EGI-Engage

ELIXIR Competence Centre

Life science requirements analysis and driver use case(s) with implementation roadmap

M6.3

Date	29/02/2016
Activity	SA2
Lead Partner	EGI.eu
Document Status	FINAL
Document Link	https://documents.egi.eu/document/2675

Abstract

The ELIXIR Competence Centre (CC) of the EGI-Engage project facilitates collaboration between EGI and ELIXIR service developers and service providers. During its 18 month lifetime the CC collects, analyses and compares life science community requirements with EGI technical offerings, then designs and implements pilot e-infrastructure setups for the ELIXIR community based on EGI services. The whole process needs to be driven by scientific use cases that are selected from the ELIXIR and its partner communities. This document is the first milestone of the CC: the description of the scientific use cases that will drive CC activities, an initial analysis of the derived e-infrastructure requirements; and a technical roadmap to implement the science cases with the use of EGI services. The document also provides a roadmap for the integration of core EGI services into the ELIXIR Compute Platform, which is expected to underpin not only use cases covered by this report, but also future use cases of the ELIXIR community.

The EGI-Engage project is co-funded by the European Union (EU) Horizon 2020 program under Grant number 654142 http://go.egi.eu/eng

COPYRIGHT NOTICE

The EGI-Engage project is co-funded by the European Union Horizon 2020 programme under grant number 654142.

DELIVERY SLIP

	Name	*Partner/Activity*	*Date*
From:	Gergely Sipos	EGI.eu-SZTAKI/SA2	26.02.2016
Moderated by:	Małgorzata Krakowian	EGI.eu/NA1	
Reviewed by	D. Scardaci	INFN/JRA1	26.02.2016
Approved by:	AMB and PMB		3.03.2016

DOCUMENT LOG

Issue	*Date*	*Comment*	*Author/Partner*
v.1	12/Jan/2016	ToC with initial text for ELIXIR-CC	G. Sipos / EGI.eu-SZTAKI
v.2	31/Jan/2016	Text added about EGI developments for ECP	G. Sipos / EGI.eu-SZTAKI
v.3	04/Feb/2016	Integration of Marine and cBioPortal use cases; Update section about EGI AAI pilot	K. Mattila / CSC M. Ruda / CESNET G. Sipos / EGI.eu-SZTAKI
v.4	10/Feb/2016	Merge input from members and partners	C. Blanchet / CNRS O. Spjuth / PhenoMeNal
v.5	26/Feb/2016	Updates based on reviewers' feedback Add input from EMBL-EBI	G. Sipos / EGI.eu-SZTAKI S. Newhouse / EMBL-EBI
v.6	29/Feb/2016	JetStream interoperability use case added	R. Quick/IU
FINAL	29/Feb/2016	FINAL version after external review	G. Sipos / EGI.eu-SZTAKI

TERMINOLOGY

A complete project glossary is provided at the following page: http://www.egi.eu/about/glossary/

Contents

Executive summary

ELIXIR[1] is a pan-European research infrastructure in agreement between 17 European governments to build a sustainable European infrastructure for biological information, supporting life science research and its translation to medicine, agriculture, bioindustries and society.

EGI[2] is a pan-European e-infrastructure that delivers integrated computing services to European researchers, driving innovation and enabling new solutions to answer the big questions of tomorrow.

The ELIXIR Competence Centre (CC) of the EGI-Engage project evaluates, adopts and promotes technologies and resources from EGI to the wider ELIXIR research community. This report is the first milestone of this effort. The document (1) captures 4 scientific use cases that will be used by the CC to assess EGI services, (2) provides details about the e-infrastructure requirements of the use cases and (3) presents the use case implementation roadmaps considering the evolution of both EGI services and the ELIXIR Compute Platform currently emerging from the ELIXIR community.

The ELIXIR Compute Platform is a reference technical architecture – and its implementation within the ELIXIR-EXCELERATE project – to support a vast range of data analysis activities. EGI is currently contributing to the platform development with several services and technologies from EGI – all relating to the management and access of a cloud federation.

All of the five use cases in this report require cloud services, but in different ways, so they will be perfect test cases not only for the EGI services, but also for the ELIXIR Compute Platform. The main capabilities required by the use cases are:

1. cBioPortal replication: Hosting a portal environment in the cloud.
2. Marine metagenomics: Opening up an analysis platform for international user base via the cloud.
3. Insyght Comparative Genomics: Providing a scalable platform with 'one click deployment' capability on top of a federated cloud.
4. PhenoMeNal project: Offering a cloud federation for microservices developed and maintained by a project community.
5. JetStream project: Compatibility of the US-based JetStream cloud with the ELIXIR Compute Platform and with EGI Federated Cloud.

[1] http://www.elixir-europe.org/
[2] http://www.egi.eu/

1 Introduction

ELIXIR[3] is a pan-European research infrastructure in agreement between 17 European governments to build a sustainable European infrastructure for biological information, supporting life science research and its translation to medicine, agriculture, bioindustries and society.

EGI[4] is a pan-European e-infrastructure that delivers integrated computing services to European researchers, driving innovation and enabling new solutions to answer the big questions of tomorrow.

Life science is a fast moving field. For the EGI services to become relevant and help keep European Life Sciences competitive globally, it is important to develop mechanisms that allow the research infrastructure to flexibly meet new challenges and respond to new scientific and technical developments.

The ELIXIR Competence Centre (CC) of the EGI-Engage project evaluates, adopts and promotes technologies and resources from EGI to the wider ELIXIR research community. This is achieved with an iterative approach:

1. Bringing together designated life science experts from ELIXIR and technical experts from EGI within the CC.
2. Identify life science use cases which could benefit from EGI services and could make big impact on ELIXIR and EGI communities. Analyse the e-infrastructure requirements of the use cases.
3. Implement the use cases as demonstrators based on EGI e-infrastructure services. Collaborate during implementation with relevant EGI and ELIXIR partners, such as the EUDAT[5].
4. Demonstrate and evaluate the implementations. Disseminate the experiences gained with the use cases towards ELIXIR, EGI and other relevant communities. Decide about the long-term adoption of EGI services within ELIXIR based on the pilot experiences.

This document is a milestone after stage 2 of this process. The document was written by life science and e-infrastructure experts from ELIXIR and EGI, brought together within the CC. The document captures scientific use cases, derived requirements and envisaged implementation roadmap based on EGI services.

