI meeting I (July 4th-5th 2016), it was suggested to continue on the development of specific use cases to identify the EMBRC e-infrastructure requirements, as was started during the pp1EMBRC project. Furthermore, several relevant e-infrastructure developments are scheduled or ongoing in the framework of pp2EMBRC, EMBRIC and ASSEMBLE Plus.

A. View from preparatory phase

The **main goals** of the EMBRC e-infrastructure remain as previously envisaged: To ensure that operation of EMBRC is supported by an adequate e-infrastructure. To ensure that EMBRC partners and external users have the access to each of those components to allow

marine researchers to undertake research projects and then process, store, analyze and make publicly available the large volumes of data that are generated at EMBRC marine stations and laboratories.

Deliverable **D2.5 – EMBRC Scientific Strategy Report** already provided several statements, recommendations and guidelines regarding data and e-infrastructure requirements:

- Infrastructure priorities include: (1) Creation of a web-based searchable EMBRC services, facilities and organism database; (2) Creation of a best practice database to achieve high and compatible standards in experimental methods, culture and husbandry, data collection and analysis; (3) Establishing bioinformatics platforms to meet the demand for 'omics-based research in accordance with WP3 e-infrastructure strategy (D3.3) and improved communication with RIs developing such facilities (e.g. ELIXIR); and (4) Implementation of a mechanism to coordinate the access program for mobility of researchers.
- Service priorities include: (1) Cementing EMBRC in the strategic research landscape in Europe, through close collaboration and dialogue with other RIs, EU funded projects and national and international research projects; (2) Creating a knowledge and technology transfer (KTT- platform to provide industry and SMEs with a go-to service for accessing scientific data and research relevant for their R&D; (3) Joint Development Activities (JDA); and (4) Training.
- EMBRC will become a gateway to the data that are regularly gathered into time series of marine species and habitats, many of which have been regularly sampled for decades and thus comprise unique long-term databases of environmental change.
- EMBRC should establish an EMBRC Services Database of habitats, organisms and cultures, platforms and equipment, databases and local expertise. In order to provide a fast and effective overview of available services for users, it is important that EMBRC advertises and lists the services provided. The database should be searchable online through the EMBRC website. This EMBRC database will be of great importance to EMBRC users and the research community as a whole. It should serve as the first point of call for interested users to explore the facilities and services available in the EMBRC access program. The EMBRC database could be used to inform other directories and existing databases, such as EUNIS or EUROCEAN and thereby provide a comprehensive service for a large user community. The database should be constructed and managed by suitably qualified IT expertise in connection with website development. An accurate, first version of the database linked with the EMBRC website must be a high priority of ipEMBRC.

Throughout operation the database must be regularly reviewed and updated to ensure accurate display of EMBRC services.

- EMBRC should harmonize long-term datasets and data collection. Currently the provision of information and data lack an integrated European approach and large scale visibility. Ecological databases range in space and time scales from those developed in the frame of specific research programs to large scale European databases (e.g. MarBEF). A number of initiatives are required so that standards and quality control can be readily assessed for implementation across EMBRC, for (1) metadata collection and standardization (e.g. INSPIRE directive, MEDIN), (2) taxa names (e.g. WoRMS), (3) habitats classification (e.g. EUNIS) (4) data acquisition (e.g. ICES, QUASIMEME, Genomics Standards Consortium (GSC), TARA OCEANS recommendations and protocols, or more locally Interreg Programs). Existing large scale databases such as the international databases PANGEA or MarBEF and more local initiatives (e.g. VLIZ, Flanders Marine Institute), should also be considered. Close collaboration with ESFRI initiatives, such as ELIXIR and LIFEWATCH, will play a crucial role in determining how data is managed, shared and processed within EMBRC.

- It is essential that the drive to participate in and utilize European-wide database initiatives is maintained within EMBRC. This will facilitate the integration of ecological data acquired by European marine centers and improve access to the user community. Standardizing data and collection methodology is an important step towards ensuring that Europe remains a leading force in marine ecological research. However, the adoption of standards and their implementation that may involve large-scale changes to procedures, data collection and the formulation of new guidelines, require coordination across centers and time. It was recommended that a workshop is organized at the onset of the implementation phase to discuss database and standardization options with the aim of implementing the agreed improvements during the implementation phase in the medium term by 2017.

- Many advances in biotechnology are underpinned by 'omics technologies. [...] Within Europe, there are well-established capabilities comprising research vessel, husbandry, organisms, access to ecosystems, mesocosms and long-term time series datasets that will underpin the potential of 'omics. This is backed by considerable expertise in biochemistry, metabolism, physiology and ecology of marine organisms that is essential for understanding, interpretation, and annotation of genomes and other types of high volume data. [...] A further significant driver for 'omics technologies is the unparalleled growth in capacity for data gathering and analysis and the decreasing costs of data acquisition and storage. A range of platforms for sequencing and analysis are accessible at national and international levels. There are a number of national 'omics facilities that

serve partner organizations, but these are not currently available to the EMBRC community as a whole. The majority of high-end sequencing platforms are not specifically marine-oriented. 'Omics-related facilities that provide functional analyses (e.g. microscopy, NMR imaging and spectroscopy, cell biology, culture collections, cryopreservation facilities) are often by necessity center-specific and widely dispersed. There is a clear need for better provision of advice on where to go for the best or most appropriate service. EMBRC will need to develop an understanding of the provision for sequencing and other 'omics data in Europe in order to provide advice to its users.

- There is a well-recognized and increasing dearth of expertise relating to bioinformatics analysis and biological interpretation of 'omics data among European institutions. In particular there are increasing training needs in the areas of annotation, curation, assembly (including use of software tools), genome structure and organization and experimental design and statistics. Raising awareness of the problems of data analysis is a major challenge. There is also a perceived lack of integration of dispersed resources and an increasing need for standardization of collection and analysis. There is a lack of international data sharing agreements and a potential longer term problem relating to data transfer and storage which underlines the need for input into the development of the EMBRC e-infrastructure network.
- Better integration between long term and historical datasets, biological resource centers and 'omics repositories is needed.
- EMBRC should develop tools and technologies for 'omics. The coordination of 'omics tools and capability throughout Europe and establishment of marine-oriented 'omics standards for design and analysis needs to be facilitated to provide a framework upon which 'omics approaches will be developed throughout the infrastructure. These include: (1) The development of standardized 'omics related technologies, specific for marine organisms, such as sample preparation, storage, quality control and validation; and (2) The development of coordinated 'omics data handling tools and expertise.
- Training in data handling and analysis (curation, assembly (including use of software tools), genome structure and organization and experimental design (including arrays, RNA sequencing) and statistics) should be prioritized over data generation in the immediate and medium terms. This should be done in conjunction with ELIXIR.
- Joint Development Activities (JDAs) will be highly instrumental in integrating the partners of EMBRC into a coherent distributed international research infrastructure offering high quality services to diverse user communities. Where appropriate, they will be conducted in close consultation with complementary infrastructures (e.g. ELIXIR, MIRRI, AquaExcel, LifeWatch), in order to cross-fertilize best practices and avoid

unnecessary duplication of effort. [...] This will be achieved via a coordinated, long-term R&D program involving collaboration between all EMBRC partners, designed to: (1) Facilitate high-priority technological and methodological breakthroughs for the collection, long-term ex situ maintenance and transport of live unicellular and multicellular marine organisms and for development of culture facilities capable of better simulating natural environments; (2) Create genetic resources such as mutant collections and transgenic lines for flagship eukaryotic and prokaryotic model organisms; (3) Adapt and develop 'omics and related methods for high throughput environmental biodiversity studies and functional exploration of marine models; (4) Develop the transversal e-infrastructure component of EMBRC that will be critical both as a means of managing the resources and as a major tool to facilitate exploitation of the resources; and (5) Develop new models with high ecological relevance (e.g. representing key functions in the ecosystem) along with the tools needed for advanced research including permanent cultures of ecotypes and inbred lines, full genome sequence information, genetic tools for functional genomics, pipelines for phenotypic characterization, and a database providing access to relevant genetic and ecological data.

