EMBRC
EUROPEAN
MARINE
BIOLOGICAL
RESOURCE
CENTRE

Plan for the requirements of the EMBRC e-infrastructure

January
2014

Deliverable D3.3

The EMBRC preparatory-phase

Contributers:

Chuck Cook/ccook@ebi.ac.uk
Janet Chenevert/chenevert@obs-vlfr.fr
Cymon Cox/cymon.cox@gmail.com
Philippe Dru/dru@obs-vlfr.fr
Claire Gachon/claire.gachon@sams.ac.uk
Evelyn Houliston/evelyn.houliston@obs-vlfr.fr
Tomas Larsson/tomas.larsson@embl.de
Peter Lenart/lenart@embl.de
Nick Pade/nipa@mba.ac.uk
Remo Sanges/remo.sanges@szn.it

SEVENTH FRAMEWORK PROGRAMME

D3.3 – PLAN FOR THE REQUIREMENTS OF THE EMBRC E-INFRASTRUCTURE

Q1 – Is the present report addressing the objectives described in the DoW?

Yes

Q2 – Are there any deviations to the DoW? If so, why?

No

Summary (internal for the MGT team)

See executive summary below.

Contents

EXECUTIVE SUMMARY 2

1. OBJECTIVE 2

2. BACKGROUND 3

3. E-INFRASTRUCTURE DEFINITION 3

4. SUMMARY OF PREVIOUS WP3 ACTIVITIES 4

5. GUIDING PRINCIPLES FOR THE EMBRC E-INFRASTRUCTURE 5

6. BLUEPRINT FOR THE EMBRC E-INFRASTRUCTURE 6

6.1 MODEL FOR AN E-INFRASTRUCTURE 6

6.2 FUNCTION OF THE E-INFRASTRUCTURE 7

6.3 AN EMBRC BIOINFORMATICS FACILITY: WHEN AND WHERE? 8

7. E-INFRASTRUCTURE: PEOPLE 8

7.1 DATA MANAGEMENT AND BIOINFORMATICS FUNCTIONS 9

7.1.1 FIMS/LIMS SUPPORT 9

7.1.2 GENOMICS AND SEQUENCING SUPPORT 9

7.1.3 ENABLING DATA ACCESS 10

7.1.4 TRAINING 11

8. HARDWARE REQUIREMENTS 11

9. PUTTING FLEXIBILITY AT THE HEART OF THE EMBRC E-INFRASTRUCTURE 12

10. INTEGRATION WITH OTHER INFRASTRUCTURES AND PROJECTS 12

Executive Summary

This report collates the results of two workshops in Heidelberg, a survey of EMBRC partners on current e-infrastructure and projected needs, and the conclusions drawn from the previous two deliverables (D3.1 “Report on e-infrastructure requirements and e-workflow scenarios” and D3.2 “Detailed evaluation of potential e-workflow scenarios”).

The goal of the EMBRC infrastructure will be to provide storage and analysis facilities for high-throughput marine biological data, e.g. molecular data generated by many research projects. The guiding principle for EMBRC’s e-infrastructure should be to identify, and join up with locally available resources to fill in the service and hardware gaps that fall between EMBRC partners and national and supra-national resources.

This EMBRC e-infrastructure should have four components:

1) A minimum level of network access for each partner.

2) Computational hardware and software for data storage and analysis.

3) Appropriate and up-to-date software for data storage and analysis.

4) Bioinformaticians and data scientists who will train or aid users to design experiments, analyse results, implement best-practice data management policies and transfer knowledge to academic, government, and private sector users.

To accomplish this, the EMBRC should establish a bioinformatics facility with a core team of bioinformaticians and data scientists, who will be tasked with providing services to all EMBRC users. This core team should be established by a few EMBRC partners, with one lead partner, and make use of existing, complementary expertise or hardware available at several nodes. We suggest that the structure of this facility may be determined through a bidding process. However, this process should be managed carefully by the EMBRC director and board, and modified if necessary, in consideration of possible funding opportunities.

In relation to the e-infrastructure EMBRC should also adopt and efficiently disseminate best practice data management. The policy will be based on the requirements for data management plans from funding agencies and on comparison to the policies of other RIs and in consideration of the Research Infrastructures’ data management charter that has been drafted by BioMedBridges.

