EXCELERATE Deliverable D4.4

Project Title:	ELIXIR-EXCELERATE: Fast-track ELIXIR implementation and drive early user exploitation across the life sciences	
Project Acronym:	ELIXIR-EXCELERATE	
Grant agreement no.:	676559	
	H2020-INFRADEV-2014-2015/H2020-INFRADEV-1-2015-1	
Deliverable title:	ELIXIR Technical Services document	
WP No.	4	
Lead Beneficiary:	20 - CSC	
WP Title	Technical Services	
Contractual delivery date:	31st August 2019	
Actual delivery date:	29 August 2019	
WP leader:	Tommi Nyrönen (FI) and Ludek Matyska (CZ)	35 - MU & 20 - CSC
Partner(s) contributing to this deliverable:	1 - EMBL, 6 - NBIC, 6.3 - RUG, 6.4 - SARA, 8 - CRG, 12 - BSC, 17 - INES-ID, 23 - UIT, 25 - SIB, 26 - CNRS, 29 - IP, 35 - MU, 36 - CESNET, 38 - DTU, 21 - ATHENA RIC, 45 - SU	

Authors and Contributors:

Christophe Blanchet (IFB-CNRS), **Mikael Borg (NBIS - SE)**, Jinny Chien (EMBL-EBI), Laia Codo (BSC - ES), Jose Ll. Gelpi (BSC-ES), Montserrat Gonzalez (EMBL-EBI), **Jarno Laitinen (CSC - FI)**, Ilkka Lappalainen (CSC - FI), **Mikael Linden (CSC - FI)**, **Ludek Matyska (MU - CZ)**, **Steven Newhouse (EMBL-EBI)**, **Tommi Nyrönen (CSC - FI)**, **Michael Prochazka (MU - CZ)**, **Mirek Ruda (CESNET - CZ)**, Harri Salminen (CSC - FI), Christine Staiger (DTL - NL), **Jonathan Tedds (ELIXIR-Hub)**, Juha Törnroos (CSC - FI), **Susheel Varma (EMBL-EBI)**, Tony Wildish (EMBL-EBI)

Reviewers:

ELIXIR-EXCELERATE WP Leaders

Table of contents

1. Executive Summary

This version of the ELIXIR Technical Services Roadmap (ELIXIR-EXCELERATE deliverable D4.4) has been written at the end of PY4 of the ELIXIR-EXCELERATE project (August 2019) and represents the final update of the roadmap within the project. Previous versions of the roadmap document can be used to track the evolution of the ELIXIR Compute Platform (ECP) over the last 4 years. This version of the roadmap focuses on the overall status of the ECP at the end of the ELIXIR-EXCELERATE project and the plans put in place to sustain the work of the ECP in the next phase of the ELIXIR Science Programme - 2019-2023. The ELIXIR Technical Services Roadmap focuses on the requirements of four ELIXIR communities (i.e. marine metagenomics, plants, rare diseases and human data) and training. As these communities also evolve quickly, the roadmap has been established as a living document that provides advice as to the current and future implementation activities and is subject to change between versions and is publicly accessible and commentable.

The ELIXIR AAI[1] has grown to be a full production Infrastructure Service with currently 63 production services and 63 services in the testing environment. ELIXIR AAI is stable and provides a way to federate and integrate scientific services. ELIXIR AAI provides user access management across organisational and national borders. By using different sources of identity users are structured into groups to support authorisation decisions by individual services. Use of the ELIXIR AAI has grown beyond internal ELIXIR services to include public ELIXIR services such as the ELIXIR Germany de.NBI cloud[2], Helix Nebula Science Cloud (to access commercial cloud providers), EUDAT's B2ACCESS service (to access other EUDAT services) and the EMBL-EBI Unified Submissions Interface.

Discussions around the planned Life-Science Identity across the BMS RI community have been consolidated within the EOSC-Life proposal which started in March 2019. The ELIXIR collaboration with the Global Alliance for Genomics in Health (GA4GH, ga4gh.org) has now expanded into a broader strategic partnership. The ECP co-leads GA4GH Data Use and Researcher Identity workstream from 2018 on in order to establish global standards in AAI to improve international interoperability. ECP also coordinates the ELIXIR Cloud and AAI Driver Project[3] alongside ELIXIR Beacons which utilises ELIXIR AAI identity management.

Moving files (i.e. data) between sites is a key capability for the ECP. The Reference Data Set Distribution Service that is being developed in collaboration with the EUDAT2020 project was introduced into the ECP during PY3 with the support of an ELIXIR IS. Furthermore, new protocols such as FTS3 and htsget, a genome data specific standard, have been integrated into the platform. The ECP deployed GA4GH standards for genomic

[1] https://f1000research.com/articles/7-1199/v1
[2] http://de.nbi
[3] https://elixir-europe.github.io/cloud/

data transfers in May 2018 and demonstrated transfer sensitive of human datasets to remote secure cloud infrastructure services.