[3] http://www.elixir-europe.org/
[4] http://www.egi.eu/
[5] http://www.eudat.eu/

2 Scientific use cases

This section provides information about the use cases that have been identified by the Competence Centre. These use cases represent scientific workflows that can be ported to the EGI Federated Cloud. At this early stage in the establishment of the ELIXIR Computer Platform and its exploration of the EGI Federated Cloud, the search for use cases was restricted to workflows that require only 'non-sensitive' data, because this simplifies complexity, and also makes the ELIXIR Competence Centre effort complementary to the BBMRI Competence Centre activities (task SA6.4 of EGI-Engage), where the focus is on handling sensitive data with EGI services. Each of the use cases are described from three perspectives:

1. Scientific
2. E-infrastructure
3. Impact

These aspects together provide a comprehensive view on the use cases and help the Competence Centre focus its limited effort on those cases that would offer the best value vs. implementation and operational cost.

2.1 cBioPortal replication use case

2.1.1 Scientific use case description

The EurOPDX Consortium[6] is an initiative of translational and clinical researchers from 16 academic cancer centres and universities across 10 European countries, with the common goal of creating a network of clinically relevant models of human cancer, and in particular PDX models. The main objectives of the EurOPDX Consortium are to:

- create a virtual collection of genomically and histologically characterised PDXs;
- harmonise working practices; and
- leverage the collection to investigate novel therapeutic strategies and uncover predictive biomarkers for personalised cancer treatment, through the performance of more effective and reproducible multicentre PDX studies with high predictability for success in the clinic.

The Consortium is requiring possibility to provide a clone of their cBioPortal to serve its existing and future user communities. The cBioPortal for Cancer Genomics provides visualization, analysis and download of large-scale cancer genomics data sets[7]. The portal would be provided as a Docker container in the cloud to simplify installation, and to standardise integration across sites of the EGI Federated Cloud infrastructure. Initial request is relatively small in terms of capacity (1 standard node with 2 CPUs, 8+ cores, 128+ GB of RAM, 10-20TB disk space). The initial user-base is approx.

[6] http://europdx.eu/

[7] http://www.cbioportal.org/

15-20 scientists, but the portal can be relevant for many more in various typical life science use-cases.

2.1.2 Scientific use case description

Responsible person within the CC	Miroslav Ruda, CESNET
User Story	Hosting the cBioPortal in the EGI Federated Cloud: 1. A Docker image is prepared from the cBioPortal. 2. Basic installation is performed on one cloud site of the EGI Federated Cloud. 3. Installation, database and analysis software are fine-tuned by scientists together with site administrator. 4. Data from scientists are uploaded into the portal and made ready for analysis. 5. Testing whether the authentication mechanism of the portal can be integrated with ELIXIR AAI solutions.
(Potential) User base	EurOPDX community.

2.1.3 E-infrastructure requirements

HW Resources	1 standard node with 2 CPUs, 8+ cores, 128+ GB of RAM, 10-20TB on disk space.
SW Resources	Software is provided by users in Docker image form. It is a copy of the cBioPortal.
Cost of delivery	1-2 PMs for deployment support, 1PM for the analysis of AAI integration and for the integration.
Operational aspect	Site maintains hardware and basic operating system, Docker image is managed by the user-group. NGI_CZ is willing to provide resources and support.

2.1.4 Impact

Business plan 1	Site maintains hardware and basic operating system, Docker image is managed by the user-group. NGI_CZ is willing to provide resources and support.

2.2 Marine metagenomics use case

2.2.1 Introduction

The effectiveness of current sequencing technologies has made sequencing a commonly used tool in all the fields of biological sciences. In environmental biology, environmental samples are sequenced to provide metagenomics data: i.e. information about the taxonomical diversity and functional profile of microbial community found in the samples.

As the actual sequencing has become more effective the bottleneck in metagenomics has moved from generating data to managing and analysing the data. Many European research groups do not have enough computational power and storage space needed to fully utilize the metagenomics data. In addition to hardware, setting up and maintaining a metagenomics analysis environment requires expertise in both system level software components and the bioinformatics tools that are used to perform the analysis.

To overcome this situation tools, like EBI-Metagenomics, Metagenomics-Rapid Annotations using Subsystems technology MG-RAST and Integrated Microbial Genomes and Metagenomes (IMG/M) have been developed. However, these services do not fulfil the needs of all domains of metagenomics. In addition, many of these tools run on a server administered by a single organization and thus available for only a limited user community.

META-pipe, developed at the University of Tromsö, is an analysis pipeline that is designed to fulfil the needs of marine metagenomics data analysis. META-pipe integrates existing biological analysis frameworks, and compute and storage infrastructure resources to provide an easy to use but effective analysis platform. META-pipe is also an important component in one of the four scientific use cases in the ELIXIR-EXCELERATE H2020 project[8].

At the moment the pipeline is available only for Norwegian academic user. The use case proposes (1) integration of META-pipe, or its computationally demanding parts with the EGI Federated Cloud and (2) extending META-pipe with a new AAI layer based on security mechanisms that can be accessed using the ELIXIR AAI to allow controlled, shared access to the services. The use case would demonstrate in practice, how an ELIXIR use case can utilise EGI security and cloud services and resources.

2.2.2 Scientific use case description

Responsible person within the CC	Kimmo Mattila, CSC
User Story	Integrating the META-pipe analysis pipeline with cloud services from EGI: To expand the potential of metagenomics for the research community and biotech industry, especially within the marine domain, the metagenomics

[8] https://www.elixir-europe.org/news/elixir-accelerates-major-horizon-2020-funding

	methodologies need to overcome a number of challenges related to standardization, development of relevant databases and bioinformatics tools. The ELIXIR-EXCELERATE Work Package 6 Use Case “Marine metagenomics infrastructure as driver for research and industrial innovation” will develop a sustainable metagenomics infrastructure to enhance research and industrial innovation within the marine domain. This will be achieved by: • Development and implementation of selected standards for the marine domain • Development and implementation of databases specific to marine metagenomics • Evaluation and implementation of tools and pipelines for metagenomics analyses • Development of a search engine for interrogation of marine metagenomics datasets and • Establishment of training workshops for end users One of the primary tools to be used in this research case is the META-pipe analysis pipeline developed in the university of Tromsö. A copy of this pipeline will be set up to the EGI federated cloud with the possibility of controlled, shared access to members of the metagenomics community.
(Potential) User base	Marine research, including metagenomics, is carried out by hundreds of research institutes in Europe so the potential user community is large ranging from individual university researchers to projects such as TaraOcean[9] and Ocean Sampling Day[10]. As part of ELIXIR, the META-pipe service will be available for the whole ELIXIR community.