1. Objective

The objectives of this deliverable is to “propose a plan to meet the requirements of the EMBRC e-infrastructure and define a set of standard work-flows and hardware and software benchmarks (with minimum and optimal requirements for EMBRC users) for physical infrastructure, data integration and processing pipelines as well as data storage and user access.”

2. Background

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 (D2.5). In summary, EMBRC has six aims:

1) Become the major European provider of marine biological resources, research infrastructure, and related services.

2) Foster mobility of European researchers between member states and between industry and academia.

3) Provide co-ordination and leadership of marine research through a strong management structure and through a pan-European programme of knowledge exchange, joint development activities, and meetings.

4) Ensure standardisation of data collection, storage and transmission, workflows and skills training along with implementation of best practise in research methodologies and technologies.

5) Develop a knowledge transfer platform that will facilitate innovation. EMBRC will help to increase economic well being and a higher quality of life through job creation and by bringing research discoveries to applied services and products.

6) Provide high quality infrastructure for training and education.

Fulfilment of these goals requires that EMBRC partner institutions have the infrastructure and expertise to undertake state-of-the-art research. Many aspects of this infrastructure encompass the generation, storage, and analysis of high volume—primarily molecular—data, for instance nucleotide sequences, whole genomes, proteomes, metabolomes, and images.

This report defines and describes these aspects of the EMBRC electronic infrastructure—the e-infrastructure.

3. E-infrastructure definition

In D3.1 we defined e-infrastructure as “network access, storage capacity, computational resources, software, and human resources” and stated that the goal of EMBRC’s e-infrastructure “should be 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, analyse and make publicly available the large volumes of data that are generated at EMBRC marine stations and laboratories.”

4. Summary of previous WP3 activities

To date work package 3 has surveyed EMBRC partners' e-infrastructure needs through a questionnaire, a small workshop with 16 EMBRC participants, and a larger international workshop with 50 participants. A summary of the first workshop, "EMBRC_Jun11_WP3_workshop _summary.pdf" is on the EMBRC website, and the results of the second workshop are presented in deliverables D3.1 "Report on e-infrastructure requirements and e-workflow scenarios" and D3.2 "Detailed evaluation of potential e-workflow scenarios", also available from embrc.eu.

The primary findings from the questionnaire were that all EMBRC partners are engaged in a diverse array of high throughput projects, that all anticipate future increases in similar data-intensive projects, and that data flows will increase significantly—probably exponentially—as technologies continue to improve. Many partners are already experiencing difficulties storing and/or analysing their research data, and even those who are currently able to manage these data anticipate future problems. An e-infrastructure that aids partners in the processing, storage, and analysis of high throughput data would greatly increase EMBRC partners' research capacity. Additionally, there is a notable shortage of funds to support bioinformaticians and data scientists at most partner marine stations. The questionnaire also showed that EMBRC partners hold a wide range of biological and historical collections that are for the most part not well catalogued. Thus, there is wide scope for digitization and then dissemination of these historically and scientifically important materials, and such work requires much the same hardware for data storage and analysis, as well as expertise in database management and the handling of large data sets, that molecular work requires.

In D3.2 we described four workflow scenarios, or use cases, that sampled the broad range of activities undertaken at EMBRC partners, and also served to highlight the e-infrastructure needs of marine researchers. These four workflows describe sequencing a genome, developing new model organisms, modelling marine species biodiversity, and combining historical data sets with 'omics data to understand ecological change. As with D3.1 these workflows emphasized the need for a robust and sustainable e-infrastructure that will allow EMBRC users to move, store, and analyse large volumes of scientific data. D3.2 also emphasised the need to design a flexible infrastructure to accompany fast-evolving conceptual needs driven by ongoing technological mutations.