- Due to the many initiatives involved with data management, handling and management, co-ordination of e-infrastructure between RIs is of key importance. EMBRC must keep abreast of developments in e-infrastructure and ensure compliance and engagement with relevant initiatives, in particular ELIXIR.

During pp1EMBRC, the work packages on e-infrastructure requirements (WP3) produced three deliverables: D3.1, D3.2 and D3.3. As the blueprint of the EMBRC e-infrastructure, **three possible models** were suggested in deliverable **D3.3 - Plan for the requirements of the EMBRC e-infrastructure**:

1. Central facility for e-infrastructure
2. Evenly distributed e-infrastructure across EMBRC nodes
3. Two-level integrated e-infrastructure (mixed model)

The third model (mixed model) was recommended as the preferable approach for the EMBRC e-infrastructure and bioinformatics facility. WGEI recognizes the fact that e-infrastructure work and requirements are dictated both at the central and the local level. Because of the distributed nature of EMBRC, a large part of the e-infrastructure needs are situated at the local level and are catered for by EMBRC operators. However, EMBRC has the ambition to improve and streamline some of the e-infrastructure related processes within its community and towards users and related RI. Therefore the third model (mixed model) indeed seems the most appropriate way forward.

Furthermore, D3.3 described that the EMBRC e-infrastructure should have **4 functions**:

1. Providing advice to users on the design and implementation of experiments and analysis of data
2. Providing data storage and computational infrastructure for analysis of active research projects
3. Training: running courses and providing instruction to enable EMBRC users to process their own data
4. Creating and maintaining a data policy for EMBRC

All functions have been discussed by the WGEI. Functions 2,3 and 4 are considered crucial functions of the e-infrastructure to be developed. Function 1 relates to function 3, but would need the establishment of specific expert groups since the expertise is widely spread within the EMBRC community.

The previously **identified needs** in terms of network access, storage capacity, computational resources, software, and human resources are largely confirmed. However, where the pp1EMBRC reports define the EMBRC data collection as primarily molecular data and focus on the need for a EMBRC bioinformatics facility, WGEI sees a broader and more diverse data collection and hence data management needs.

Figure 2 shows the **strategic priorities** of the EMBRC e-infrastructure, as they were presented in the pp1EMBRC Strategy Report. Several of these strategic priorities relate to e-infrastructure components and should be part of the e-infrastructure model.

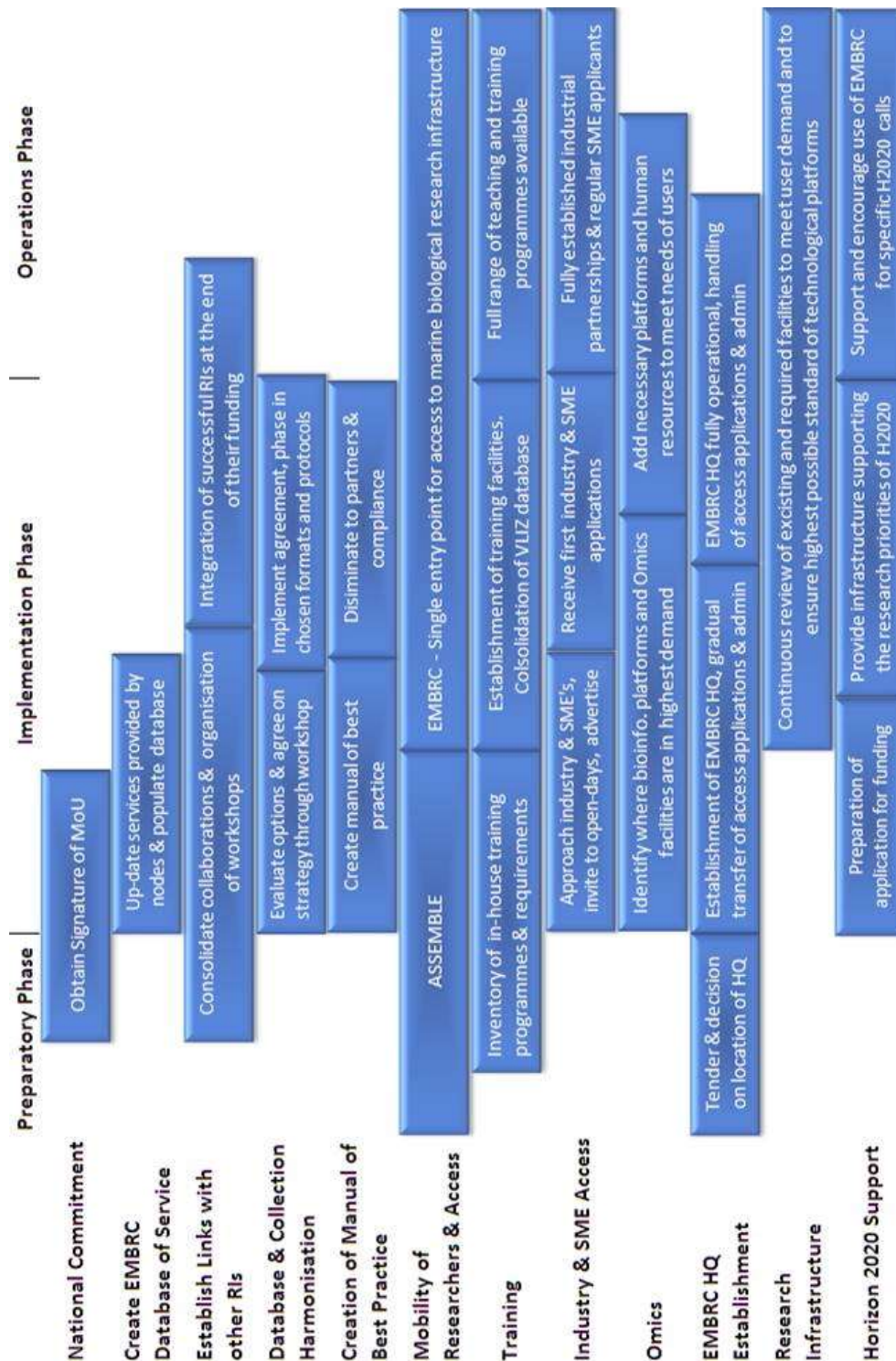


Figure 2 – Strategic priorities of the EMBRC Infrastructure, as presented in the pp1EMBRC Strategy Report

Use cases and required components

B.1. Selection and analysis of use cases

As part of the followed procedure to identify the e-infrastructure requirements, a number of scientific and managerial **use cases** were selected from a variety of sources: pp1EMBRC, pp2EMBRC, EMBRC WGEI, EMBRIC, ASSEMBLE Plus JRA's, etc. (Table 1).

For an overview of the specific e-infrastructure requirements for each use case, we refer to Annex 1.