2.2.3 E-infrastructure requirements

HW Resources	In the use case we plan to set up a system, which has hardware resources similar to the current META-pipe server that is running at the University of Tromsö. Estimate of the needed resources: Computing requirements • 16 medium sized (8 core) virtual machines machines plus 4 small VMs for the cluster front end and other functions. In total 16*8 +(1+3)*4 = 144 computing cores. • RAM: 16*30.720 + 4*15.360 = ~553 GB Storage requirement: • Hadoop Distributed File System or object storage: At least 10 TB

[9] http://www.embl.de/tara-oceans/start/index.html

[10] https://www.microb3.eu/osd

SW Resources	META-pipe consists of several layers, each including several software components. The software resources needed depend on how META-pipe will be implemented in EGI Federated Cloud. Possibilities range from full installation including computing, storage and user interfaces, to a scenario where only computationally demanding parts are installed in EGI environment and linked to the current META-pipe. In any case, all the components used in the META-pipe are open source. **Most essential Scientific tools:** PreProsessing: - Prinseq - FastQC - Mira - MetaRay Taxonomic components: - rRNAselector - LCA-classifies + Silva DB - Kona Tools CDS Analysis: - Glimmer - MGA - Priam - BLAST+ UniprotKB - Interproscan5 + databases - Metarep **System level components:** - Apache Spark - Web interface - META-pipe specific job processing scripts
Cost of delivery	At this stage it is hard to give an estimate for the work needed to set up this system in EGI federated cloud. It will take at least 4-8 PM effort, but possibly more in case of porting of some of the SW components cause unexpected complexity.
Operational aspect	The use case described here aims to demonstrate and study the use of EGI services to operate META-pipe. The cost of operation will be assessed based on the experience and feedback of this demonstrator.

2.2.4 Impact

Business plan	A European marine metagenomics pipeline would enhance and make easier research and collaboration within the marine research community. A common analysis platform would reduce the need to setting up local analysis platforms and would guide researchers to use methodologies and data formats and would enable re-use of previously generated data. The use case will be setup as a demonstrator. The cost of operation will be assessed based on the experience and feedback of this demonstrator. If the use case would be converted to a production level service it would be operated by the ELIXIR community including at least ELIXIR-Norway, ELIXIR-Finland and EMBL-EBI.

2.3 Insyght Comparative Genomics

2.3.1 Introduction

High-throughput sequencing technologies produce an ever-increasing number of newly sequenced genomes. Faced with this huge mass of data, biologists need efficient and user-friendly tools to assist them in their analyses. In this context, tools that facilitate comparative genomics analyses (i.e. conservation of gene neighbourhood, presence/absence of orthologous genes, phylogenetic profiling, etc.) of large amounts of data are much needed. Insyght is a comparative genomic visualization tool[11] that tightly integrates three complementary views:

(i) a table for browsing among homologs

(ii) a comparator of orthologs' functional annotations and

(iii) a genomic organization view that combines symbolic and proportional graphical paradigms to improve the legibility of genomic rearrangements and distinctive loci.

Insyght benefits from an easy and smooth navigation between these 3 views and provides users with a powerful search mechanism.

Its underlying database contains the cross comparison of 2660 bacterial proteomes (Lacroix et al., 2015). It would be interesting to extend the Insyght database to all the representative bacterial genomes (as defined in RefSeq for instance) to better cover the microbial diversity.

Firstly, a virtual machine has been developed by the authors with the complete computing environment required by Insyght. Secondly, the VM was ported to the cloud in the context of the European Project CYCLONE (European Commission Horizon 2020 framework, grant number 644925).

2.3.2 Scientific use case description

Responsible persons	Christophe Blanchet, CNRS IFB

[11] Lacroix *et al.*, 2014

within the CC	Jean-françois Gibrat, INRA IFB
User Story	• As a bioinformatician, I have access to the required appliances, published in the marketplace, to help me comparing my bacterial data • As a bioinformatician, I can deploy and synchronize the required public reference data collections in the targeted clouds • As a bioinformatician, I can deploy a distributed cluster of VMs to compare microbial proteomes • As a bioinformatician, I can run a VM to visualize my genomic data • As a bioinformatician, I can deploy in one-click the complete environment (cluster of VMs for the computations and one VM for the visualization)
(Potential) User base	Any life-scientist who needs to perform bacterial comparative genomics analyses.
Cost of delivery	The estimate for the work needed to set up this system in EGI federated cloud is 6 PM effort.

2.3.3 E-infrastructure requirements

HW Resources	Updating the Insyght database requires N*(N-1)/2 bacterial proteome comparisons, where N is the number of bacterial species considered. In theory, this implies sending millions of jobs to the IT infrastructure, although the number of jobs can be tuned as desired by clustering bacterial proteomes. To ease the computations, they should be distributed to several Cloud facilities.
SW Resources	A virtual machine has been developed that integrates all the elements needed by Insyght: 1. A database. Data is stored in a PostgreSQL relational database. This database contains three types of data: (i) primary data such as genomic annotations extracted from genome files (obtained from EMBL-EBI's Ensembl Bacteria), (ii) secondary data that results from the cross comparison of the proteomes using BLASTp (Altschul *et al.*, 1997), and (iii) tertiary data such as the synteny regions. 2. A pipeline. The database is populated by a pipeline of Perl scripts that (i) process the genome files, (ii) run the BLASTp jobs on a cluster, (iii) parse the results, and (iv) execute the program that determines the syntenies between all the pairs of bacterial proteomes 3. A Web interface.

2.3.4 Impact

Business plan	The Insyght comparative genomic visualization tool will provide scientists with an efficient and user-friendly tool to assist them in comparative genomics analyses (i.e. conservation of gene neighbourhood, presence/absence of orthologous genes, phylogenetic profiling, etc.) of large amounts of data. The use case will be setup as a demonstrator. According to the results, the business plan will be assessed based on the experience and feedback of this demonstrator.