As described in D3.1 new 'omics technologies have revolutionized all fields of biology. Related technologies that allow storage and analysis of very large imaging, oceanographic, and climatic datasets provide similar new opportunities for oceanographers and earth scientists, and will become increasingly critical to contextualise biological knowledge in the marine environment. For marine biologists the combination of these technological breakthroughs brings many new opportunities for analysis and understanding of biodiversity, developmental biology, marine resources, and ecology. Novel applications, including some that were not possible even when the ppEMBRC survey was conducted early in 2011, are continuously emerging, and promise to create entirely new disciplines. For example, spatially and temporally resolved genomics (e.g. along sediment cores) opens the door to deconvoluting evolutionary processes from geographical and physicochemical environmental variations *in situ*. **EMBRC's scientific edge will rely on the capacity to bridge the technological gap between these novel scientific ideas, to integrate computationally demanding biological data with**

heterogenous metadata, and to integrate of data management standards and policies across biology and marine sciences.

5. Guiding principles for the EMBRC e-infrastructure

The EMBRC e-infrastructure should ensure that EMBRC users have access to the storage, computational resources, and advice needed to undertake their research projects.

This does not, though, mean that EMBRC will supply all components of the e-infrastructure. All EMBRC partners have some computational infrastructure and some have in-house staff who contribute to the e-infrastructure. Additionally, many EMBRC host nations have e-infrastructure facilities that are available to all researchers, including those within EMBRC. These include, for instance, ELIXIR nodes and other national bioinformatics resources, such as BILS (the Bioinformatics Infrastructure for Life Sciences) and SciLifeLab-WABI in Sweden.

The guiding principle for EMBRC's e-infrastructure should be to identify, link and fill in the service and hardware gaps that fall between what is locally available at EMBRC partners and national and supra-national resources.

This task is not straightforward: EMBRC partners and their host countries vary widely in their installed e-infrastructure capacities and staffing. Some EMBRC partners have larger e-infrastructure deficits than others, and some EMBRC partner countries already provide a national e-infrastructure while others have much more limited shared resources: the EMBRC e-infrastructure will have to serve these varying needs. Furthermore, the EMBRC e-infrastructure will have to adapt to increasing demands for hardware and expertise as new technologies are developed and incorporated into research workflows. In sum, the goal of the EMBRC e-infrastructure should be to enable users at all partners to provide data handling and analysis support that is adequate for them to undertake research, but without duplicating what is available locally or from higher-level providers.

In section 3 of D3.1 we described a preliminary outline for an EMBRC e-infrastructure, and the workflow scenarios described in D3.2 validated this outline. The components of this infrastructure are: network access, storage capacity, computational resources, software, and human resources.

Network access is required to allow transfer of molecular and imaging data to and from EMBRC partners and sequencing facilities, off-site computational facilities, public databases, and collaborators. Whereas maintaining a basic network connection is a core responsibility of each EMBRC partner, the e-infrastructure should ensure a minimal level of network access to enable such work. In D3.1 we recommended that all partners should aim to have connection speeds of at least 100 Mbps in order to handle high-volume data transfers and remote computational requirements. This requirement is likely to increase as new technologies allow ever-faster data generation. However, some marine stations are located in remote locations where access to the network is significantly more expensive compared to central, urbanised locations.

Network access, while part of the e-infrastructure, is also a component of each partner's physical infrastructure, much the same as buildings, electricity, and a water supply. The responsibility for installing and maintaining the network access infrastructure therefore lies with each partner and not

with the EMBRC. The 100 Mbps target should be achievable for all EMBRC locations except perhaps for those on islands (e.g. Kristineberg, Ischia).

Storage. High throughput molecular data, as well as imaging and video data, require storage and computational resources (racks of processors) to manipulate and analyse those data. At present nucleotide sequences, in particular raw sequence data, comprise the largest demand on both storage and computational resources, although images and in particular videos could in principle demand even more storage.

Computational resources. These include hardware in the form of storage, RAM, and computing cores as well as the software required to store, process, and analyse experimental data.

Data/analyses experts. The proliferation and quickly changing landscapes of bioinformatics algorithms, data repositories, data and metadata sharing technologies, and standards puts an enormous burden on marine scientists. It is impossible for most marine scientists running sampling campaigns or experimental studies to also stay up-to-date for making informed and sustainable decisions on data management, analyses and archival/publication. The need for dedicated personal for supporting the increasingly interdisciplinary marine science projects of the next decades on this side has been strongly emphasized during our discussions.

A major resource for the EMBRC to tap from is the locally funded marine science experts based in several nodes who have already developed expertise in various fields. Their leading expertise could be made available to other EMBRC users via the e-infrastructure. An economical model would be for e-EMBRC to match-fund of their salary to free up time for networking activities, such as remote service provision, hosting EMBRC users or training provision.