Table 1 – Overview of identified use cases

Use case	Identified by
Sequencing genomes and/or transcriptomes using second and third generation sequencing technologies, with <i>Platynereis dumerilii</i> as an example. In this scenario it is assumed virtually nothing is previously known about the genome/transcriptome in question.	pp1EMBRC
Developing new model organisms, with <i>Clytia hemisphaerica</i> , <i>Phallusia mammillata</i> and <i>Patiria minata</i> as examples.	pp1EMBRC
Species distribution modelling based on collection data and molecular data, adapted to high latitude pelagic habitats.	pp1EMBRC
Using historical datasets to augment 'omics data to understand ecological change over time.	pp1EMBRC
Ocean Sampling Day (OSD)	MicroB3/OSD
Service access system	pp2EMBRC

Improving virtual access to marine biological stations data, information and knowledge	ASSEMBLE Plus
Configurator (connecting several databases: nucleotide data – proteomic data – microbial data – chemical data)	EMBRIC WP4
Microbial pipeline from environments to active compounds	EMBRIC WP6
High performance microalgae for blue technological applications	EMBRIC WP7
Ecosystem assessment and mapping (MSFD-related) use cases	EMBRC WGEI
Digital library – Retrieving or providing overview of EMBRC related publications	EMBRC WGEI
Management – Central management of EMBRC	EMBRC WGEI
Long-term monitoring	EMBRC WGEI
Education and training – Building capacity and applying standards within the EMBRC community.	EMBRC WGEI

B.2. Identification of general and specific requirements

Many recurring requirements within the use cases could be listed as general requirements for all use cases. The following categories for **general requirements** were selected:

- **Tools and services (user applications):** Generic services for research providing support for research workflows using combinations of the above (virtual research environments).

- **Data (data layer and services):** Middleware components to enable the seamless use of the above services, including authentication and authorization.
- **Computing:** Access to high performance computing (supercomputing) and high-throughput computing.
- **Storage:** Access to high end storage for ever increasingly large data sets.
- **Networking and connectivity (including security):** Advanced networking services to connect computing and storage resources to users and instruments.

In a second step, a list of 35 unique **required e-infrastructure components** was distilled based on the analysis of the use case requirements (see below). For a definition of each component, we refer to Annex 3. This list of required e-infrastructure components will be further referred to as the “**EMBRC e-infrastructure shopping list**”.

Components were identified for the following categories: Administrative tools; Registers and catalogues; Knowledge output module; Dataset and raw data file repositories; Integrated thematic databases; Analysis tools; Human resources; Local databases; Data storage and computing; Networking and connectivity; and Training.

Administrative tools:

- Service request system
- Service allocation system
- Financial administration system
- Project management system
- Document management system
- Event management system
- User registry

Registers and catalogues:

- Central service register
- Sample register:
- Literature register
- Expert register

- Cultures register
- Analysis methods register
- Datasets register
- Training register

Knowledge output module

Dataset and raw data file repositories

Integrated thematic databases:

- Sequence data database
- Reference molecular data database
- Taxon observation data database
- Ecological and environmental data database

Analysis tools:

- Sequence data processing tools
- Virtual analysis platform

Human resources:

- Bio-informaticians
- Data scientists
- IT staff
- Liaison officers

Local databases:

- Local monitoring databases
- Lab or Field Information System

Data storage and computing capacity:

- Local
- Shared

Networking and connectivity:

- Local
- Central

Training:

- Online training platform
- Trainers

B.3. Consultation with related e-infrastructures

A guiding principle for the establishment of the EMBRC e-infrastructure is that EMBRC will not develop or supply all components of the required e-infrastructure itself. Instead EMBRC will **make beneficial use of existing services and components** that can be provided on a national level or by related European e-infrastructures. A number of **experts** were **consulted** in order to map the EMBRC e-infrastructure requirements with the current landscape of e-infrastructures in Europe and describe the relevant services they can provide.

During WGEI Meeting I (July 4th-5th 2016), several experts and representatives of related e-infrastructures were invited. The following initiatives were presented:

European e-infrastructure landscape – ICT infrastructures serving scientific communities (Jesus Marco de Lucas, IFCA)

Jesus presented the European Grid Initiative (EGI.eu), EGI Engage (Horizon 2020 project), and the competence center approach for ESFRI projects like BBMRI-ERIC, ELIXIR, LifeWatch, DARIAH, etc. The INDIGO-DataCloud was presented as an open source Cloud platform for science that is currently being developed as part of an Horizon 2020 project. PRACE was presented as the European HPC Research Infrastructure. Several other public and commercial e-infrastructure solutions were discussed. Finally the European Open Science Cloud was mentioned as the next step on the horizon.

Marine biodiversity data systems and services – an EMODnet perspective (Simon Claus, VLIZ)

Simon presented the European Marine Observation and Data Network, and more specifically the Biology component of this network. Various activities under the EMODnet umbrella such as data management, facilitating data access and creation of data products were discussed. Background e-infrastructure components such as the European Ocean Biogeographic Information System (EurOBIS), the World Register of Marine Species (WoRMS), and Marine Regions were described.

Marine biodiversity data systems and services – a LifeWatch perspective (Christos Arvanitidis, HCMR)

Christos presented the LifeWatch ESFRI as an e-infrastructure for biodiversity and ecosystem research. The presentation contained objectives and ongoing developments both in the framework of the Marine Virtual Research Environment and at national level. The LifeWatch Taxonomic Backbone services were presented. The RvLab and other mobile portals were demonstrated as supporting e-infrastructure tools for biodiversity research.

European marine genomics data infrastructure – a MPI-MM perspective (Renzo Kottman, MPI)

Renzo presented the involvement of MPI in information infrastructures on a national and European level. The challenges and benefits of a scientific e-infrastructure were discussed. Renzo presented the approach and lessons learned from the development of bioinformatics solutions for the MicroB3 FP7 project and the Ocean Sampling Day initiative. Discussion was raised on the required e-infrastructure to support an Ocean Sampling Day consortium under the wings of EMRBC.

European marine genomics data infrastructure – an ELIXIR perspective (Petra Ten Hoopen, EBI)

Petra presented ELIXIR as a life sciences ESFRI connecting the EMBL European Bioinformatics Institute (EMBL-EBI) and national bioinformatics centers in Europe. ELIXIR-EXCELLERATE was described as a Horizon 2020 project to implement the ELIXIR Research Infrastructure. EMBL-EBI and its core resources were listed. Petra also described the MicroB3 reporting and services standards and the European Marine Biological Research Infrastructure cluster EMBRIC and their activities to connect resources.

The experts and representatives of related e-infrastructures were invited a second time to the WGEI Meeting IV (December 14th-15th 2016) to have a follow-up discussion and continue the consultation process. During this meeting, the invited experts provided input on useful examples of existing systems for each e-infrastructure component (see Annex 3).

After the consultation meetings, WGEI started the exercise to compile a list of **available components within the related e-infrastructures**.

B.4. Prioritization

The e-infrastructure requirements are dictated at different levels. There are requirements that are defined by the central tasks at EMBRC HQ level and the interaction with the EMBRC operators, the users, and other Research Infrastructures.

For each component of the EMBRC shopping list, the importance was prioritized at four community levels: (1) the EMBRC management (including the liaison officers), (2) the EMBRC operators (including the EMBRC nodes where the services reside), (3) the EMBRC users, and (4) other research infrastructures (Figure 3).

At each community level a **MOSCOW analysis** was performed and four priority categories were used: Must (priority score = 3), Should (priority score = 2), Could (priority score = 1) and Won't (priority score = 0) (Figure 3).