2.4 PhenoMeNal project use case

2.4.1 Introduction

The PhenoMeNal H2020 project[12] started in 2015 with the goal to setup an integrated, secure, permanent, on-demand service-driven, privacy-compliant and sustainable e-infrastructure for the processing, analysis and information-mining of the massive amount of medical molecular phenotyping and genotyping data that will be generated by metabolomics applications now entering research and clinic. The infrastructure will be one of the key enabling e- infrastructures addressing the H2020 Societal Challenge in Health, Demographic Change and Wellbeing.

PhenoMeNal WP5 (titled "Operations and Maintenance of PhenoMeNal GRID/Cloud") will provide the foundation upon where data and analysis services can be used together on computing resources. This foundation should comprise the hardware (compute and storage) as well as middleware for federating queries and resources between sites, enabling the functions in the Virtual Research Community (VRC) portal (the portal will be developed in WP6).

Of high importance is the documentation and packaging of infrastructure resources and configurations to allow for easy setup on partner systems, enabling a federated system. The use of containers and orchestrators is envisaged for this.

2.4.2 Scientific use case description

Responsible persons within the CC	Steven Newhouse, EMBL-EBI Enol Fernandez, EGI.eu
User Story	The use case has three, interlinked user stories: 1: Cloud infrastructure for Virtual Machine and container operation: • Define and implement a cloud infrastructure in collaboration with ELIXIR, that can be used to roll-out and operate virtualised and/or containerised

[12] http://phenomenal-h2020.eu

	PhenoMeNal applications and services from a Marketplace. This work will build on the ELIXIR Compute Platform that is based on the EGI Federated Cloud approach and services. (See Section 3.2 for further details) 2: Marketplace for PhenoMeNal VMs and containers: • Setup a dedicated group in the EGI AppDB marketplace[13] for the Phenomenal community • Connect the Identity Providers of the PhenoMeNal institutes into the Marketplace (if such IdPs exist. Otherwise use the ELIXIR AAI for authentication) • Delegate VM/container developers into the group (They can upload VMs into the group) • Delegate 1/more VM/container administrator into the group (They can endorse VMs on behalf of the project) • Delegate users into the group (They can download the VMs/containers) 3: Support for VM/container development: • Prepare guidelines for VM and container preparation and contextualisation. The guidelines would ensure that PhenoMeNal VMS are secure and compatible with the cloud platforms that support their e-infrastructure. EGI recently prepared guidelines in the form of a tutorial[14] about VM preparation. This can be adopted for PhenoMeNal. • The use of Docker is expected to ship applications to sites. EGI recently prepared guidelines about how to use Docker containers[15] in the EGI Federated Cloud. This guideline can be adopted by PhenoMeNal.
(Potential) User base	Researchers working with medical molecular phenotyping and genotyping data that are / will be generated by metabolomics applications now entering research and clinic

2.4.3 E-infrastructure requirements

HW Resources	The provisioning of a reference cloud infrastructure is envisaged by the PhenoMeNal project. PhenoMeNal VMs and containers should be usable on both the reference infrastructure and the ELIXIR/EGI infrastructure, so users can scale up or move their applications between these two.
SW Resources	The software environment of the PhenoMeNal project is under discussion/development. The project will organise a hands-on workshop between Feb 29 - Mar 2 where further details about this will be presented and decided.

[13] https://appdb.egi.eu/browse/cloud

[14] Dos and Don'ts for Virtual Appliance Preparation: https://indico.egi.eu/indico/event/2544/session/46/?slotId=0#20151110

[15] https://wiki.egi.eu/wiki/Federated_Cloud_user_support#Docker_containers

Cost of delivery	In the short term, the PhenoMeNal project is looking for a flexible infrastructure able to quickly scale in the cloud to cope with any input dataset the community will need to process. At the time of writing, as the development of the container platform is just at the beginning, it is complex to reliably predict the amount of computing power needed to analyze a given dataset, and thus provide figures in terms of VMs or allocated resources. For this reason, flexibility of the underlying systems is a key requirement that must be pursued since their infancy, keeping in mind that to fulfill the project final goal it may be required to scale the PhenoMeNal Gateway up to 100s of VMs in different geographical locations.
Operational aspect	See previous.

2.4.4 Impact

Business plan	See previous.

2.5 JetStream interoperability use case

2.5.1 Introduction

Jetstream, led by the Indiana University Pervasive Technology Institute (PTI), will add cloud-based computation to the US national cyberinfrastructure[16]. Researchers will be able to create virtual machines on the remote resource that look and feel like their lab workstation or home machine, but are able to harness thousands of times the computing power. Jetstream will provide the following core capabilities:

1. Use Virtual Machines interactively
2. Researchers and students can move data to and from Jetstream using Globus Transfer
3. Use virtual desktops.
4. Publish VMs with a DOI.

Jetstream will be attractive to communities who have not been users of traditional HPC systems, but who would benefit from advanced computational capabilities. Among those groups are researchers not only in biology, but also atmospheric science, observational astronomy, and the social sciences.

2.5.2 Use case description

Responsible persons within the CC	Robert Quick, Indiana University and Open Science Grid
User Story	• I am an ELIXIR collaborator with access to JetStream computing resources in the US. I would like to use the ELIXIR Compute Platform

[16] http://jetstream-cloud.org/index.php

	on these resources. • I am a bioinformatician that would like to use ELIXIR software to execute my workflow on EGI collaborating resources. • I am a cloud infrastructure provider I would like to provide cloud based resources to the ELIXIR project.
(Potential) User base	Any life-scientist who collaborates on ELIXIR based software and scope extends internationally.
Cost of delivery	The estimate for the work needed to set up this system to interoperate with the ELIXIR compute platform and EGI Federated Cloud is 2 PM effort.

2.5.3 E-infrastructure requirements

HW Resources	Existing EGI Federated Cloud and Jetstream cloud environment. Existing international networking.
SW Resources	No software requirements.
Cost of delivery	1 PM of interoperability testing of the EGI Federated Cloud and ELIXIR Compute Platform in the Jetstream cloud.
Operational aspect	Submission interoperability. Accounting through existing EGI-OSG accounting service.

2.5.4 Impact

Business plan	International interoperability for each collaborating project (EGI Federated Cloud, Jetstream, and ELIXIR) will increase the diversity of both compute resources and user base. Each project maintains the hardware (EGI Federated Cloud and Jetstream) and software resources (ELIXIR compute platform), effort will be in porting software and creating seamless workflow submission environments.

