General requirements of ELIXIR

Context of general requirements in ELIXIR

The requirements collected for ELIXIER are based on the questionnaire but they don't follow the questions in the prescriped order. In essence only a few but important use cases are described, reflecting only several aspects of ELIXIR, and therefore clearly not offering a comprehensive view of ELIXIR which is simply too complex to be specified in each details. This format has been chosen to prevent misleading and incorrect generalisation of ELIXIR.

Summary of ELIXIR general requirements

Multi-Topic requirements for ELIXIR

https://www.elixir-europe.org/

Aim of ELIXIER is to build a sustainable European infrastructure for biological information, supporting life science research and its translation to medicine, agriculture, bioindustries and society. ELIXIR unites Europe's leading life science organisations in managing and safeguarding the massive amounts of data being generated every day by publicly funded research. It is a pan-European research infrastructure for biological information.

The European Nucleotide Archive (ENA) is one of the databases which is core to the RI. Petra Ten Hoopen is mainly working for this database.

It contains primary data with all the nucleotide sequencing information and associated information. The sequences are the data and the associated is the contextual information on the sequences such as

- Sequencing studies
- Sequence experiments
- Information of provenance of sample material
- Analysis of the sequences
- Metadata that describe the data

ENA records this information in a data model that covers input information (sample, experimental setup, machine configuration), output machine data (sequence traces, reads and quality scores) and interpreted information (assembly, mapping, functional annotation).

1. data acquisition, data curation and availability of data to users:

several sources:

1. data are coming to the European database directly by European scientists, which are the submitters of primary data (real sequences)
2. brokered data from expert database and already quality assured
3. from partners such as INSDC partners

several ways of transmission:

data is coming to the archive in small or big scale via pipelines which are pre-established for an automated way or via interactive interface, supported by direct contact with helpdesk

Once data is archived, they are processed, curated manually or in an automated way and undergo validations for quality assurance.

Data presentation/publication:

Data is presented either directly for browser, searches and download or it serves as underlying backbone of other databases as secondary sources

1. ELIXIER offers interactive support as well as programmatic support for the flow of the data with validation.
2. The users are actually owners of the data, the archive is only responsible for its presentation.

The user also decides whether the data should be kept confidential or should be made publically available.

The user can keep data confidential for two years after submission – till this deadline he/she has either to contact the RI for prolongation of the confidentiality period justifying this with documented reasons or in case of no response, he/she accepts to open access and publication of the data. The RI encourages the open access policy of data.

1. Interaction with other RIs – Microbe3 with Lifewatch with Seadatanet

Detailed requirements

Use case example 1

The European Nucleotide Archive (ENA) is one of the many diverse services within the ELIXIR RI. It counts amongst the most established (some 30+ years) and, as a generalist and comparatively low-level component of infrastructure, faces broad use across the RI's stakeholders. In this use case, we illustrate in particular marine science-facing elements of ENA.

ENA is maintained at the EMBL-EBI, which serves as one of the hub organisations of ELIXIR, and provided from the extensive EMBL-EBI computational infrastructure. As one of the largest of the ELIXIR services (e.g. multi-petabyte data content), technical activity in providing ENA sustainably is substantial.

ENVRIplus contacts have the following roles in ENA: Guy Cochrane – Head of ENA, Petra ten Hoopen – ENA data content.

ENA is a comprehensive open repository for permanent archiving of public domain primary sequence data and associated contextual information, serving both alongside the scholarly literature as the accepted and formal scientific data record for the domain and as a platform for data sharing between scientists before and outside the formal literature publication workflow.

The ENA, GenBank (http://www.ncbi.nlm.nih.gov/genbank/) and DDBJ (http://www.ddbj.nig.ac.jp/) form the International Nucleotide Sequence Database Collaboration (INSDC, http://www.insdc.org/), a global repository of all public domain primary sequence data.

The INSDC-wide supported data model (http://www.ebi.ac.uk/ena/submit/metadata-model) comprises of the following conceptual objects of metadata:

- study (describes the purpose of the sequencing or data analysis effort)
- experiment (describes experimental design of the sequencing)
- sample (describes provenance of sequenced material)
- run (describes files containing sequencing reads submitted as data files)
- analysis (contains secondary analysis results of primary sequence data)

and a variety of data files, covering raw and derived data forms.