For each component of the EMBRC shopping list, a final priority score was calculated by adding up the individual scores at the four community levels (Annex 3). **Components with a final priority score ≥ 7 , were identified as high priority.**

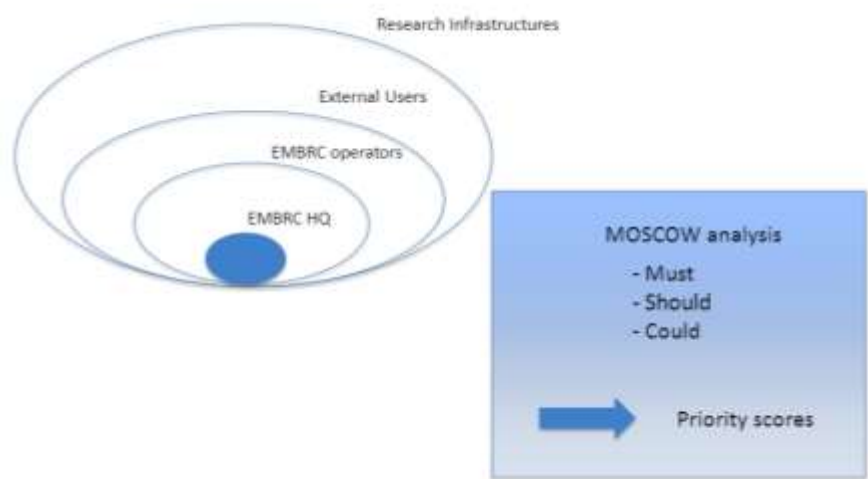


Figure 3 – MOSCOW analysis at four community levels

For each of the components, there was an **evaluation of the priority of development**, whether or not these components should be developed by EMBRC or could be co-developed or simply used from existing developments outside of EMBRC.

For each individual e-infrastructure component of the EMBRC shopping list, a **SWOT analysis** was performed at four levels:

- **Individual development:** development of components is initiated by EMBRC to address EMBRC objectives and requirements.
- **Co-development:** co-development of components is in collaboration with related initiatives aimed at realization of shared components that can address EMBRC objectives and requirements.
- **Integration:** incorporation of externally developed components as a dedicated system hosted and managed within EMBRC.
- **Interoperation:** making use of existing components managed and hosted outside of EMBRC by creating interoperability with these components.

For each component, the **preferred approach** and some **recommendations** were identified in consultation with the invited experts during Meeting IV (December 14th-15th 2016) (see Annex 3 and Chapter 5.A *Recommended EMBRC developments* below).

5. E-infrastructure architecture

A. Recommended EMBRC developments

Some important recommendations were already made as part of D3.1 - Report on e-infrastructure requirements and e-workflow scenarios, but remain valid:

- EMBRC should aid each partner to estimate its current and future storage and computing needs. Additionally, EMBRC should ensure that all partners have access to remote high performance computing facilities, either supplied by EMBRC operators or made available as a cloud service, for projects that cannot efficiently be supported locally, e.g. large phylogenetic analyses, aggregation and analysis of large, disparate ecological datasets, the digitization of historical and legacy datasets, (meta-)genome and (meta-)transcriptome assemblies and annotation, as well as imagery, HD video and microscopy outputs. A current key limitation is network bandwidth that, considering present needs, should reach a minimum of 1 Gbps for any EMBRC partner.
- EMBRC should support from central funds a team of data scientists who will support EMBRC users in creating e-workflows and maximizing the value of their data. Another task for this team will be to organize workshops, conferences, and meetings designed to promote the sharing of expertise among members. The size of this team, and how its members might be deployed, is unclear and requires further discussion.

In addition, here are some more specific recommendations for the required e-infrastructure components:

Administrative tools:

Since the administrative tools are specific to the central management of EMBRC, co-development is not an option for most of these components.

For the project management system, it is necessary to make a distinction between internal and external management, where external management means the collecting and presenting of information.

Document managing system: currently mainly Dropbox is used for this purpose. However, there might be issues related to versioning. Dropbox seems to be more useful as a repository and less as a collaborative platform. An alternative could be to use a Wiki system, which also offers version tracking, but has not the same level of control as a docs based system.

A main issue with user registries is the problem of central authentication. Developments to solve this require time and have been on-going for several years. Given the limited

time frame, the following strategy is recommended: on the short term to set up a system that fulfills the EMBRC requirements, and on the long term to make this system compatible and link with the existing systems for authentication.

Registers and catalogues:

Sample register: It is very important to register the origin of the sample, especially if it concerns resources where the Nagoya protocol applies. For these resources it is furthermore recommended to also consult legal databases, such as ABSCH. LIMS systems are good to track the samples but not for registering their origin. It would be good practices to work towards common metadata and use of persistent identifiers at the sample level. Furthermore, EMBRC should look into linking with the BioSamples database, which will become the central registry of samples for ELIXIR resources. BioSamples will provide a hierarchical structure of samples and will support recognized ontologies. BioSamples records will be linked to individual experiments (genomic, metagenomic, metabolomics, etc.).

The users of the EMBRC e-infrastructure should be stimulated to provide their publications, there should be guidelines for the users on how to refer to the EMBRC services, there could be a terms of agreement or the users could be incentivized through factor. The literature register should be used to track and assess the impact of EMBRC on science. The literature register should harvest the main publication platforms to see if the publications mentioning and acknowledging EMBRC are already in the EMBRC literature register.

The expert register needs to be focused towards specific EMBRC expertise and closely linked with the service register. Furthermore, a link with the publication register and with OrcID, LinkedIn, ResearchGate, etc. is recommended, as this can be seen as a validation of the expert. The expert register should both be at the level of persons and institutes. Some categories of expertise could be identified (e.g. scientific diving).

Analysis methods registers exist in certain specific communities; a general register is lacking however due to challenges such as high level management. Such a register is seen as valuable but challenging. It could be very innovative to have such register, but it would be very ambitious to set up. Such register would also need to refer to publications. It is recommended to not see this as a priority.

Datasets register: Both the EMODnet biology dataset and ENVRIplus catalogues are very general and are able to accommodate all datasets listed in the EMBRC inventory. A dedicated entry point from the EMBRC portal is needed. A limiting factor could be the wide range of standards EMBRC wants to be compatible with (INSPIRE, GCMD, SeaDataNet, CERIF, etc.).

Training register: Interoperation with marinetraining.eu is required, but a dedicated entry point is needed from the EMBRC portal.

Knowledge output module:

This module needs to be developed as a showroom for the EMBRC realizations. Statements on realized output need close links to literature, dataset and expert registers. EMBRC needs to identify who the specific users are, and translate the knowledge output to different levels. In ASSEMBLE Plus, EMBRC will work with AquaTT to identify who the specific users are and to translate the knowledge output to different levels.

Dataset and raw data file repositories:

Repositories are needed to keep track of the raw and processed data by the users of the EMBRC services. These repositories need to be certified and comply with specific requirements regarding long term preservation. One suggestion is for EMBRC to have a dedicated repository for data which falls out of already existing repositories or could be used as a repository for continued storage for the Joint Research Activities (JRAs) and the access projects.

EMBRC also acknowledges the importance of long term preservation of research data in trustworthy accredited repositories. Therefore EMBRC should maintain a repository for ensuring archival and long term preservation of newly generated data, in collaboration with recognized data repositories such as EurOBIS, EMODnet, EBI-EMBL, Pangaea, GEOSS and Copernicus.

Integrated thematic databases:

Very strong existing initiatives and RI's have operational components and few specific requirements are expected from EMBRC. Therefore interoperation is recommended as the correct approach.

However, there are specific domains where additional systems could be of use, for example: there is still a lot of information lacking for many model species. Therefore it would also be interesting to look at co-operation for the Reference molecular data database.

Analysis tools:

The Virtual analysis platform is a development scheduled within ASSEMBLE Plus, based on developments at HCMR and VLIZ, and building on already existing initiatives with operational components.

This Virtual analysis platform should include or link closely with the Sequence data processing tools (such as e.g. ABiMS, ELIXIR Registry of tools and services (still in progress), EMBL-EBI Embassy cloud, etc.), which are identified as another required infrastructure component in the EMBRC e-infrastructure shopping list.