3 Implementation roadmap

3.1 Introduction

During 2015 the ELIXIR community – in collaboration with various e-infrastructures and other service providers – initiated the development of the reference architecture for ELIXIR, called the 'ELIXIR Compute Platform' (ECP). The prime role of the ECP is to support the use cases of the ELIXIR-EXCELERATE H2020 project, however, the platform is expected to serve other ELIXIR-related use cases from ELIXIR and other biomedical sciences Research Infrastructures. The previously described use cases will also interact with the ELIXIR ECP, and will operate as high-level services within the platform. The next subsections describe the architecture of the ECP (Section 3.2), the ongoing activities in EGI to customise and integrate EGI services into the ECP (Section 3.3), a data replication use case that is emerging within the EUDAT2020 project with consequences in the mid-term for the EGI services (Section 3.4) and the implementation roadmap of the previously described scientific use cases in relation to the ECP (Section 3.3-3.5).

3.2 The ELIXIR Compute Platform

This section is a summary of v0.9 of the living document that defines the ELIXIR Compute Platform and is available online[17].

The need for an ELIXIR reference technical architecture was first discussed during a BioMedBridges e-Infrastructure workshop in May 2014, where reference was made to the MONARC report[18] that formed the basis of the Tiered model that was initially adopted by WLCG community to serve the needs of High Energy Physics. Following on from work by the ELIXIR Authentication and Authorization Infrastructure (AAI), Storage and Cloud Task Forces to define a set of Technical Use Cases, a workshop was held in Amsterdam (12-13th March 2015) to discuss with representatives of ELIXIR nodes, European e-Infrastructures and other service providers, how the ELIXIR-EXCELERATE Scientific Use Cases could be mapped onto the Technical Use Cases and thereby define the ELIXIR Compute Platform. Through a series of presentations and breakouts the technical aspects of the Scientific Use Cases were identified and mapped to a number of Technical Use Cases. As a result of these discussions, a number of recommendations have been made for technical solutions that together will provide an ELIXIR Compute Platform. The platform can not only support the ELIXIR-EXCELERATE Scientific Use Cases, but a vast range of other data analysis activities that will be found within the ELIXIR research community. Such as:

[17] https://docs.google.com/document/d/1gMKFrcbzuN9BSREU1VDnlml-bl6KSOnfyQbJGh20L5s/edit

[18] It is worth noting that the MONARC report is now considered outdated as following the initial experience of running the WLCG and the advances in the capability of the international networks, an evolved technical architecture is now being established that is less hierarchical in its data flows.

- Hosting portals that enable users to select and launch virtual machines onto an available cloud resource (e.g. for training activities).
- Hosting web tools that deploy a network of virtual machine images onto distributed cloud resources operated for ELIXIR users for large scientific analysis.
- Provisioning 'Desktop as a Service' where researchers are able to obtain a desktop image (e.g. BioLinux) in a cloud that they can use for their data analysis activities that is always on for their use.

The role of ELIXIR and the ELIXIR-EXCELERATE proposal is not to undertake middleware development. Instead the focus is on leveraging the investment that has already been made in services that can be integrated for our needs and steer future development priorities. Essentially, our role is to define a minimal 'neck' of an hourglass that ELIXIR Researchers and Application Developers can build upon and that ELIXIR Nodes and other infrastructure service providers can deploy and support. The ECP is envisaged to consist of the following service groups:

- Basic Identity Environment: authentication and authorization related infrastructure ("AAI") to provide user identity and access management services[19] for 'ELIXIR infrastructure services' (all other services). The basic ELIXIR AAI environment is available since the end of 2015 and further developments and refinements are coming during 2016.
- Core Enabling Infrastructure Services: provide capabilities to store and effectively transfer data (storage management and file transfer services). ELIXIR and EUDAT are working together in the EUDAT2020 project to identify, test and deploy services for this area. (See Section 3.4 for more information)
- Basic Infrastructure Services: Cloud IaaS, Cloud Storage or HTC/HPC Cluster resource may be operated from within the ELIXIR community. ELIXIR is working with EGI to implement this service area using technologies and know-how from the EGI Federated Cloud solution[20].
- Integrating Infrastructure Services: providing a federating structure that ensures a consistency of operation and behaviour across all resources and services of the ECP. ELIXIR and EGI are working together to implement this service area using technologies and know-how from the EGI Federated Operations solution[21].
- Higher-Level Services: solutions that expand the platform to better serve specific use cases or use case categories. Competition among similar solutions is expected in this area. ELIXIR is working with EGI to bring in solutions, primarily in connection to the Federated Cloud, into this area (for example Virtual Machine Marketplace)

[19] ELIXIR AAI – Requirements and Design: https://docs.google.com/document/d/1CMY1np3GyvPD8LcKvXljXcRO04V2zu3n_Jcg19jgNOw/edit
[20] https://www.egi.eu/solutions/fed-cloud/index.html
[21] https://www.egi.eu/solutions/fed-ops/index.html

3.3 EGI services in the ELIXIR Compute Platform

This subsection provides further details about the development activities that are required and are already ongoing in EGI to customise and integrate services into the ECP. The development work was triggered by the previously mentioned 'The ELIXIR Compute Platform: A Reference Technical Services Architecture for supporting Life Science Research' document (v0.9, May 18, 2015), and by additional recommendations that recently emerged from ECP-related development activities (e.g. the setup of ELIXIR AAI at the end of 2015).

As it was mentioned in the previous section, EGI is expected to contribute to the ECP with

1. Federated Cloud services (in Basic Infrastructure Services area)
2. Operational tools (in Integrating Infrastructure Services and Basic Identity Environment areas)
3. Virtual Machine marketplace and other optional services (in Higher-Level Services area)

The deployment of 'Federated Cloud services' has started at EMBL-EBI in late 2015. The goal is to integrate the EMBL-EBI Openstack Kilo site into the EGI Federated Cloud, using the installation documentations[22] currently available from EGI. The first feedback from this experience was reported to EGI in the second half of February. The feedback pointed out weaknesses in the installation manual and in some of the software components. The relevant EGI teams are now implementing improvements in the manual and the software based on this feedback. New versions will be available by the end of March. EMBL-EBI and other members of the CC will perform a reassessment of the technology in April and will document the experiences in the D6.10 deliverable, titled 'Infrastructure tests and best usage practices for life science service providers' by the end of May.

