General requirements for EMBRC

Context of general requirements in EMBRC

Complete EMBRC report available at:

https://envriplus.manageprojects.com/projects/requirements/notebooks/470/pages/40

EMBRC (the European Marine Biological Resource Centre [1]) is a distributed European RI which is set up to become the major RI for marine biological research, covering everything from basic biology, marine model organisms, biomedical applications, biotechnological applications, environmental data, ecology, etc. Having successfully completed a 3-year Preparatory phase (2011-2014), it is now in its Implementation phase (2014-2016), and operation is planned to start in 2016-2017. It has 9 European countries and associated countries as full members, the stations and laboratories of which contribute their facilities, equipment and human capital to the infrastructure. Apart from ENVRIplus, EMBRC is involved in the biomedical cluster CORBEL ([2]), and in the marine cluster, EMBRIC ([3]).

The main purpose of EMBRC is to promote marine biological science and the application of marine experimental models in mainstream research by providing the facilities (lab space), equipment (e.g. electron microscopes, real time PCR machines, crystallography, lab equipment, equipment for accessing the environments such as research vessels, scientific divers, ROVs etc.), expertise and biological resources that are necessary for carrying out biological research. Users (scientists, the private sector, SMEs) who are interested in working on a particular marine organism can browse through the EMBRC catalogue of labs and facilities and submit an application for visiting one or more sites (from one of the EMBRC member countries). If their application is accepted, during their visit they can either collect organisms using the EMBRC equipment, or EMBRC can collect them for them, and train them on working on them. They can also set up cultures for which EMBRC provides the access, or EMBRC can set them for them. The users can then perform the experiments that they like, depending on the purposes of their research. They can take away the preserved organisms, or they can work in the EMBRC labs to produce the data that they need for their research.

In what concerns data, the role of EMBRC is to generate and make it available. It does not usually do any analysis on the data, unless it is contracted to do so. Data is usually generated through sensors in site in the sea or samples that are collected and then measured in the lab. *Envir onmental data*, which is largely produced by the marine biology community through contracts paid by national research councils or environmental agencies, is mostly provided free of charge in public databases (e.g. PANGAEA [4] or the NERC [5] database named BODC [6] in the UK, EMODNET [7]). EMBRC acquires the data and submits it in raw form, depending on the project, to these national or international open access databases. In what concerns the *molecular data* (anything having to do with omics) that is generated by the EMBRC or by its users, the scientists from member institutes or the users of EMBRC usually do some work on the data to curate it and, if part of a bigger project, they may perform some annotation and assembly. As part of the data policy, users who are scientists and generate molecular data will be required by the EMBRC to deposit it in an open access database. They may impose some timing restrictions depending on the purposes of their research (e.g. until they publish), but they are usually given a deadline to submit the data to the database, after which the EMBRC will submit it. Private sector users retain the IPR of the data that they generate, and the EMBRC cannot impose that they deposit their data, or where to deposit it.

References:

1. http://www.embrc.eu/
2. https://www.elixir-europe.org/about/eu-projects/corbel
3. http://cordis.europa.eu/project/rcn/198465_en.html
4. http://www.pangaea.de/
5. http://www.nerc.ac.uk/
6. http://www.bodc.ac.uk/
7. http://www.emodnet.eu/

Summary of EMBRC general requirements

EMBRC would like to achieve several objectives through participation to ENVRIplus:

- **Establishing collaborations with the environmental community**, which would benefit from their environmental and ecological data: EMBRC has been heavily involved with the biological and biomedical communities so far, but not with the environmental community.
- **Developing and learning about new standards and best practices in terms of standards**: the EMBRC community are very interested in considering new standards in terms of methodologies, workflows and data. They would like to work towards increasing the compatibility of the data generated from their various member institutes such that its analysis can become comparable, but at the same time not lose the compatibility of this data with data which was already generated, sometimes as far back as 100 years ago.
- **Developing new standards within INSPIRE [8], which can be used for other datasets**: While INSPIRE is particularly focused on spatial data, its disadvantage is that it does not apply for other types of data. Through participation to one of the ENVRIplus work packages, the the RI REP advised that his organisation- the MBA (Marine Biological Association [9]), a member of the EMBRC, will attempt to develop standards within INSPIRE which can be used for other datasets by EMBRC.
- **Exploring new data workflows which make use of marine biological and ecological data.**
- **Networking with other RIs**.