Local databases:

EMBRC needs to realize and improve virtual access to the local data. Part of this work implies an optimized local database infrastructure. EMBRC could either provide guidelines and training on how to set up relational databases and improve connection with integrated thematic data systems.

Data storage and computing capacity:

Data storage and computing capacity are in general issues to be dealt with by the EMBRC nodes. However, EMBRC should map the needs and look into co-operation with European initiatives to empower the local and central services.

EMBRC should take action to ensure sufficient storage capacity. This can be done in cooperation with European e-infrastructures, such as EGI, EUDAT, INDIGO. Expertise in this domain could also be developed in collaboration with Euro-BioImaging.

Networking and connectivity:

Networking and connectivity are in general issues to be dealt with by the EMBRC nodes, except for some central services that might be developed. Locally, each user needs only one single stop-over point, which support the idea of the EMBRC virtual access system.

EMBRC should look into co-operation with European initiatives such as e.g. the E-Infrastructure Commons, European Open Science Cloud, INDIGO-DataCloud, etc., to make use of offered cloud infrastructure. Further recommendations are to set up a pilot to connect to this cloud (e.g. shared Roscoff bioinformatics pipelines on INDIGO HPC Cloud).

Training:

The European Marine Training Portal should be further developed. It should centralize the training offer and work towards an online training platform. Resources should be made available for EMBRC operators to involve in training activities.

B. Architecture model

Based on the combination and interaction of the identified e-infrastructure components, a general architecture model was drawn up and put forward. This model is based on the guiding principle – already put forward in pp1EMBRC – that the general strategy approach for EMBRC regarding e-infrastructure is to work towards interoperability with existing thematic initiatives and to pursue co-development regarding lacking components.

The proposed architecture recommends the implementation of a number of components at the central level that can support a more integrative approach regarding the processes at the local level and the interaction between the local operators and thematic initiatives.

The e-infrastructure architecture scheme in Figure 4 suggests a two-way interaction between EMBRC and related e-infrastructures such as LifeWatch, ELIXIR, EMODnet, and

Pangaea. The EMBRC nodes receive the necessary training to comply with standards and make use of European e-infrastructures and their components. These e-infrastructures should take into account the specific requirements of the EMBRC community and provide visibility to EMBRC output.

At the same time EMBRC needs to be able to provide direct user access to its knowledge output. This can be realized by setting up a virtual access system that has a number of central services and has a register that keeps track (making use of persistent identifiers) of resources that have been contributed to external systems in the framework of EMBRC. It should be possible to keep track of the provenance and the use of these resources. In this model the virtual access system is part of the EMBRC portal. Added value products could be created in the framework of EMBRC as part of the knowledge output.

Networking, computing and storage are transversal components that need to be shaped in co-operation with European initiatives to empower the local and central services.

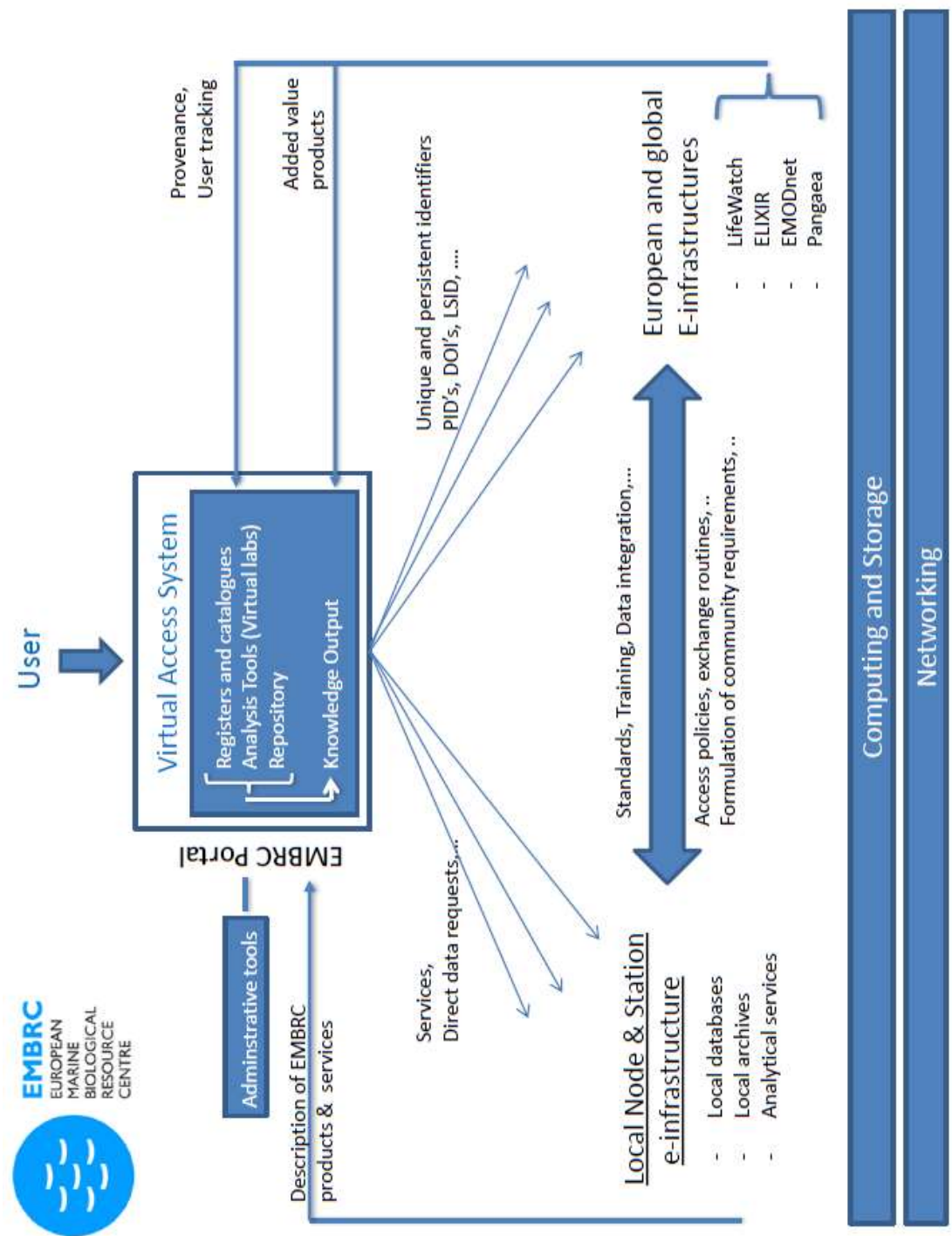


Figure 4 – Suggested e-infrastructure architecture scheme for EMBRC

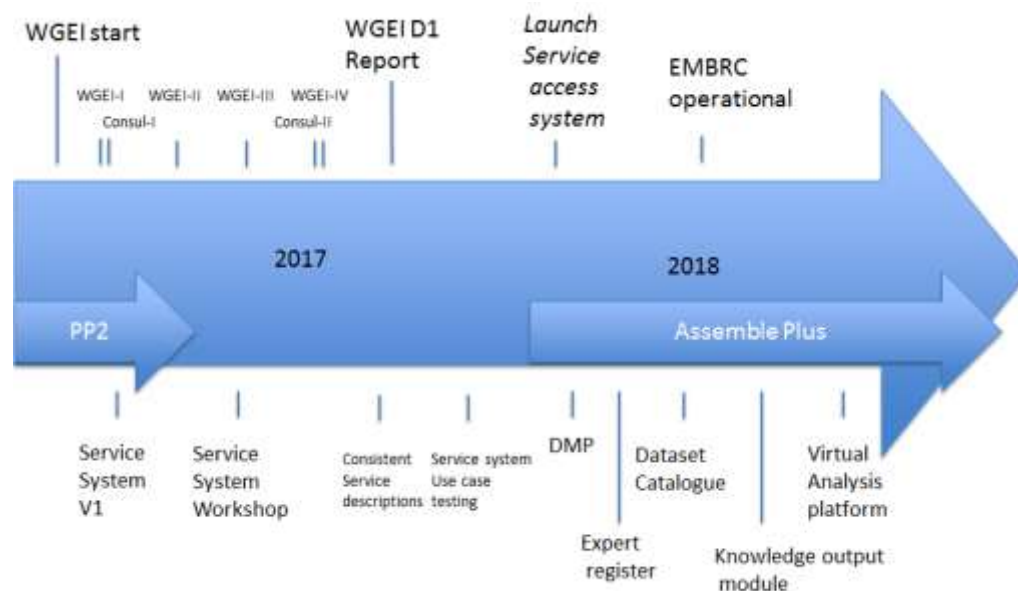
6. Short term implementation plan and long term vision

A. Short term implementation

The short term implementation plan needs to focus on those components of the architecture model that are considered high priority and were identified as such in Chapter B.4. Prioritization (see also Annex 3).