Work within the ECP to integrate the cloud resources affiliated with the ELIXIR Nodes continues. The EGI Federated Cloud model has to date not been widely adopted within the ECP but remains under evaluation by the ECP, especially within the context of the European Open Science Cloud (EOSC) projects EOSC-Hub and EOSC-Life, where it may be used as the federation model. The ECP is engaging with EOSC through an ELIXIR Competency Centre that is funded as part of the EOSC-Hub project which builds upon the work that has been undertaken in the EGI-Engage project. ECP leads are steering the work packages in EOSC-Life for cloud and access control targeted to the community of the Biomedical Science ESFRIs. Experience gathered during ELIXIR-EXCELERATE in 2015-2019 suggests that the nodes, Biomedical Science ESFRIs and e-Infrastructures could be willing to form a federation of life science user identities and update their internal services policies to allow trans-national European access.

Technical discussions with the ELIXIR-EXCELERATE Use Cases (the four scientific use cases and the training activities) continued through PY4. Marine Metagenomics, the ECP focus during PY1 continued to develop and consolidate its technical activity. The ECP focus in PY2 on Human Data showcased two key functionalities: secure transfer of sensitive human data and making datasets originating from the European Genome-phenome Archive available on a secure cloud. The demonstrator leveraged ECP technologies and GA4GH standards (htsget) for transferring the data, and relied on ELIXIR AAI for user identification and dataset access control. In PY3, the human data community use case continues to be a key scientific driver for technology integration. A new use case relating to plants was added to the focus in January 2018, and by August 2018 the technical experts from WP4 and WP7 had formulated a technology demonstrator focusing on transfer of distributed datasets. The solution delivered through EXCELERATE project effort was reported in the ECP milestone M4.3[4]. The PY4 focus has been on the rare diseases community, where a combination of data transfer and workflow execution support has been successfully demonstrated during the ELIXIR All Hands Meeting in June 2019 in Lisbon as the ECP milestone M4.4[5].

To support the reader in reading this report a comprehensive glossary of technical terms is provided in Appendix A.

2. Impact

The key project results of the work within WP4 are:

- The establishment of the Common ELIXIR Authentication and Authorisation Infrastructure (AAI) service toolkit in 2018-2019 which is the basis of the European Life Science AAI in the EOSC-Life project. By August 2019 the ELIXIR AAI had

[4] https://drive.google.com/file/d/11vMhgJzdAIzptww4Rj6Zb5kxVmDptkpE/view
[5] https://drive.google.com/file/d/149_pxAEF_A_Z25Z3xSFhakcMTFhagCoS/view

enabled 2923 users organised in 503 groups from 721 institutions to use 126 relying scientific services.

- Providing visibility of the considerable cloud & compute e-Infrastructure services for Life Science that are available through the ELIXIR Nodes. In June 2019 this comprised 80,000 compute cores; 50,000 TB storage; 3,100 users. These services are fully utilised, and typically access can be gained by complying with national or organisational access policy, specific international collaboration project, or payment.
- Exploring a number of data transfer and data distribution mechanisms including established data transfer protocols (e.g. GridFTP, FTS) and data distribution and replication services (e.g. Reference Data Set Distribution Service[6] or B2* services)
- Public description of the platform roadmap[7] has been distributed extensively, and should be updated once again during 2019.
- Using ELIXIR Implementation Studies to integrate new user requirements into the ELIXIR Compute Platform. These projects have included supporting the development of the ELIXIR AAI, deployment and use of storage data transfer services, and a standards based platform for task based workflows. Many of these have been undertaken in collaboration with other ELIXIR platforms and provide a foundation for sustaining the ELIXIR-EXCELERATE deliverables in the context of ELIXIR's core funding as well as in other H2020 projects

Collaboration within the ELIXIR-EXCELERATE project:

- WP6 marine metagenomics Metapipe service demonstrator linked user portal Tools (WP 1 & 2) to the ELIXIR Compute Platform (WP4), integrating ELIXIR AAI in 2016. The service has been deployed in ELIXIR nodes (FI, CZ), supported outside the original node (NO), and made available for scientists in the international scientific community.
- WP7 demonstrator to support the plants community which coordinated transfer of distributed non-sensitive user datasets to ELIXIR Compute and was delivered in 2018.
- WP8 demonstrator with Rare Diseases integrated a container workflow orchestration over distributed ELIXIR Compute sites and was delivered in 2019.
- WP9 ELIXIR Compute and Human Data platform demonstrated technology integration across European nodes to transfer sensitive human datasets to the remote secure cloud infrastructure service in 2017, and work has evolved since through, e.g. the ELIXIR - GA4GH strategic partnership.