ENA records captures and presents input information (sample, experimental setup, machine configuration), output machine data (sequence traces, reads and quality scores) and interpreted information (assembly, mapping, functional annotation).

ENA receives content through its submission services that face expert database brokers and direct sequence data submitters (both small scale data from individual researchers and large scale data from sequencing centres). These services support a ‘push’ model, where the data submitter initiates and drives data flow. Further content is received from daily data exchange with INSDC partners.

While many ENA data submitters are located in Europe, the submitterbase is global. Typically, ENA collaborates – as has been the case with various marine projects – with external parties to establish project-specific standards and to provide dedicated curation and other data coordination support.

All sequence data and metadata are validated, processed and archived. The majority of deposited data is validated against formalised standards that are developed with heavy user community consultation and deployed into the host of submission and validation applications that are made available from ENA (see http://www.ebi.ac.uk/ena/about/reporting-standards).

The focus of scientific curation of the data is in the development and implementation of standards and metadata conventions through the configuration of data submission and presentation applications. The scientific team also provides a helpdesk and training programme to submission and other processes.

ENA content is made available for search, browse and download through a variety of web and programmatic services. Data output services are provided both to the end-user scientific community directly and through a wealth of external open data resources that aggregate, integrate, curate and otherwise add value on top of ENA content. An example ENA record is available at http://www.ebi.ac.uk/ena/data/view/FN123456. Data are also provided to collaborating users throgh the EMBL-EBI Embassy cloud infrastructure.

The ENA provides public access to several software products: (1) the flat file validator (http://www.ebi.ac.uk/ena/software/flat-file-validator), (2) Webin data streamer (http://www.ebi.ac.uk/ena/software/webin-data-streamer) and (3) CRAM toolkit (http://www.ebi.ac.uk/ena/software/cram-toolkit).

INSDC data depositors are deemed the ‘owners’ of the INSDC records, defining the release date of their data into the public domain and retaining editorial control and responsibility for content. While many data sets flow directly into public view, data can be kept confidential at the request of the submitter for a period of time to allow for pre-publication analysis by the submitter. ENA has an open access data policy and encourages depositors to release their data as soon as possible and is active in the brodaer open data sharing discussion.

Specifically around our marine projects, ENA has built interoperability structures with the pan-European oceanographic network of marine stations, the SeaDataNet (http://www.seadatanet.org/) and the European node of the international Ocean Biogeographic Information System (EurOBIS, http://www.eurobis.org/about), which is part of the central taxonomic backbone of LifeWatch (http://www.lifewatch.be/) , an infrastructure for biodiversity research. Standards supported under this work include the ISO8601 compliant date and time reporting and support of the Opean Geospatial Consortium standard for access to sequence data associated with geographic information.

Use case example 2

LifeWatch Greece (https://www.lifewatchgreece.eu/) e-Services and virtual Labs (vLabs) (part of ESFRI LifeWatch).

Genetics + omics e-services (http://rvlab.portal.lifewatchgreece.eu/) – These services include an SSH access to over 100 bioinformatics packages available in our PC cluster (see below for specifications. Pipelines have been set up to cover metagenomics, population genetics, phylogeny, comparative genomics and Next Generation Sequencing (NGS) data analysis. Moreover, in-house parallelised pipelines have been developed for: population genetics, metagenomics, RNA-seq data analysis and functional annotation (parallel BLAST, Interpro search and GO terms). Local ENSEMBL databases for fishes, humans and some ENSEMBL Metazoa have been set up as well as data mining with ENSEMBL APIs.

These services primarily target marine microbial communities. Moreover as part of the genetics+omics e-services a virtual environment for the QIIME (Quantitative Insights into Microbial Ecology) or QiimeVlab has being developed, also available through the LifeWatch Greece portal. QIIME is an open-source bioinformatics pipeline for performing microbiome analysis from raw DNA sequencing data. QIIME is designed to take users from raw sequencing data generated on the Illumina or other platforms through publication quality graphics and statistics. This includes demultiplexing and quality filtering, OTU picking, taxonomic assignment, and phylogenetic reconstruction, and diversity analyses and visualizations. QIIME has been applied to studies based on billions of sequences from tens of thousands of samples.