Some of these **top priority developments** have already started during pp2EMBRC, e.g. the EMBRC portal (www.embrc.eu), the service discovery and access request system and the European Marine Training Portal (www.marinetraining.eu). The service discovery and access request system is being integrated with the www.embrc.eu portal and will be further upgraded within the framework of ASSEMBLE Plus.

Other priority developments such as a Literature register, Expert register, Datasets register, Knowledge output module, Dataset and raw data file repository and Virtual analysis platform are planned in the framework of the ASSEMBLE Plus Horizon 2020 project. The integration of an Event management system is ongoing. See also Annex 2 for the specific e-infrastructure requirements for the planned developments within ASSEMBLE Plus.



The **remaining components** that were considered high priority but have not been planned yet are a Service allocation system (Booking system), an User registry, Local and

shared data storage and computing capacity, and Local and central networking and connectivity (see Annex 3). Other high priority components include IT staff, Liaison officers and Trainers.

Cost assessment of the remaining components were considered extremely difficult. Especially since the functional requirements of these components have not been studied in detail. E.g. for the administrative tools, factors such as edition (server-based versus cloud-based), payment plan (free tools, monthly versus yearly costs or one-time payment), and the amount of intended users need to be taken into account, as these can severely alter the pricing. Also additional costs for training, deployment and maintenance should be considered. A very rough cost estimate for the administrative tools ranges between very low costs to several tens and even hundreds of thousands. More follow-up discussions are needed to assess the additional components to be developed at the central level regarding registers and catalogues, integrated thematic databases and analysis tools. Networking and storage are considered a more local issue until the central services are fully developed.

Important efforts and actions are planned or ongoing to advance the EMBRC e-infrastructure on the short term. Positive dialogues have been initiated with related e-infrastructures. Mutual opportunities should be identified and formalized in formal agreements or Service Level Agreements. The European e-infrastructure landscape is rapidly evolving and EMBRC should be on the look out to take advantage of all arising opportunities that can fulfil the EMBRC community requirements and can support the EMBRC e-infrastructure offer. A continuation of the WGEI as platform for e-infrastructure related discussions and positioning is seen as essential.

B. Long-term vision

The short term plan will satisfy the more immediate needs that were identified by WGEI. At the same time it will improve the articulation of EMBRC e-infrastructure demands and resources, which is crucial for the development of a more long term vision. Very likely an operational EMBRC will uncover gaps and bring forward additional requirements in terms of e-infrastructure.

At the same time new opportunities and possible solutions will arise. The landscape of e-infrastructures at European level is rapidly evolving. Cluster projects like EMBRIC, CORBEL and ENVRiplus are working towards defining overarching needs and solutions in collaboration with the participating RI's, however they are, by definition 'generic'. Therefore it is key to interact closely with these initiatives and to make best use of the developments and standards that result from that work.

For a lot of the identified e-infrastructure resources required at the local level and central level, the EMBRC strategy should be to make optimal use of European e-infrastructure developments such as the E-infrastructure Commons and the European Open Science Cloud.

GLOSSARY OF ACRONYMS

ABiMS	Analysis and Bioinformatics for Marine Science (http://abims.sb-roscoff.fr/)
ABSCH	Access and Benefit-Sharing Clearing-House (https://absch.cbd.int/)
ASSEMBLE Plus	Association of European Marine Biological Laboratories Expanded. ASSEMBLE Plus will provide scientists from academia, industry and policy with a quality-assured program of Transnational Access (TA) and Virtual Access (VA) to marine biological stations offering a wide variety of marine ecosystems, unique marine biological resources, state-of-the-art experimental and analytical facilities with integrated workflows, historical observation data, and advanced training opportunities. The goal is to stimulate European excellence in fundamental and applied research in marine biology and ecology, thereby improving our knowledge- and technology-base for the blue economy, policy and education purposes. The sum of the actions envisaged in ASSEMBLE Plus will ultimately increase the number of users of marine biological stations and shape a novel business strategy perspective, to be based on effective integration and efficient complementarities, resulting in a key contribution to their long-term sustainability.
BBMRI-ERIC	Biobanking and BioMolecular Resources Research Infrastructure – European Research Infrastructure Consortium (http://www.bbmri-eric.eu/)
CERIF	The Common European Research Information Format (http://www.eurocris.org/cerif/main-features-cerif)
CORBEL	Coordinated Research Infrastructures Building Enduring Life-science Services (http://www.corbel-

	project.eu/)
DARIAH	Digital Research Infrastructure for the Arts and Humanities (http://www.dariah.eu/)
EGI	European Grid Initiative (https://www.egi.eu/)
EGI Engage	Engaging the EGI Community towards an Open Science Commons (https://wiki.egi.eu/wiki/EGI-Engage:Main_Page)
E-Infrastructure Commons	The e-infrastructure Reflection Group uses the metaphor of the e-Infrastructure Commons for the e-Infrastructure resources and related services, which among others refer to networking, computing, storage, data and software, along with digital tools and collaboration opportunities. (http://knowledgebase.e-irg.eu/commons)
ELIXIR	A distributed infrastructure for life-science information (https://www.elixir-europe.org/)
ELIXIR-EXCELLERATE	An EC funded project to help ELIXIR coordinate and extend national and international data resources to ensure the delivery of world-leading life-science data services.
EMBRC	European Marine Biological Resource Centre (http://embrc.eu/)
EMBRC HQ	Headquarters of EMBRC, located at UPMC in Paris.
EMBRIC	European Marine Biological Research Infrastructure Cluster (http://www.embric.eu/)
EMODnet Biology	Biology Project of the European Marine Observation and Data Network (http://www.emodnet-biology.eu/)
ENVRIplus	A Horizon 2020 project bringing together Environmental and Earth System Research Infrastructures, projects and networks together with

technical specialist partners to create a more coherent, interdisciplinary and interoperable cluster of Environmental Research Infrastructures across Europe (<http://www.envriplus.eu/>)

EOSC	European Open Science Cloud
ESFRI	European Strategy Forum on Research Infrastructures (http://www.esfri.eu/)
EurOBIS	European Ocean Biogeographic Information System (http://www.eurobis.org/)
FIMS	Field Information Management System
GCMD	Global Change Master Directory (https://gcmd.nasa.gov/)
Horizon 2020	The European funding program for research and innovation in Europe
INDIGO-DataCloud	INDIGO - DataCloud develops an open source data and computing platform targeted at scientific communities, deployable on multiple hardware and provisioned over hybrid, private or public, e-infrastructures. By filling existing gaps in PaaS and SaaS levels, INDIGO-DataCloud will help developers, resources providers, e-infrastructures and scientific communities to overcome current challenges in the Cloud computing, storage and network areas. (https://www.indigo-datacloud.eu/)
INSPIRE	This European Spatial Data Infrastructure will enable the sharing of environmental spatial information among public sector organizations, facilitate public access to spatial information across Europe and assist in policy-making across boundaries. (http://inspire.ec.europa.eu/)
JRA's	Joint Research Activities within ASSEMBLE Plus.