6. Blueprint for the EMBRC e-infrastructure

6.1 Model for an e-infrastructure

The discussions held during the WP3 workshops identified three possible EMBRC e-infrastructure models:

Model 1: Create a central bioinformatics facility at one of the partners to serve the needs of all partners. This facility would have hardware (computational facilities and storage), software for analysis, and a full time staff dedicated to serving the bioinformatics and computational needs of all EMBRC partners. The advantage of such a facility is that it is more efficient than distributing resources higgledy-piggledy at different partners, and can have staff with enough expertise to cover most of the needed work. A downside however is the lack of proximity with all other partners, which would only be partially filled by staff exchanges or remote access arrangements.

Model 2: Create an evenly distributed e-infrastructure across the EMBRC nodes: Such an infrastructure would build on existing facilities and local bioinformatics expertise. This model would meet the desire for proximity to scientists that has been expressed by partner scientists in WP3 discussions and surveys. It would also facilitate interaction with a diversity of EMBRC stakeholders and users across all scientific areas covered by the EMBRC nodes, underpin a thorough collective

understanding of locally available, complementary electronic infrastructures, as well as interdisciplinary requirements for data management policies. Towards this model, a number of current or aspiring EMBRC partners already have a bioinformatics or computational facility.

Given the cost, particularly for staff, it is unrealistic to expect all partners to have a bioinformatics or computational facility or to provide expertise across many different fields: Sample and laboratory records management, DNA and 'omics technologies, genome and molecular data analysis, data storage, systems administration, and programming skills. No one person can have all of these skills, so a facility with only one or two staff members will always find itself short of needed expertise. Another major drawback of a widely distributed e-infrastructure would be to entail significant managerial challenges.

Model 3: Create a two-level integrated e-infrastructure :

This model distinguishes between EMBRC partners equipped with significant e-infrastrucure related resources (network access, storage and computing capacity, IT and bioinformatics skills), and already involved in bioinformatics initiatives at a national level, and partners with less locally available resources. Partners of the first group could then act as suppliers or as hosting facilities for projects carried out by partners of the second group. The duplication of skills is reduced by dedicating nodes either to a technology part of the e-infrastructure (data archiving, HPC, bio-analysis skills, etc.) and/or to bioinformatics skills (phylogeny, sequencing analysis, etc.). This hybrid model between the central and fully distributed models improves potential exchanges between all nodes.

For maximum cost efficiency, as well as leveraging local match-funding opportunities, any facility should build upon the hardware and complementary expertise available at other partners. The EMBRC e-infrastructure should also act to maximize the expertise and facilities already in place at other EMBRC nodes by coordinating sharing of these resources whenever possible.

We propose that EMBRC should create a partially distributed bioinformatics facility involving a restricted number of partners that will serve the needs of all partners. This facility will have hardware (computational facilities and storage), software for analysis, and a staff dedicated to serving the bioinformatics and computational needs of all EMBRC partners. One lead partner should be responsible for the overall management and coordination of the facility, and will likely have most of the full-time EMBRC bioinformatics staff. Other nodes would share their computational resources and would formally support the central bioinformatics facility by allocating time from locally funded bioinformaticians and data scientists to external users.

6.2 Function of the e-infrastructure

The e-infrastructure will have four primary 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

The first three of these functions involve direct interaction with EMBRC users, and will likely require travel for EMBRC users to the core facility(ies), and/or to other nodes with the appropriate facilities, as well as frequent travel for the e-infrastructure staff to EMBRC nodes.

6.3 An EMBRC bioinformatics facility: when and where?

The decision on when, and whether, to establish an EMBRC e-infrastructure should be made by the EMBRC executive director and board. However, it seems likely that the bioinformatics facility will be an important part of EMBRC's provision of services to its users, and it would ideally commence when EMBRC begins its operational phase in 2016, or soon thereafter. The biggest bottleneck in creating the facility will be funding, so planning to secure funding will be an early task for the EMBRC construction phase. In particular, the question of whether to include a request to fund the e-infrastructure into the primary H2020 EMBRC application, or whether to keep these separate, should be addressed. Identification of funding sources, and the decision on integration or separation of e-infrastructure funding, will be tasks for the EMBRC executive director during the construction phase. However, we note that H2020 funding schemes, in particular INFRAIA-1- 2014/2015, include specific calls for construction of e-infrastructures that are likely to be suitable for an EMBRC application.