MicroCT Services (http://microct.portal.lifewatchgreece.eu/)

Micro-tomography (micro-computed tomography or microCT) is a method of non-destructive 3D x-ray microscopy, which allows the users to create 3D models of objects from a series of x-ray projection images, similar to the conventional clinical computer tomography. The MicroCT Service will offer a collection of virtual galleries of taxa which will be displayed and disseminated through a web-based framework, and will allow the user to manipulate the 3D models through a series of online tools or to download the datasets for local manipulations.

Literature mining services - SPECIES (http://species.hcmr.gr,http://species.jensenlab.org)

SPECIES: a standalone command line application capable of identifying taxonomic mentions in documents and mapping them to corresponding NCBI Taxonomy database entries. Given a folder with plain text files, SPECIES based on its taxonomic name and synonym dictionary reports the taxonomic mentions (start, end position in each document), the detected term and the corresponding NCBI Taxonomy database record identifier. Besides binomials following the Linnaean naming convention, recognised taxonomic mentions include acronyms, common names and abbreviations, as well as misspellings and the rest of the naming types supported by the NCBI Taxonomy.

Literature mining services - ORGANISMS (http://organisms.hcmr.gr,http://organisms.jensenlab.org)

ORGANISMS: a web resource supporting taxonomic-mention-based document retrieval of abstracts from the Medline database. A simple web interface enables the user to query for any organism from the NCBI taxonomy and view the corresponding Medline abstracts with highlighting of the relevant organism names. The underlying architecture takes into account both synonyms and the hierarchical structure of the taxonomy. For this reason a search e.g. for Metatheria (marsupials) will retrieve both abstracts that explicitly mention the taxon and all abstracts that mention taxa within it, e.g. the Tammar wallaby.

Virtual environment - MedOBIS (https://medobis.portal.lifewatchgreece.eu/)

MedOBIS: A biogeographic information system that acts as the Regional Repository of Marine Biodiversity Data, freely available on the web, and as a communication / dissemination forum for the Eastern Mediterranean and the Black Sea. MedOBIS provides a common platform for the integration of efforts devoted to Marine Biodiversity in the region. Data are provided to the Ocean Biogeographic Information System - OBIS (http://www.iobis.org) and to the Global Biodiversity Information Facility - GBIF eventually (http://www.gbif.org).

Virtual environment - GBIFgr (http://lifewww-00.her.hcmr.gr:8080/gbifgreece/)

Greek GBIF National Node: A biogeographic information system that acts as the national repository of species occurrence data in Greece, freely available on the web. It was agreed that LifeWatch Greece in HCMR would be become the GBIF national node for Greece to serve eventually the Global Biodiversity Information Facility - GBIF (http://www.gbif.org). The technical setup is done with already several datasets, but the public opening of the service is pending a governmental signature (as GBIF is a multi-governmental organisation).

Virtual environment - GTIS/SpeLoG (http://speciesgreece.myspecies.info/)

Greek Taxon Information System / Species List of Greece: A biodiversity information system that will list all valid species recorded in Greece. The effort to list all species in Greece was started in the 1990s with the last funded effort done as the project of the Greek Biodiversity database (http://greek-biodiversity.web.auth.gr/). However, the latter system is not updated since 2008, even if it was published in 2010 the latest. Locally we developed a database called GTISdb to continue to establish and validate the list(s) group by group that will be progressively disseminated under a ViBRANT/Scratchpad named SpeLoG, a system based on the CMS Drupal managing biodiversity data mainly at species and ecosystem level. SpeLoG will be opened to public in December 2015 with about 15 lists.

Internally, we also have network access to installed copies of the databases of the Catalogue of Life - CoL (http://www.catalogueoflife.org/) and of the World Register of Marine Species - WoRMS (http://www.marinespecies.org/).