LifeWatch	A virtual laboratory for biodiversity research (http://www.lifewatch.eu/)
LIMS	Laboratory Information Management System
MicroB3	Marine Microbial Biodiversity, Bioinformatics, Biotechnology, EU FP7 project (http://www.microb3.eu/)
OSD	Ocean Sampling Day (http://www.microb3.eu/osd)
pp1EMBRC	First preparatory phase for EMBRC
pp2EMBRC	Second preparatory phase for EMBRC
PRACE	Partnership for advanced computing in Europe (http://www.prace-ri.eu/)
RI	Research Infrastructure
SeaDataNet	a pan-European infrastructure to ease the access to marine data measured by the countries bordering the European seas (https://www.seadatanet.org/)
WGEI	EMBRC Working Group on E-Infrastructures
WoRMS	World Register of Marine Species (http://www.marinespecies.org/)

ANNEX 1: Specific e-infrastructure requirements for each use case

pp1EMBRC USE CASE 1: Sequencing genomes and/or transcriptomes using second and third generation sequencing technologies, with *Platynereis dumerilii* as an example. In this scenario it is assumed virtually nothing is previously known about the genome/transcriptome in question.

- Database comparing the currently available sequencing methods.
- Database listing available sample material.
- Database listing available gene libraries.
- Database listing available commercial laboratories for sequencing.
- LIMS and FIMS software to track the samples and material.
- Storage for the raw sequence data files.
- Software packages for quality checks, trimming, cleaning, error correction.
- Assembly algorithms and software.
- Software packages for validation of assemblies.
- Software packages for annotation.
- Software packages for visualization of assemblies.
- Public databases to deposit the sequence data and assembly.
- Human resources.

pp1EMBRC USE CASE 2: Developing new model organisms, with *Clytia hemisphaerica*, *Phallusia mammillata* and *Patiria minata* as examples.

- Database listing experts who can consult on model organisms.
- Database listing available culture organisms.
- Database listing available culture facilities.
- Facilities for high-throughput molecular work.
- Storage facilities.
- Human resources: expertise in clean-up, assembly and analysis of the data.

pp1EMBRC USE CASE 3: Species distribution modelling based on collection data and molecular data, adapted to high latitude pelagic habitats.

- Databases with taxon observation records.
- Databases with ecological/environmental data.
- Methods for enabling integration of molecular and traditional biodiversity data.
- Widely accepted and used metadata deposition standards.
- Workflow to harvest the observation data.
- Workflow to exploit molecular data from reference database and sequence database.
- Local database to store the harvested observation data and exploited molecular data.
- Software for experiment designing and model building.
- Experiment catalog and model catalog.
- Software to prepare and map the data.
- Servers.
- Human resources.

pp1EMBRC USE CASE 4: Using historical datasets to augment 'omics data to understand ecological change over time.

- Searchable data system listing the available historical samples and (meta)data.
- Searchable data system listing the available historical samples & (meta)data.
- Software to digitize metadata of the samples.
- Database listing available gene libraries.
- Database listing available commercial laboratories for sequencing.
- LIMS and FIMS software packages to track the samples and material.
- Storage for the raw sequence data files.
- Software packages for quality checks, trimming, cleaning and error correction of the sequence data.
- Assembly algorithms and software.
- Software packages for validation of assemblies.
- Software packages for annotation.
- Software packages for visualization of assemblies.

- Public databases to deposit sequence, assembly and annotation data.
- Databases with environmental data for data integration.
- Human resources.

Ocean Sampling Day (source: MicroB3/OSD)

- Database listing (My)OSD participants and (My)OSD participant sites, and which data from previous (My-OSD editions is available where.
- Online system to sign up and order sampling kits.
- Workflow for sending the (My)OSD samples to the sequencing facilities.
- Bio-repository.
- Data-repositories for sequencing data and environment (meta)data.
- Reference gene catalogues and libraries for sequencing.
- Human resources.

Service access system (source: pp2EMBRC)

- Access services database containing all research services provided by the partners to the EMBRC network.
- Central web-based service-access system where end-users can search the available services.
- Back-end of the service-access system will allow communication with EMBRC national liaison officers and EMBRC HQ.
- Human resources: EMBRC partners keep their services up to date. Technically skilled people to keep the access system up and running.

Improving virtual access to marine biological stations data, information and knowledge (source: ASSEMBLE Plus)

- Knowledge Transfer Platform - Knowledge Outputs module.
- Knowledge Transfer Platform - Publications module.
- Virtual open access entry point to data resources.
- Data access and standardization of genomic and long term marine biodiversity observation.
- Virtual analysis platform for long-term biodiversity and genomics observatories data.

Configurator (connecting several databases: nucleotide data – proteomic data – microbial data – chemical data) (source: EMBRIC WP4)

- Data storage facilities (servers)
- Nucleotide database (WP4 will use already existing database)
- Proteomic database (WP4 will use already existing database)
- Microbial database (WP4 will use already existing database)
- Chemical database (will be built by WP4 with the input from WP6 and WP7)

Microbial pipeline from environments to active compounds (source: EMBRIC WP6)

- Database listing different chemical compounds at different stages of isolation.

High performance microalgae for blue technological applications (source: EMBRIC WP7)

- Database listing different chemical compounds at different stages of isolation.

Ecosystem assessment and mapping (MSFD-related) use cases

- Databases with taxon observation records.
- Databases with ecological/environmental data.
- Workflow tools to calculate and map indicators and descriptors.

Digital library – Retrieving or providing overview of EMBRC related publications

- Online system for retrieving or providing overview of EMBRC related publications.

Management – Central management of EMBRC

- Financial administration system.
- Project management system.
- Document management system.
- Event registry system.

Long-term monitoring

- LIMS and FIMS software to track the samples and material throughout the process.
- Local database system for efficient storage and exchange of long term data series.
- Reporting and visualization workflows.

Education and training – Building capacity and applying standards within the EMBRC community.

- Register of relevant trainings.
- Webinar/Online training platforms.
- Hardware and networking equipped training rooms for physical training.

ANNEX 2: Specific e-infrastructure requirements for the planned developments within ASSEMBLE Plus

Set up a single-access point to the offered infrastructure.

Task NA1.2 within NA1: Improving Access Provision

- EMBRC service access system.
- Access services database containing all research services provided by the partners to the ASSEMBLE Plus network.

Design DMP and interoperability with related e-infrastructures

Task NA2.1 within NA2: Improving virtual access to marine biological stations data, information and knowledge.

- Analysis of the work field.
- Overlap, gaps, links with other initiatives.
- Standards and protocols.

Creation of the Knowledge Transfer Platform

Task NA2.2 within NA2: Improving virtual access to marine biological stations data, information and knowledge.

- Information database containing info on researchers, their research, expertise, publications (Integrated Marine Information System).
- User search interface where end-users can search through the information database (Knowledge Output Module, Publication Module).

Set-up of a virtual open access entry point to data resources

Task NA2.3 within NA2: Improving virtual access to marine biological stations data, information and knowledge.

- Information database containing metadata on datasets (Integrated Marine Information System).
- Data repository for archival purposes.
- Data infrastructure components specific for genomic observatories derived from previous projects.
- Workflow to harvest and quality control data.