  Additionally, support for bioinformatics with WP11 Training has been tested using ELIXIR clouds with the additional funding support from an ELIXIR Implementation Study in 2018-2019.

Collaboration with European e-Infrastructures projects:

- The ELIXIR Compute Platform partnered with the AARC2 project and contributed to the general AARC2 AAI blueprint. ELIXIR AAI endorses and is compatible with the blueprint architecture. ELIXIR has been a major contributor to the

[6] https://github.com/EMBL-EBI-TSI/RDSDS

[7] http://drive.google.com/file/d/0B0KXZdVao0kqUE9BbXVrc3ZLY1E/view

development of a Life Science AAI requirements (developed in CORBEL, GA No 654248) and made a plan to deploy the key components with the European e-infrastructures in EOSC-Life project, again fully compliant with the AARC2 AAI blueprint.

- The ELIXIR Compute Platform participation in European Open Science Cloud (EOSC) activities has been established with the use of public sector cloud resources through EGI and commercial cloud resources through the HelixNebula Science Cloud, and generic EU data services through EUDAT. Application oriented work that was supported in the now completed EOSCpilot project continues in the life-science focused EOSC-Life project (with ELIXIR Compute Platform leadership in the AAI (WP5) and Cloud (WP7) activites) and is complemented by ongoing infrastructure related engagement in EOSC-Hub.

3. Project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following objectives:

No.	Objective	Yes	No
1	Establish the technical infrastructure across ELIXIR to enable effective data deposition, access, exchange and compute	x	

4. Delivery and schedule

The delivery is delayed: Yes • No☑

5. Adjustments made

N/A

6. Background information

Background information on this WP as originally indicated in the description of action (DoA) is included here for reference.

Work package number	4	Start date or starting event:	Month 1

Work package title	**Technical Services**
Lead	Tommi Nyrönen (FI) and Ludek Matyska (CZ)

Participant number and person months per participant

1 – EMBL 34.00, 6 - NBIC 2.00 (LTPs: SARA 13.00, RUG 2.00) 8 - CRG 8.5, 12 - BSC 7.00, 17 - INESC-ID 6.00, 20 - CSC 48.00, 23 - UiT 2.00, 25 - SIB 2.00, 26 - CNRS 12.00, 29 - IP 6.00, 35 - MU 30.00, 36 - CESNET 24.00, 38 - DTU 6.00, 41 - ATHENA RIC 8.00, 45 - UU 0.00 (LTP: SU 10.10).

WP4 - Technical Services [Months: 1-48]

MU, EMBL, NBIC, CRG, BSC, INESC-ID, CSC, UiT, SIB, CNRS, IP, CESNET, DTU, ATHENA RIC, UU

The role of the ELIXIR-EXCELERATE Technical Services WP is the practical integration of existing TechnicalServicesavailable for ELIXIR in the Nodes and e-Infrastructure by testing and contributing to documentation and integration with small-scale programming and scripting where needed. Development is managed outside the WP. As a result of the tasks described below, WP4 will provide a generic integrated platform that can be tailored further for the ELIXIR-EXCELERATE scientific Use Cases (WP6 to 9), Training activities (WP11), and other ELIXIR pilots and projects to meet their specific needs. This includes user support, documentation and guidance to enable and promote technology adoption.

Work Package uses a mechanism of renewal of focus with the ELIXIR Heads of Nodes committee as necessary. If scientific needs change or disruptive technologies emerge that change the technical objectives heads of Nodes committee supports linking of the changed landscape of technical services implementation with the other Work Packages (e.g. ELIXIR resource governance, training, data resources, service registry). Involvement of ELIXIR heads of Nodes is used for securing physical information technology resources from Nodes, and making experts available for collaborative work.

Objectives

The Technical services Work Package (WP) links the ELIXIR scientific programme 2014-2018 to the day-to-day technical service work in the distributed Nodes. The research platform for life science will be achieved through the

following objectives:

- Develop a sustainable and supported research platform for implementing geographically and organisationally distributed Cloud, Compute, Storage and Authentication and Access infrastructure services collected in the ELIXIR registries.
- Manage external technical dependencies with e-Infrastructures and Nodes with ELIXIR technical coordinator group for services delivered as a priority for the

ELIXIR-EXCELERATE Use Cases.

• Close collaboration with translational, bio-banking and imaging infrastructures at both the European and national level to ascertain that there are effective services to securely access and exchange data.