The bioinformatics core facility should be embedded in a restricted number of EMBRC partners, under the coordination of a leading node. This will ensure that the bioinformatics team is in close contact with marine biologists and wet lab scientists and enable them to keep abreast of the needs of "wet" scientists and of new laboratory and field technologies. We suggest that the EMBRC executive director and secretariat should design an open call for hosting the bioinformatics core facility, similar to the call made for hosting the EMBRC secretariat itself. The decision on where to place the bioinformatics core will be conditional on prospective funding, and could prove complex: both tasks should be initiated early during the construction phase if the EMBRC plans to establish a bioinformatics team early during the operational phase.

The criteria for choosing the location of the core facility will rest with the executive director and the board. However, these should include consideration of the available physical infrastructure. Does the core facility have fast and reliable internet access, enough space, an uninterruptable power supply, and sufficient IT staff to support the facility?

7. E-infrastructure: people

The term "e-infrastructure" conjures images of cables, hard disks, and computer racks. In reality, it is people who are the key components of the system. The proliferation and quickly changing landscapes of bioinformatics algorithms, data repositories, data and metadata sharing technologies, and standards puts an enormous burden on marine scientists. It is impossible for most marine scientists running sampling campaigns or experimental studies to also stay up-to-date for making

informed and sustainable decisions on data management, analyses and archival/publication. The primary component of the EMBRC e-infrastructure, then, will be the staff who can advise users.

The EMBRC e-infrastructure team should include experts who can give advice, participate in research projects, and provide training to EMBRC users. It is for this reason that "providing advice" is listed in section 6 as the first of the EMBRC e-infrastructure functions. The three primary functions of the e-infrastructure team will be to provide advice on sample management, data generation/analysis, and training. These are described in detail in section 7.1. In principle, the EMBRC e-infrastructure will employ at least one person with expertise in every category. In reality, this is unlikely, and the hiring team will have to rank expertise needs and hire first those whose expertise has the highest user need. As additional funds become available experts in other categories can be taken on.

7.1 Data management and bioinformatics functions

The guiding principle of the e-infrastructure team is to fill the gaps between local expertise and national expertise. This will require that the team keep track of the skills available at each EMBRC partner and at the various national service providers. Additionally, the team members will have to keep abreast of current developments in relevant technologies, keep up-to-date with the literature in relevant fields, and network within EMBRC and the greater data management and bioinformatics community. Below is an overview of the services that the e-infrastructure team should provide.

7.1.1 FIMS/LIMS support

Field and laboratory information management systems track biological material and laboratory samples from the point of collection through all data generating experiments. Field information management systems (FIMS) are used to track samples collected in the field, as for instance marine samples collected from aboard a ship, and then to track those samples as they are brought to a laboratory and processed for various experiments. Laboratory information management systems (LIMS) track samples during laboratory experiments. This tracking may become quite complex, as a particular biological sample is usually subdivided for different types of experimental work: for instance morphological analysis, chemical assays, DNA sequencing, and photography. Each subsample may be sent to a different laboratory for processing, and data from different experiments may be deposited in different data repositories. Ideally, each datum (DNA sequence, image, biogeographical record, morphological finding, etc.) should be associated with the original sample, with the metadata from that sample (lat/long, temperature, depth or elevation, weather conditions, time, etc.) and with all other data derived from that same sample in other databases. EMBRC data scientists can fill a need by aiding EMBRC users in choosing the right FIMS/LIMS tools and then guiding users in their implementation.

7.1.2 Genomics and sequencing support

EMBRC scientists are engaged in a large number of projects that use data from genomics and nucleic acid sequencing as tools to address scientific questions. These projects using metagenomics, DNA barcoding, genome sequencing/analysis, transcriptomes/RNA-seq, and others, are much more

valuable if they are planned from the beginning so that sample collection, laboratory methods, data management, and data analysis are chosen in light of the experimental questions being asked.