The EGI contributions to the ECP must be compatible with the 'Basic Identity Environment' (aka. ELIXIR AAI) to allow seamless support of users and use cases in the platform. There are two key aspects of the ELIXIR AAI that EGI must consider during service integration:

1. The ELIXIR AAI is implemented as an ELIXIR Identity Provider (IdP) at CESNET. Every ELIXIR user will have a user account and user attributes at this IdP. ELIXIR users will be assigned with roles in this system. (For example 'cloud site manager', 'VM manager', 'admin of central services', etc.) ECP services – including those contributed by EGI – must be able to authenticate and authorise users by their ELIXIR accounts and role attributes.
2. The ELIXIR IdP provides only the hashed eppn for SPs that are not committing to the GEANT Code of Conduct (CoCo)[23]. The ELIXIR IDP provides the following set of attributes for Service Providers that are committing to the CoCo:
    a. eduPersonPrincipalName - ELIXIR ID
    b. displayName

[22] https://wiki.egi.eu/wiki/Federated_Cloud_resource_providers_support#Join_as_a_Resource_Provider
[23] https://wiki.edugain.org/Data_Protection_Code_of_Conduct_Cookbook

c. email

The following table provides details about the development activities that are ongoing in EGI to customise and integrated EGI services into the ECP. The table is an updated version of the 'ELIXIR Compute Platform Timeline of EGI developments' document[24].

EGI service	Service area in ECP	Requirement	Priority	Status of development
Federated Cloud site deployment	Basic Infrastructure Services	Improve installation guide and certain software components to simplify deployment, especially at sites that are not yet part of the EGI federation.	High	Improvements are under development and will be released by the end of March. Assessment of these in the CC will be in April, with final conclusions to be documented in D6.10 by the end of May.
VO Membership Management tool (PERUN)	Basic Identity Environment	ELIXIR members should be able to join the ELIXIR VO with their ELIXIR ID.	High	Completed and deployment is in place for the vo.elixir-europe.eu VO.
Applications Database Virtual Machine Marketplace (AppDB)	High-Level Services	ELIXIR users should be able to login to AppDB with ELIXIR IDs. AppDB should recognise ELIXIR VO managers and allow them to add control the Virtual Machine image list that's associated to the VO.	High	Direct integration with the ELIXIR IdP is completed and deployed in the AppDB development instance[25]. The setup can be reassessed in 2016 Q2, when the EGI AAI proxy[26] will become available for early adopters.
Operations Portal	Integrating Infrastructure Services	ELIXIR cloud providers and VO admins should be able to login to the Operations Portal with their ELIXIR identity to send service	High	Need to investigate and decide about integration approach: direct integration with the ELIXIR AAI or

[24] https://docs.google.com/document/d/1J3XPAvX0jVhJ_pFex5gXWYBazRweKNKbqRfTUJF56M0/edit#heading=h.idg60lt9pvvi

[25] https://appdb-dev.marie.hellasgrid.gr

[26] Se section 3.2.2 for further details.

		downtime broadcast messages to members of the ELIXIR VO.		through the EGI AAI pilot.
Service monitoring (ARGO)	Integrating Infrastructure Services	Create a site listing feature under 'Site status reports'. (Currently it's only a search function so the user must know what to look for.)	Medium	Requires only service configuration. This will be done after there is at least one ELIXIR cloud site in the ELIXIR VO.
		Add the first set of ELIXIR sites to the site list. Prime candidates are: • Members of ELIXIR-EXCELERATE WP4 (EMBL-EBI, CSC, CESNET, SURFsara). • Other cloud providers from the CC: GRNET, CNRS, ,		
		In the Availabilities/Reliabilities menu introduce a new subcategory: ELIXIR REPORT[27]. The ELIXIR sites should be listed in this subcategory.		
Service registry (GOCDB)	Integrating Infrastructure Services	ELIXIR cloud providers should be able to login to GOCDB to register/update information about their site.	Low	Direct integration with the ELIXIR IdP is completed and deployed in the GOCDB development instance.
Accounting system (APEL)	Integrating Infrastructure Services	The EXCELERATE project will define the ELIXIR Metrics database and portal. APEL should gather and send accounting data from the ELIXIR sites to this ELIXIR metrics database.	No	Development can start approx. in spring 2016, when the ELIXIR metrics database and portal are specified.

[27] ELIXIR would like to have its own identity on the portal even if it reuses sites from EGI and EUDAT at the infrastructure level.

3.3.1 New EGI AAI Pilot

Several European Research Infrastructures recently decided to operate their own AAI (Authentication and Authorisation Infrastructure) to provide user identity and user attributes for their community members. These RI-specific AAIs simplify and harmonise access to online services across institutional and national borders. ELIXIR recently also established its own AAI, hosted at CESNET from late 2015.

The trend of dedicated RI AAIs made EGI reassess its own AAI architecture. The community concluded that it needs to evolve its own AAI architecture to allow coordinated linkage of EGI services with externally operated RI AAIs. The EGI community – within the JRA1.1 task[28] of the EGI-Engage project – started the design and development of the new EGI AAI in March 2015. The work aimed at a pilot system that would

1. Simplify the process of connecting EGI services (e.g. AppDB, Operations Portal, GOCDB, etc.) with AAI architectures operated by external infrastructures, such as the ELIXIR.

AND

2. Harmonise the integration of EGI services across multiple, externally operated RI AAIs. (e.g. AppDB would be connected to the ELIXIR AAI, the DARIAH AAI, the EPOS AAI in a harmonised way).

The design of this new EGI AAI pilot system has finished in 2015 in close collaboration with the AARC H2020 project[29]. In the heart of the pilot system there is an 'IdP/SP Proxy' component, which is based on SAML technology (See Figure 1). This component acts as a Service Provider (SP) for the supported identity federation (e.g. the ELIXIR IdP), while at the same time, it will act as an Identity Provider (IdP) for the EGI services (e.g. AppDB, Operations Portal, GOCDB, etc.). The IdP/SP Proxy will be responsible for mapping an external user identity to an 'EGI identifier' which will be used for the same user across all the EGI services. The IdP/SP Proxy will be able to import attributes from external attribute authorities (e.g. from ELIXIR IdP) and assign these to the internal EGI user identifier. Based on the imported attributes the EGI services can authorise users across the whole EGI network in a coherent way. (e.g. an ELIXIR site manager will be recognised in both GOCDB and Operations Portal).