The following are some important priorities for EMBRC:

1. **Setting up an e-infrastructure within the next year**, which would need to provide appropriate connectivity in order to facilitate the movement of data between the labs, in cases where users want to sample data from different locations, and send it to various repositories.
2. **Developing several web resources (information about various species, availability of genomes, trasnscriptomes, mutants etc) and platforms**: In particular, they are working towards building their access portal, which will include a catalogue of all of their lab locations, equipment, organisms, e-infrastructure services which users (researchers within or outside the marine sector, private sector, SMEs) can search and apply for. The EMBRC community is working on connecting external resources to their access portal, such that information on all of their available molecular resources is available from the same place. They expect their access portal to be ready within at most one year.
3. **Negotiating with other RIs within ENVRIplus, the biomedical cluster Corbel and the marine cluster EMBRIC how they could work together to provide their services.**

The biggest challenge for EMBRC will be looking at the different standards and workflows of the 3 clusters that it is involved in, and deciding on common ones, both for its member institutes, but also with other RIs across domains to facilitate collaboration.

References:

8. http://inspire.ec.europa.eu/index.cfm/pageid/48

9. http://www.mba.ac.uk/

Detailed requirements

EMBRC has plans to develop an e-infrastructure which would need to provide appropriate connectivity in order to facilitate the movement of data between the labs, in cases where users want to sample data from different locations, and send it to various repositories. The EMBRC community have also discussed about having a dedicated data group, which could advise users with their experimental designs (how and where to analyse their data, where to deposit it) and provide support for sequence assembly and annotation.

EMBRC are discussing with the CORBEL and EMBRIC clusters the possibility of users going to multiple RIs to generate and analyse data, and investigating how difficult this would be to implement. They are considering using ELIXIR [10] for data curation, in particular working with the marine node from Tromso, Norway.

The EMBRC community do not use common software or standards. Some member labs use specialised in-house, non open-access, software, while others open-access one. In terms of standards, the RI REP advised that his organisation, the MBA, is using GBIF based on Darwin Core [11], and the MEDIN [12] metadata standard in the UK which is compliant with the European INSPIRE directive.

EMBRC does not have any non-functional constraints for data handling and exploitation. In what concerns security and access, the RI REP could only tell me about the situation of the MBA, who are using the ISO 27001 Information Security Management standard [13].

EMBRC can share with ENVRIplus or with other RIs instrumentation in terms of a number of buoys that are connected to various labs. It can also provide detectors and lab equipment, which users will be able to apply for through an access portal. Expertise in areas such as taxonomy and specific model organisms is spread across its different member labs, and EMBRC is currently making an inventory of it and of the willingness of the different labs to provide it as a service. It is also investigating the option of having a dedicated data group, which could provide help and support on experimental design, and analytical help on sequence assembly and annotation. As several stations from member countries have small libraries with large amounts of grey literature, some going back hundreds of years, EMBRC intends to run a workshop in the spring of 2016 to investigate how it could connect these libraries in order to make more of this grey literature available.

The biggest challenge for EMBRC will be looking at the different standards and workflows of the 3 clusters that it is involved in, and deciding on common ones, both for its member institutes, but also with other RIs across domains to facilitate collaboration. The community would like to organize a meeting between ENVRIplus, EMBRIC and CORBEL to discuss these issues. One option could be to the rolling out of the ENVRI Reference Model within other RIs.

Questionnaire answers from EMBRC/St Andrews

1. What is the basic purpose of your RI, technically speaking? university
    a. Could you describe one or several basic use-cases involving interaction with the RI that cover topics 1-7 ?