EMBRC e-infrastructure bioinformaticians and data scientists should be tasked with keeping up with current laboratory techniques, data types, and analysis methods in order to provide users with advice on designing experiments, data management workflows, data storage, data analysis, and long term data storage, which will in most cases mean deposition in public repositories.

7.1.3 Enabling data access

The European Commission and almost all national funding bodies require grant recipients to adhere to a data access policy. Such policies usually state that, whenever possible, data from publicly funded research should be placed in public repositories for public use. In general the only exceptions to this policy are for medically related data from human patients. EMBRC will be legally required to abide to these policies for publicly funded activities.

In addition, BioMedBridges and the BMS Research Infrastructures are actively cooperating to draft a data management and sharing charter. EMBRC WP3 has been involved in this process, and we strongly support EMBRC participation in this cross-infrastructure initiative to promote data sharing and coordination of policies.

We recommend that EMBRC should have a formal data policy that complies with the legal requirements of funders and should also actively participate in drafting the cross-infrastructure data management charter, with the aim of fully adopting the measures recommended in this charter. Drafting the policy, as well as participation in drafting the data management charter, should be the responsibility of the Director and the secretariat with technical help from the bioinformatics team as required. Drafts of the current version of the data management charter are available from the BioMedBridges project manager (www.biomedbridges.eu). A specificity of EMBRC is that marine environmental datasets are particularly rich in a broad diversity of metadata, which also require ad hoc data management. A critical aspect of the EMBRC data management activities will be to understand and iron out the data management issues arising from interdisciplinary projects.

On a practical level, EMBRC bioinformaticians and data scientists should then be tasked with understanding the EMBRC's data management policy, and with aiding EMBRC users in implementing it. This will probably mean creating a data management plan for each EMBRC project that will describe what types of data will be generated, which public repositories these data will go to, and a plan a timeline for deposition of those data. In general molecular data will likely go to various EMBL-EBI resources, biodiversity data to EurOBIS and LifeWatch, georeferenced data to PANGAEA and image data to the proposed Euro-Bioimaging respository. As a matter of principle the EMBRC bioinformatics facility should be used for storage of data for ongoing projects only: EMBRC should not create public data repositories for its users' data—this would be prohibitively expensive, and would duplicate larger and better-funded resources.

Note that this policy may not apply to commercial users whose projects are not publicly funded. They will have the right and expectation that data they generate remain private, at least until any commercial potential is exhausted.

7.1.4 Training

The primary task of the EMBRC e-infrastructure team will be to transfer knowledge to EMBRC users. In addition to advising users on specific projects as described above, the e-infrastructure should:

- Keep track and promote to EMBRC users relevant courses at other organizations (such as EMBL, ELIXIR).
- In collaboration with all of the EMBRC nodes, develop and organize marine-specific training courses in topics for which there is a perceived need.

Some EMBRC nodes, for example SAMS, have considerable experience in organising such training and knowledge exchange activities, and have developed cost-efficient mechanisms to develop training materials targeted at different audiences. One model is to subsidise the organisation of workshops or summer schools targeted at specialists, where cutting edge training materials are contributed by international leaders. The contents of these courses could then used by local e-EMBRC staff as a basis to develop undergraduate and postgraduate modules, as well as short continuous professional development courses. The French partners provide bioinformatics training in several domains such as sequence analysis (RNA-seq, etc.) and analysis pipelines, and the SZN bioinformatics staff has more than 10 years experience teaching in courses in large-scale data analysis and specific bioinformatics application program interfaces (API).

This variety of formats not only reach out to a broad audience of stakeholders, but also provides a mechanism to ensure the long-term financial sustainability of the training activities, and ultimately helps to meet the long-term demand throughout Europe for a skilled labour force.