Virtual environment - Statistical R (http://rvlab.portal.lifewatchgreece.eu/)

The R vLab makes use of “R” which is a statistical processing environment widely used by scientists working in many biodiversity related disciplines. It supports an integrated and optimized (in respect to computational speed-up and data manipulation) online R environment. This vLab tackles common problems faced by R users, such as severe computational power deficit. Many of the routines operating under the R environment, such as the calculation of several biodiversity indices and the running of the multivariate analyses, are often of high computational demand and cannot deliver a result when the respective datasets are in the form of large matrices. This vLab allows for a predefined, commonly used set of R functions to run on the LifeWatch Infrastructure in order to support large-scale computational and modeling activities.

Data Services (metacatalogue.portal.lifewatchgreece.eu**)** - login required

Data Services provide the users with tools in order to: a) publish their datasets and make them available to the community by providing information that allows a user to locate and access the resource and its curator/creator, b) import their datasets to the LifeWatch Greece Infrastructure and to GBIF or MedOBIS, c) perform biodiversity data and information quality improvement, and d) search about datasets of interest by providing an efficient way of querying semantic networks. The schema of the data that is provided by the users is mapped to the semantic model of the LWI and the data is transformed to LWI format before it is stored to the Infrastructure. The semantic model is based on CIDOC CRM (http://www.cidoc-crm.org/), CRM dig, CRM geo, CRM sci and MarineTLO (http://www.ics.forth.gr/isl/MarineTLO/).

All of the above services are integrated under a common online system available through the LifeWatch Greece portal (http://microct.portal.lifewatchgreece.eu/). The portal requires registration and login and offers users a wide array of virtual computational environments (Rvlab, MedObis vlab) and services (Literature mining services, Data services, MicroCT Services and Genetics services) as well as mobile applications. Users are provided with and integrated, open access, multidisciplinary environment to manage as well as obtain useful information related to their data, as well as overly their data with other similar datasets available via the portal. This however brings complexity of use of this environment alluding to the need for training courses to familiarize users with the portal. Data is initially private but users are encouraged to release it as soon as possible.

Environmental and ecological data are shared with the other LifeWatch projects from all over Europe. Specifically LifeWatch Greece is involved in an initiative to create an ICT Core Construction Committee and integrate this with the European Grid Infrastructures (EGI).

Software developed in JAVA, C++, as well as vector-based languages like R, are available. Also web based applications using HTML, PHP are developed to construct integrated application programming interfaces (APIs), all of which are open source and available through LifeWatch Greece core group upon request.

A local Beowulf Cluster (108 cores, 784GB RAM) serves more than 100 bioinformatics packages and pipelines covering population genetics, phylogeny, comparative genomics and NGS data analysis, as mentioned above.

Some of the Genetics and Molecular Biotechnology instruments currently hosted by HCMR, Greece:

- 3 Next Generation Sequencing machines (Roche 454, Junior FLX, and an Illumina Miseq).
- ABI 3730 automated sequencer
- Dual OpenArray Real Time PCR Platform (Applied Biosystems) for SNP genotyping and gene expression
- 2 Real Time thermal cyclers (PCR machines), (DNA Engine Opticon and 1 Light Cycler 1.5)
- 10 thermal cyclers (PCR machines), (MJ Research, BIORAD)
- Microarray scanner (GenePix 4100)
- Robotic sample manipulator BIOMEK 2000 (Beckman)
- Robotic workstation for automated purification of DNA, RNA, or proteins (QIAcube)
- Photometers (QuantiFluor™-ST, Nanodrop 1000, and other)
- Agilent 2100 Bioanalyzer
- TissueLyser
- Coulter counter (Beckman) for counting particles and cells
- Laminar flow cabinet (Telstar Bio-IIa)
- Computer Cluster for genome sequencer FLX (GS FLX Titanium Cluster)
- Gel electrophoresis apparatuses
- Gel documentation system
- Standard laboratory equipment like centrifuges, autoclaves, speed-vac, deep freezers, ovens, pHmeters etc.
- Skyscan 1172 micro-computertomograph (MicroCT)

Biodiversity and Ecosystem Management instruments:

a) Biological material (e.g. sorting, biomass and abundance measurements, identification of species):

- Precision balances
- stereoscopes
- light microscopes (with still and video facilities)
- rich literature containing taxonomic books and papers

b) Sedimentological material:

- mechanical sieve shaker
- a large water bath for pipette analysis
- incinerators
- ovens

The laboratories are also equipped with the proper facilities for:

- underwater still and video images (underwater still cameras with a variety of lenses and lighting system, underwater housing, etc)
- Skyscan 1172 micro-computertomograph (MicroCT)

The final analysis of the obtained data is carried out using a number of multivariate techniques by means of software packages related to the analysis of environmental and ecological data (PRIMER, DECORANA, SPSS, etc) along with some packages developed within the IMBG.