        > We engage in a variety of different types of research although no long term environmental data acquisition directly ourselves to my knowledge

    a. Could you describe how data is acquired, curated and made available to users?
    > Fairly ad hoc depending on project. There is no insititutional policy.

    b. Could you describe the software and computational environments involved?
    > They vary. Some people use servers for big data analysis others just desk tops and laptops.

    c. Could you describe the software and computational environments involved?

        > They vary. Some people use servers for big data analysis others just desk tops and laptops.

2. Possiblity data collection, often data analysis

    a. Do you have any use case involving interactions with other RIs (in the same or different domains?)

    > Not to my knowledge

3. What datasets are available for sharing with other RIs ? Under what conditions are they available?

    > People may share data on a personal basis but I know of no institute to institute sharing arrangments

> Yes. It is specialist software for the analysis of population genetic/genomics and environmental data. The software is open source and is being used widely throughout the world.

4. Apart from datasets, does your RI also bring to ENVRIplus and/or other RIs:

a. Software? In this case, is it open source?
> Like other universities we do produce software (e.g. in our case *Distance* as well as statistical analysis packages (i.e. R libraries)

b. Computing resources (for running datasets through your software or software on your datasets)?
> No

c. Access to instrumentation/detectors or lab equipment? If so, what are the open-access conditions? Are there any bilateral agreements?

> We build pods for attachment to marine mammals to give details of environment and movement.

d. Users/expertise to provide advice on various topics?

> Yes

f. Access to related scholarly publications?

> Yes

g. Access to related grey literature (e.g. technical reports)?

> Yes

5. What objectives would you like to achieve through participation to ENVRIplus?

> Knowledge of best practice in handling data

6. What services do you expect ENVRIplus technology to provide

> None

7. What plans does your RI already have for data, its management and exploitation?

a. Are you using any particular standard(s)?

> http://www.st-andrews.ac.uk/staff/research/data/policiesandguidelines/university/Strengths and weaknesses

b. Are you using any particular software(s)?

> Lots!

c. Are you considering changing the current:

> We are always open to change.

8. What part of your RI needs to be improved in order:

a. For the RI to achieve its operational goals?

> Unsure

b. For you to be able to do your work?

> Centralised back up of data might be useful

> There is a need for a computer engineer who could maintain the software packages already produced and help improve them and integrate them into a single platform. At the moment everything is done by PhD students, most of whom stop maintaining the programs they developed once they finish their studies.

9. Does your RI have non-functional constraints for data handling and exploitation? For example:

> Maintenance costs , Operational costs

> Security and Privacy we have freedom of information issues that cause problems

10. Do you have an overall approach to security and access?

> yes, https://www.st-andrews.ac.uk/itsupport/security/classification/

11. Are your data, software and computational environment subject to an open-access policy?

> some are

12. What are the big open problems for your RI pertinent to handling and exploiting your data?

> None that I am aware of

A perspective from within SMRU, just one small part of the EMBRC.

Being the European Marine Biological Resource Centre, EMBRC data may often fall within more of a "biological" rather than an "environmental" category. However, there will be many projects which do collect at least some definitely "environmental" data.

The Scottish Oceans Institute (SOI [14]) at the University of St Andrews is only partner within the EMBRC .The Sea Mammal Research Unit (SMRU [15]) is just one part of the SOI. These notes are only intended to give a really quick overview of things from the perspective of some of the areas of activity within SMRU. They specifically don't cover the SMRU/SOI projects that involve the more traditional areas of bioinformatics, such as genetics and genomics (which are likely to be a common thread across the EMBRC partners) and many of the seal pool based studies. Given the number and diversity of the partners to get any really EMBRC wide view it would probably be necessary to start by contacting the EMBRC central office (info@embrc.eu).

An overview of the research within the Sea Mammal Research Unit can be found here [16]

Types of data collected

SMRU collects quite a variety of types of data for a range of purposes including:

Telemetry data collected using tags, mainly built within SMRU:

SMRU Instrumentation tags (locations, dives, CTD, etc.) [17]

DTAGS (sound, acceleration, etc.) [18]

Photos and video from surveys, ID studies, etc.