8. Hardware requirements

A number of EMBRC partners now have, or have funding to install, local data centres and computational infrastructure. For instance, the French partners (Banyuls, Roscoff and V/mer) provide a national bioinformatics platform with, 700 Tb secured storage, 10 Gbps Internet connection, a dedicated 800-core cluster (1500 in 2014 (Institut Français de Bioinformatique - French ELIXIR node and EMBRC-France funding) and a Galaxy workflow platform; SZN plans to install a system with 40 Tb of storage and three 64-core computational nodes, with associated connectivity; SAMS has recently been awarded NERC funding to install a 320-core cluster, 65 Tb of secure replicated storage, and a 2-year full time position to develop a data management policy for marine environmental data. These two computational centres are designed to serve the bioinformatics needs of a single partner at a cost of €100,000-400,000 each, including some salary support for IT and scientific staff. National and supra-national bioinformatics providers, with much larger hardware installations, will provide both storage and computational infrastructure for some users. The EMBRC e-infrastructure should aim to provide storage and computational power that enables partners who lack their own infrastructure to undertake projects, and that store data and analysis results from ongoing projects. The EMBRC e-infrastructure should specifically not act as a long-term centre for data storage. Data from completed projects should be transferred to the

appropriate public data repository. This will allow EMBRC to maintain a good service without the need to continuously add storage capacity.

It is not now possible to make specific recommendations for the hardware needs of the EMBRC bioinformatics facility. Given that rapid development of storage and computational hardware the equipment available in two or three years will be different from what is available now. Instead, we recommend that EMBRC partners and other experts be consulted when it is time to develop these plans. Furthermore, the e-infrastructure is not a fixed entity—additional hardware and staff can be added, assuming funding is available, if demand warrants expansion.

9. Putting flexibility at the heart of the EMBRC e-infrastructure

As highlighted above, both the technical specifications of the hardware and the e-infrastructure landscape that EMBRC is expected to address change extremely rapidly. Most of these parameters (e.g. creation of a national facility in a member country) are beyond EMBRC control, but they need to be taken into account in order to meet the guiding principles described in Section 5.

It is therefore critical to design, from its inception, a mechanism that embeds flexibility in the EMBRC e-infrastructure. Unless such a mechanism exists, the EMBRC e-infrastructure would be vulnerable to losing relevance in the face of emerging user needs. At the same time, it is important to guarantee some degree of long-term stability to the e-infrastructure, in line with the vision that is at the core of the ESFRI concept.

An option might be to devote 80-85% of the e-infrastructure budget to provide core services, including training, and ring-fence the remainder of the funds (e.g. 15-20%) of the e-infrastructure budget for an annual call open to all EMBRC partners on a competitive basis. This annual call would enable meeting evolving user demand, might fund specific system upgrades at any one node (e.g. hardware or connectivity improvement), and could quickly respond to opportunities such as the provision of match-funding to leverage additional resources from a variety of funders. This call might also explicitly favour joint initiatives in order to foster networking between the partners. Thanks to its inclusive nature, such a mechanism would provide an incentive for all EMBRC partners to strongly engage with EMBRC. Proposal evaluation would stay with the steering committee, possibly with consultative input from the core bioinformatics facility to provide technical advice, general coordination and oversight on the EMBRC portfolio.

A second, non-exclusive, mechanism would be to for the steering committee to review the e-infrastructure on a regular time frame commensurate with the fast evolution of the field (i.e. every 2-4 years).

10. Integration with other infrastructures and projects

A major task for the EMBRC bioinformatics team will be to ensure that EMBRC users take full advantage of national level facilities, such as ELIXIR nodes, publicly-supported bioinformatics facilities, and EC-funded cloud initiatives, as well as commercial services that are likely to be more cost-effective than locally administered hardware for some projects. These facilities, where

available, are provided to aid biomedical researchers in storing and handling large data sets: using such facilities when they are available will allow EMBRC to concentrate on its core remit of supporting marine science. This does mean, though, that the EMBRC bioinformatics facility staff must maintain knowledge of what national facilities exist and how to use them.

It is likely that most of the national facilities will be ELIXIR nodes: the EMBRC e-infrastructure team should work together with the various Elixir nodes in order to understand the Elixir infrastructure and should work to create services that complement, but do not duplicate, those available from ELIXIR. As of the end of 2013 the various ELIXIR nodes are still starting up so it is not possible to say what services they will provide. However, it seems likely that ELIXIR nodes will give users access cpu time for running analyses and to some storage. The EMBRC team should aid EMBRC users in taking advantage of these resources. Additionally, the Norwegian ELIXIR node at University of Tromsø, who may join the Norwegian EMBRC node, will focus on marine biology so the EMBRC team should work with the Tromsø team to avoid duplication of effort.