[28] https://wiki.egi.eu/wiki/EGI-Engage:WP3#TASK_JRA1.1_Authentication_and_Authorisation_Infrastructure
[29] https://aarc-project.eu/

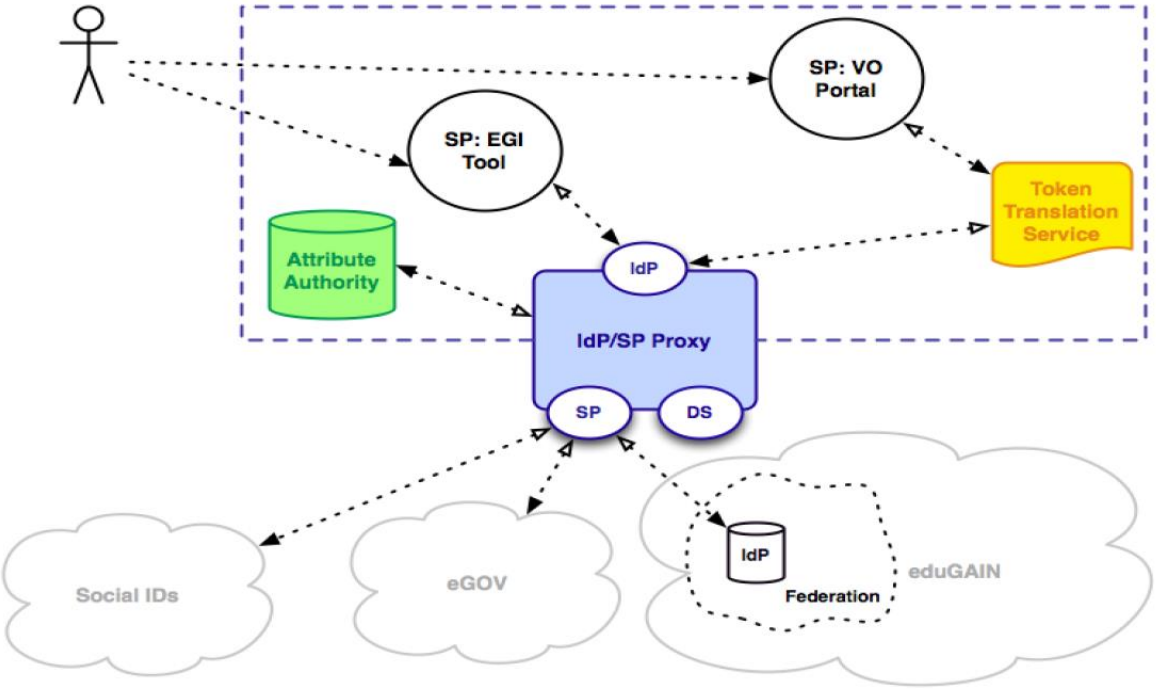

Figure 1. Architecture of the EGI AAI pilot

The new AAI pilot will be available for early adopter use cases by the end of Q1 2016. During Q2 the JRA1.1 activity will work with early adopter use cases to test the system with real use cases. The ELIXIR AAI would be a perfect early adopter because of the need of ELIXIR to interact with multiple EGI services in a coherent and consistent way. The involvement of ELIXIR in the early adopter programme will be discussed in the CC in March.

Integration of ELIXIR AAI with EGI through the EGI AAI proxy would lower the cost of maintenance and further development of the collaborative ELIXIR – EGI setup in the long term. Besides, the work could speed up the integration of those EGI services with the ELIXIR AAI that has not achieved this connection yet (for example the EGI Operations Portal).

3.4 Strategic data distribution and computing use case on the ECP

ELIXIR and EUDAT are currently working together in the EUDAT2020 project to establish a data distribution service. The primary use case of this service would be to replicate large and frequently used datasets from EMBL-EBI premises to strategic partner sites across Europe (See dashed arrows in Figure 2). This use case could decrease the data egress from EMBL-EBI by redirecting downloads to one of the partner sites for either individual files or complete data sets.

In a second phase the use case would be extended with cloud capabilities, by coupling the setup with the 'Basic Infrastructure Services' of ECP to enable rapid setup of analysis on the distributed datasets. This complete scenario – a joint use case of ELIXIR, EUDAT and EGI – would look like this (See also Figure 2):

- Data Set Owner (provider)
  - Create the data set on EMBL-EBI resources and provide meta-data to promote discovery
  - Create a data set placeholder for release
  - Add files/directories to the data set
  - Release the data set (no more files can be added to this version)
- Strategic partner sites:
  - Subscribe to a dataset (automatic distribution or just notified when a new data set is available)
  - Define where the data should be placed on your site
  - Identify who should be notified once the transfer has completed
- Researchers:
  - Notified when a new data set is available on a particular site (OR) Discover the availability of particular data set versions within the infrastructure on a site
  - Discover the availability of an application (a Virtual Machine, VM)
  - Access selected cloud resource and launch application
  - 'Mount' the local data set copy and run analysis