- Portal for data, information on community data policies and standards, experimental protocols and the outcomes of benchmarking exercises (VA platform).
- Monitoring tools for system access, data download, etc.

Improvement of data access and standardization of genomic and long term marine biodiversity

Task NA2.4 within NA2: Improving virtual access to marine biological stations data, information and knowledge.

- Workflow for harvesting, standardizing, annotating of biodiversity and genomic data.
- Data repositories for biodiversity and genomic data.

Set-up of a virtual data platform for data analysis

Task NA2.5 within NA2: Improving virtual access to marine biological stations data, information and knowledge.

- Virtual analysis platform (R, Taverna) that enables user to select data and perform predefined analyses.
- Databases feeding the platform.
- Predefined Workflows: biodiversity analysis workflows, bioinformatics pipelines, etc.
- Monitoring tools for system access, data download, data use.

ANNEX 3: List of required e-infrastructure components (“EMBRC shopping list”)

			MOSCOW analysis at 4 community levels						SWOT analysis
Required component	Definition	EMBRC Management	EMBRC Operators	EMBRC users	Other RIs	Final priority score	Implementation status (L0-L3)	Level of SWOT analysis	
ADMINISTRATIVE TOOLS	Service request system	Online system that allows a user to request EMBRC services and that allows a nodes and HQ to approve or reject requests.	Must (3)	Must (3)	Must (3)		9	L3: Implemented (pp2EMBRC)	Individual development Co-Development Integration Interoperation
	Service allocation system (Booking system)	System that allows a node or station to plan, schedule and allocate services and resources.	Could (1)	Must (3)	Must (3)		7	L0: To be planned	Individual development Co-Development Integration Interoperation
	Financial administration system	Software for bookkeeping and financial administration					4	L0: To be planned	Individual development Co-Development

		EMBRC HQ.							Integration
									Interoperat n
	Project manage ment system	Software for project management, task descriptions, and follow up of activities.	Mus t (3)	Coul d (1)	Coul d (1)		5	L0: To be planned	Individual developme Co- Developme Integration Interoperat n
	Docume nt manage ment system	Software to store and share documents.	Mus t (3)	Coul d (1)	Coul d (1)		5	L0: To be planned	Individual developme Co- Developme Integration Interoperat n

58

[illegible]

REGISTERS AND CATALOGUES	Central service register	Central catalogue of services offered by EMBRC.	Must (3)	Should (2)	Must (3)		8	L4: Implemented (pp2EMBR C)	Individual development Co-Development Integration Interoperation
	Sample register	Register of taken and stored samples (cfr. Nagoya).	Should (2)	Must (3)	Could (1)		6	L0: To be planned	Individual development Co-Development Integration Interoperation
	Literature register	Register of publications produced in the framework or based on the service provision of EMBRC.	Must (3)	Should (2)	Could (1)		6	L1: Planned (ASSEMBLE Plus)	Individual development Co-Development Integration Interoperation

61

62

TO									
	Training register	Register of trainings organised in the framework of EMBRC or linked to its service provision.	Mus t (3)	Shou ld (2)	Mus t (3)		8	L0: To be planned	Individual developme Co- Developme Integration Interopera n
	Knowledge output module	Online system displaying knowledge output of EMBRC activities.	Mus t (3)	Coul d (1)	Mus t (3)		7	L1: Planned (ASSEMBL E Plus)	Individual developme Co- Developme Integration Interopera n
TO	Dataset and raw	Repository storing data	Mus t (3)	Mus t (3)	Mus t (3)		9	L1: Planned	Individual developme

	data file repositor ies	files and datasets produced or made available in the framework of EMBRC or making use of its service provision.						(ASSEMBL E Plus)	Co- Developme Integration	Interopera n
INTEGRATED THEMATIC DATABASES	Sequenc e data database	Integrated thematic database that allows users to access sequence data collected in the framework of EMBRC or its service provision.	Coul d (1)	Must (3)	Coul d (1)		5	L0: To be planned	Individual developme Co- Developme Integration	Interoperat n
	Referenc e molecula r data database	Integrated thematic database that allows users to access reference molecular data collected in the framework of	Coul d (1)	Must (3)	Coul d (1)		5	L0: To be planned	Individual developme Co- Developme Integration	Interoperat

		EMBRC or its service provision.							n
	Taxon observation data database	Integrated thematic database that allows users to access taxon observation data collected in the framework of EMBRC or its service provision.	Could (1)	Must (3)	Could (1)		5	L0: To be planned	Individual development Co-Development Integration Interoperation
	Ecological and environmental data database	Integrated thematic database that allows users to access ecological and environmental data collected in the framework of EMBRC or its service provision.	Could (1)	Must (3)	Could (1)		5	L0: To be planned	Individual development Co-Development Integration Interoperation

ANALYSIS TOOLS	Sequence data processing tools	Tools that allow users to process sequence data (qc, cleaning, assemble, annotate, etc.).	Could (1)	Must (3)	Could (1)		5	L0: To be planned	Individual development Co-Development Integration Interoperation
	Virtual analysis platform	Online environment that allows a user to run predefined calculations, mapping and analysis workflows on a combination of personal and shared data resources.	Should (2)	Should (2)	Could (1)		5	L1: Planned (ASSEMBLE Plus)	Individual development Co-Development Integration Interoperation
HUMAN RESOURCES	Bioinformaticians	Human resources that have expertise in bioinformatics and can support EMBRC activities.	Could (1)	Must (3)	Could (1)		5	L0: To be planned	Individual development Co-Development Integration Interoperation
	Data scientists	Human resources that have expertise in data science and can support EMBRC activities.	Could (1)	Must (3)	Could (1)		5	L0: To be planned	Individual development Co-Development Integration Interoperation
	IT staff	Human	Mus	Mus	Coul		7	L0: To be	Individual

		resources that have expertise in ICT and can support EMBRC activities.	t (3)	t (3)	d (1)			planned	development Co-Development Integration Interoperability
	Liaison officers	Human resources that are appointed as liaison officers to liaise with EMBRC and stations.	Mus t (3)	Mus t (3)	Coul d (1)		7	L1: Planned	Individual development Co-Development Integration Interoperability

LOCAL DATABASES	Local monitoring databases	Local database systems for efficient storage and exchange of long term data series.	Should (2)	Must (3)	Could (1)		6	L0: To be planned	Individual development Co-Development Integration Interoperation
	Lab or Field Information System	LIMS and FIMS software to track the samples and material throughout the process.	Should (2)	Must (3)	Could (1)		6	L0: To be planned	Individual development Co-Development Integration Interoperation
DATA STORAGE AND COMPUTING CAPACITY	Local	Appropriate institutional storage and computing capacity at the nodes and stations.	Must (3)	Must (3)	Could (1)		7	L0: To be planned	Individual development Co-Development Integration Interoperation
	Shared	Appropriate shared storage and computing capacity.	Must (3)	Must (3)	Could (1)		7	L0: To be planned	Individual development Co-Development Integration Interoperation
AND	Local	Appropriate institutional	Must (3)	Must (3)	Could (1)		7	L0: To be planned	Individual development

TRAINING		networking and connectivity at the nodes and stations.							Co-Development Integration Interoperability
	Central	Appropriate shared storage and networking and connectivity.	Must (3)	Must (3)	Could (1)		7	L0: To be planned	Individual development Co-Development Integration Interoperability
	Online training platform	Webinar/Online training platform.	Should (2)	Could (1)	Could (1)		4	L0: To be planned	Individual development Co-Development Integration Interoperability

								n
Trainers	Human resources that are available to provide training supporting EMBRC activities.	Must (3)	Must (3)	Could (1)		7	L0: To be planned	Individual development Co-Development Integration Interoperation