Existing large-scale sampling structures/facilities/equipments of HCMR, Greece:

- RVs “AEGAIO” and “PHILIA”
- deep sea vessel “THETIS”
- ROVs
- small crafts
- RoxAnn (mapping of the seabed)
- CTDs (measuring in situ of hydrographic parameters in the water column, e.g. PAR, temperature, salinity)
- Niskin water bottles (5l) (determination of nutrients, organic carbon and chloroplastic pigments in the water column)
- hyperbenthic sledges (sampling of BBL fauna close to the seabed and mapping of the seabed, including the upgraded BBL sampling towed sledge system
- grabs, corers (sampling of benthos in the sediment which are also used for grain size analysis of the sediment and determination of chemical parameters, e.g. organic carbon, chloroplastic pigments)
- otter trawls (sampling of demersal fish mostly for further stomach content analysis and genetic analysis)

- WP2 plankton nets (sampling of zooplankton in the water column)
- diving equipment, digital cameras and videos

These facilities are freely available, provided that funding and bilateral agreements are supplied.

HCMR employs experienced, trained personnel as well as experts in microbial ecology, biotechnology, bioinformatics, software engineers and developers. Thus encompassing all the necessary expertise to facilitate usability of all facilities as well as provide advice, consulting and training in the form of workshops, training schools and conferences.

HCMR has numerous academic licenses for high caliber journals and books allowing for full access to scholarly publications.

Technical reports are available through HCMR's technicians and laboratory experts, as well as through online cloud and storage facilities.

We aim to become a training node for ELIXIR Greece, to specifically address issues of environmental microbiology, biotechnology and bioinformatics with a focus on metagenomics and next generation sequencing tools and analysis.

Currently our researchers are well supported to achieve the goals of their work, but given the explosion of data being generated due to high throughput methods currently in use, constant increase in e-infrastructures is a necessity. Our main need is additional e-infrastructures and also consulting from experienced personnel that could potentially provide invaluable training capabilities in the fields described above.

Standards and software will change in according to the upcoming community updates. HCMR follows the data standards via the Genomics Standards Consortium (next meeting to be held in HCMR in June 2016). Software will also be modified to address new needs and requirements proposed and guided by our users.

We aim to train our research and technical community via in-house workshops and training courses, as well as support them to attend other international and national events focusing on training and knowledge transfer. Our institute has already acted as a training host for numerous international relevant workshops and training courses, such as:

- **Computational Molecular Evolution - Practical course - European Molecular Biology Organization (E) - http://www.imbbc.hcmr.gr/content/embo-practical-course-computational-molecular-evolution-come-5-14-may-2014-imbbc-hcmr-2**
- **MicroB3 Summer school - from sampling to analyzing microbial diversity & function - https://www.microb3.eu/events/workshops/micro-b3-summer-school-crete**
- **European Marine Science Educators Association (EMSEA) conference - EMSEA 2015 - https://emsea2015.hcmr.gr/**
- **Workshop on Microbial Diversity, Genomics and Metagenomics - http://www.marbigen.org/content/microbial-diversity-genomics-and-metagenomics**

Virtual environments and e-services related to biodiversity and genetics are being developed and students, technicians as well as researches with backgrounds in ecology, biotechnology, molecular biodiversity and bioinformatics are being trained on these e-infrastructures as part of the LifeWatch Greece project. The training delivery methods are workshop, conferences, hands-on tutorial or hackathons.

Data managers ensure that data acquisition and curation for the LifeWatch Greece project occur in an organized and supervised manner, allowing for high quality data to be deposited in our online databases.

Note these are only use cases, reflecting only several aspects of ELIXIR, and therefore clearly not offering a comprehensive view of ELIXIR.

Formalities (who & when)

Go-between	(barbaramagagna)
RI representative	@Petra ten Hoopen
Period of requirements collection	September - December 2015
Status	finished