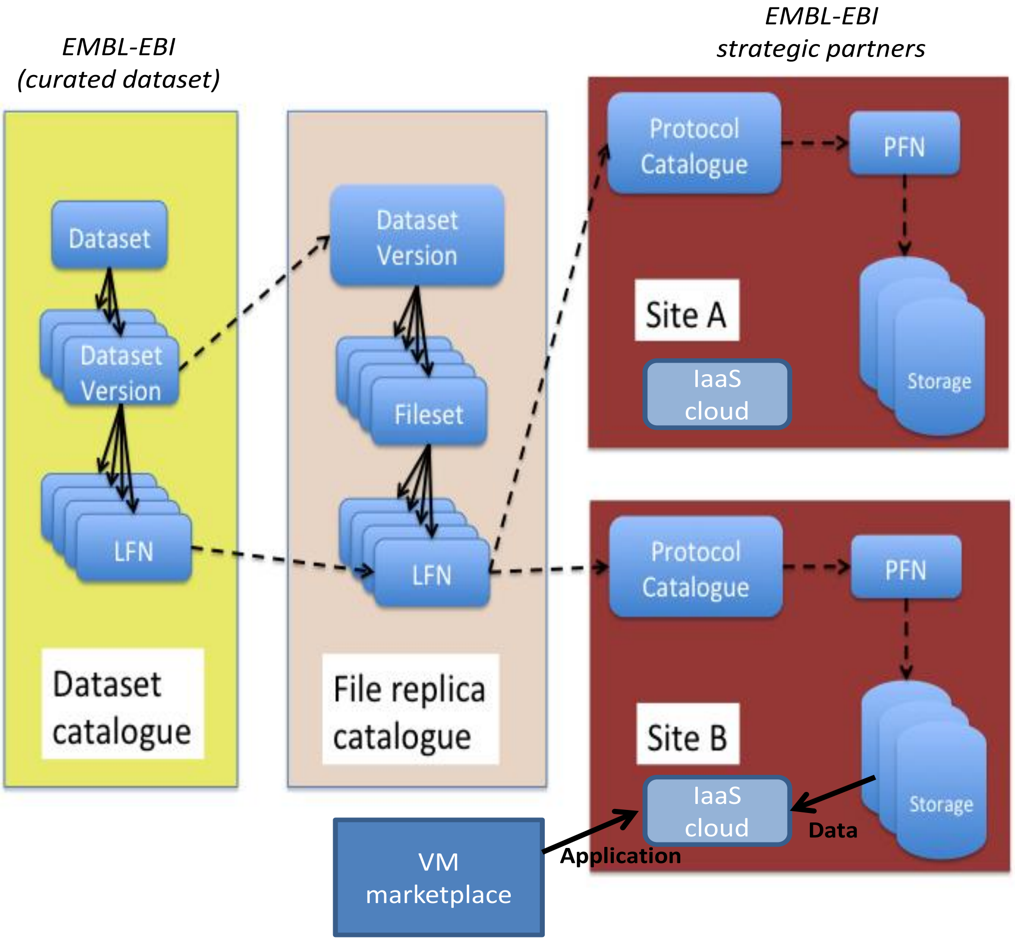

Figure 2. Strategic data distribution and computing

3.5 cBioPortal replication use case

The first version of the Docker-ised cBioPortal image will be provided by the EurOPDX community in March 2016. The initial deployment should be completed in one month, using the NGI_CZ site of the EGI Federated Cloud. Fine tuning of the installation will take probably several months. The status of the activity will be reviewed at future ELIXIR-CC teleconference meetings. Public report will be provided during the EGI Forum in autumn 2016.

3.6 Marine metagenomics use case

In Marine metagenomics use-case, the main target is to study, how the META-pipe analysis pipeline can be installed in a cloud environment. The installation can be achieved in different ways. These different options will be studied to find the best practices to run and access the service. Tests will be performed both on the OpenStack cloud environment of CSC and the Federated Cloud environment of EGI.

The META-pipe installation process includes:

- Setting up a medium sized virtual Linux cluster (approximately 20 nodes).
- Setup of object storage for the data. (There is about 10TB of data initially but the storage needs will increase in the future.)

One of the challenges is that the data structure and the access patterns are not yet fully known, so experimentation with multiple formats is expected. HDFS is strongest technology candidate here but we may also be able to use Swift from OpenStack. Ability to use Apache spark environment needs be included to the virtual cluster too.

Once the computing and storage platforms are available, the application programs and the work flow scripts used by META-pipe will be installed. Some of these have complex dependencies. The installation process should be automatized so that it can be easily reproduced. There are several possibilities for automatization, for example Virtual Machine libraries, Docker containers, Ansible playbooks.

For the end users the system should also run the META-pipe web interface, or be linked to the current META-pipe web interface. The user interface will use the ELIXIR AAI service for user authentication and thus the chosen AAI method on EGI should be compatible with this. This compatibility is expected to be achieved by the ELIXIR Compute Platform, which will include cloud resources for ELIXIR based on the technologies from the EGI Federated Cloud.

3.7 Insyght Comparative Genomics use case

A virtual machine has been developed with the complete computing environment required by Insyght. This VM has been tested on the IFB-core cloud site and on the CYCLONE project testbed infrastructure.

In the CC-ELIXIR we will

1. Evaluate the interoperability of the IFB cloud appliances between the IFB-core cloud site and the EGI Federated Cloud and the ELIXIR Compute Platform.
2. Evaluate the features available for the deployment of a complex application (many virtual machines of different flavours) on the EGI Federated cloud (with the use of several sites).
3. Evaluate the scalability of the application with representative scenarios and datasets, for example by generating a database containing the cross comparison of 2660 bacterial proteomes.

3.8 PhenoMeNal project use case

The PhenoMeNal use case is very similar to the work that is ongoing to establish the ELIXIR Compute Platform:

1. Establish a cloud federation using technologies and possibly resources from the EGI Federated Cloud. Compatibility with the PhenoMeNal ‘micro-services’ architecture is of key importance here.

2. Configuring the EGI AppDB marketplace to enable distribution of endorsed PhenoMenal applications to the sites of the community federated cloud. Support for containers is of key importance here.
3. Support the community in the development of new virtualised applications (VMs, containers) for new use cases.

The setup of the ELIXIR Compute Platform is ongoing. The first cloud site (OpenStack site from EMBL-EBI) is expected to become certified in the EGI cloud federation during Q2 2016 – reaching a baseline implementation of point #1 above. The PhenoMeNal project is working on its 'reference architecture', based on cloud computing environments and microservices.

PhenoMeNal will organise a workshop at the end of Feb (29/2-02/03) where the microservice architecture will be presented and discussed with external infrastructure providers. EGI and ELIXIR are invited to the event and will analyse further the integration/compatibility of the EGI – ELIXIR – PhenoMeNal architectures. The need for expanding support of orchestrators (e.g. Terraform, Ansible) and containers (Docker) are expected to come up as technical requirements for the EGI – ECP developments.

3.9 Role of CC members

	cBioPortal replication use case	**Marine metagenomics use case**	**Insyght Comparative Genomics use case**	**PhenoMeNal project use case**		**ELIXIR Compute Platform developments**
CSC		• Use case owner • Test on CSC OpenStack site • Test on EGI FedCloud				
CESNET	• Use case owner • Test on EGI FedCloud					Liaison for ELIXIR AAI
EMBL-EBI				• Use case owner • Liaison (ECP compatibility)		• Coord. of devel. • Site integrator
CNRS IFB-core			• Use case owner • Test on EGI FedCloud			

SURFsara	To be discussed					
GRNET	Cloud resource provider					
University of Indiana (Open Science Grid)					Liaison for use case Intergration of ECP into Jetstream environment.	
EGI.eu				Liaison (EGI FedCloud compatibility)		Integration of EGI tools into ECP