Audio recordings using hydrophones, DTAGs

Sightings (e.g. time, location, number, species) which may be visual or automatic (e.g. pods)

Specific examples of more general "environmental" data would include:

Temperature and CTD records from SMRU Instrumentation tags

Measurements of ocean noise made by DTAGs and hydrophone arrays

Photos of the coastline taken as part of the aerial surveys of seal populations

Data management and dissemination

As parts of the University of St Andrews SMRU and SOI come under the University's general research data policy [19] and management guidelines [20].

Much of SMRU's research, particularly the long term projects, have been funded by the UK's Natural Environment Research Council and so more specifically fall under NERC's data policy [21]

Any current and new NERC funded projects now require have an appropriate NERC approved data management plan [22]

For NERC funded projects the designated data archive centre is the British Oceanographic Data Centre (BODC [23]) and by default such data should be considered for being made publically accessible within 2 years of the end of data collection. There is also official lab liaison program between SMRU and BODC [24]

As part of that collaboration the near term goals include beginning to increase the discoverability and public accessibility of at least SMRU's NERC funded data by creating more EDMED/MEDIN records and when appropriate adding datasets to the BODC Published Data Library [25]

An in-house system, SMRUDAS, being developed to help manage data and metadata with the Unit and facilitate passing individual datasets onto places like BODC for longer term archival and public access. One aim being to collect enough metadata during the course of a project so that at the end, it will be fairly straight forward to generate records to be included in metadata catalogues such as EDMED [26] and MEDIN [27]. As such internally SMRUDAS makes a use of a number of controlled vocabularies (NERC, ICES, WoRMs, etc.)

Case study – SMRU Instrumentation Group telemetry data

The Instrumentation Group has designed and built tags for use by biologists and oceanographers from around the world for many years. As a result it has had to develop one of the more advanced systems within SMRU for handling, storing and disseminating data.

Incoming data from tags is automatically processed on the group's servers and then stored in an Oracle database maintained by the University's central IT Service.

Data from many SMRU CTD tags goes out in near real-time over the GTS, via BODC, for use in forecasting [28]

Telemetry data from the tags is also made available (currently as MS Access databases, KMZ and ODV files) to the scientists who deployed them via a dedicated, password protected website [29]

Quality controlled versions of much of that CTD data eventually also becomes publically available via the MEOP (Marine Mammals Exploring the Oceans Pole to Pole) website [30]

References:

10. http://www.elixir-europe.org/

11. http://tools.gbif.org/dwca-assistant/

12. http://www.oceannet.org/marine_data_standards/medin_disc_stnd.html

13. http://www.iso.org/iso/home/standards/management-standards/iso27001.htm
14. http://soi.st-andrews.ac.uk/
15. http://www.smru.st-andrews.ac.uk/
16. http://www.smru.st-andrews.ac.uk/pageset.aspx?psr=136
17. http://www.smru.st-andrews.ac.uk/Instrumentation/Overview/
18. http://soundtags.st-andrews.ac.uk/
19. https://www.st-andrews.ac.uk/staff/policy/research/researchdata/
20. http://researchdata.wp.st-andrews.ac.uk/
21. http://www.nerc.ac.uk/research/sites/data/policy/
22. http://www.nerc.ac.uk/research/sites/data/dmp/
23. http://www.bodc.ac.uk/
24. http://www.bodc.ac.uk/partners/research_centres/smru/
25. https://www.bodc.ac.uk/data/published_data_library/
26. https://www.bodc.ac.uk/data/information_and_inventories/edmed/
27. http://www.oceannet.org/
28. https://www.wmo.int/pages/prog/www/TEM/GTS/index_en.html
29. http://www.smru.st-and.ac.uk/protected/technical.html
30. http://www.meop.net/

Formalities (who & when)

Go-between	Cristina Adriana Alexandru
RI representative	Nicolas Pade
Period of requirements collection	September-October 2015
Status	finalised