Work Package Leads: Tommi Nyrönen (FI) and Ludek Matyska (CZ)

Description of work and role of partners

Task 4.1: Leadership (52PM)

Subtask 4.1.1: Management and Coordination (25PM)

This task is responsible for coordinating technical work in the ELIXIR-EXCELERATE project and wider ELIXIR research infrastructure with WP12 building on the emerging community of technical experts in the ELIXIR task forces. In addition, the task establishes appropriate management and technical interfaces into the services and organisations the technical activities are dependent upon.

Partners: FI, CZ

Subtask 4.1.2: Provide a gateway to use European e-Infrastructure services for ELIXIR (GÉANT, EGI, EUDAT, PRACE) (13PM)

Regular requirements gathering from the Use Cases in WP6 to 9 and elsewhere in the ELIXIR community will define biological information service requirements and identify areas and activities that could be sourced by the European e-Infrastructures. Any planned service integration into the ELIXIR Technical Services will be identified in the regular Roadmap documents that will define a technical architecture and technology insertion roadmap. This should include defining the relevant 'account managers' in each public sector e-Infrastructure.

Partners: FI, EMBL-EBI

Subtask 4.1.3: ELIXIR technical community building and knowledge exchange (14PM)

Task grows the community of ELIXIR branded resource providers and sustains community of ELIXIR technical experts (i.e. ELIXIR technical coordinators and Node personnel) through engagement in major e-Infrastructure events, technical workshops, audio/video conferencing and other collaboration mechanisms. Working groups and task forces bring in relevant experts from outside ELIXIR such as e-Infrastructures. This task will be made in collaboration with WP1 (Tools) and WP12 (Management).

Partners: ELIXIR Nodes

Task 4.2: User Facing Support (69PM)

This task interacts with the individuals and projects that are users of the ELIXIR Technical Services platform through their defined Use Cases. The main consumers are the ELIXIR-EXCELERATE Use Cases (WP6 to 9) and other ELIXIR activities (e.g. ELIXIR-EXCELERATE WPs, ELIXIR pilots, and external projects like EC funded Centres of Excellence or Virtual Research Environments).

Subtask 4.2.1: Technical requirements (27PM)

Gather and analyse the technical requirements for the Technical services platform in order to define the detailed technical specifications and interfaces of the technical service platform. One outcome is the classification of the different Use Cases with technical terms (e.g. small compute, large data input-output; large compute, data access management, etc.). This work will feed into WP12 concerning requirements requested and procured from any external service providers.

(a) EXCELERATE Use Cases WP6 to 9.

(b) ELIXIR tools registry, ELIXIR training events, data transfers to/from ELIXIR data resources, and authentication and authorization.

Partners: EMBL-EBI, FI, CZ, ES, NL, NO

Subtask 4.2.2: User support and integration (42PM)

Provide a support structure that can be applied to adopters of the ELIXIR Technical Services. This will be focused on the use of ELIXIR Technical platform e.g. for supporting organizing a training event. Task provides operational support for ELIXIR-EXCELERATE activities and externally funded ELIXIR activity (technical pilots and projects). This will take place through a single-point-of-contact ‘helpdesk’ function and ‘hackathons’ where users and the providers of the ELIXIR Technical Services work together to integrate functionality across AAI, cloud and data. As a result of this work

a set of ‘recipes’ focused around user activities will be collected into a ‘cook book’ to enable community adoption (e.g. to run a Galaxy workflow environment on an ELIXIR-affiliated Cloud Resource with accounting if necessary).

Partners: EMBL-EBI, FI, SE, FR, NL

Task 4.3 Technical infrastructure integration (99.5PM)

This task focuses on the integration of Technical Services being delivered by individual ELIXIR Nodes and from the public e-Infrastructures in Europe to meet the requirements of the ELIXIR community (e.g. by establishing account manager relations with each e-Infrastructure). The strategy within this task is to focus on the integration of existing mature and stable services to ensure that these services are easier to uptake by bioinformatics Use Cases (WP6 to 9).

Subtask 4.3.1: ELIXIR AAI - Authentication, authorization (access) integration (27PM)

ELIXIR needs a service based on European federated identity that authenticates an individual is a member of a group (or has a particular role within a group) that can be managed remotely. Group management needs to enable delegate decision making to multiple individuals within a particular community (e.g. institutional representative within a project) and queries from other services.

(a) Establishing an ELIXIR Identity: Federated identity technologies are fairly mature, as are many of the related tools (e.g. REMS, PERUN). This task ensures that ELIXIR research community is fully covered (including users whose home organization does not provide federated identities) and acts as a single IdP for ELIXIR branded services technical work.

The task continues to integrate the existing services and ensures that they provide the interfaces needed for adoption within this Work Package, the project and externally (e.g. BMS RIs).

(b) Providing additional AAI services: eduGAIN IdPs, Common IdPs, guest login, Proxy IdP, ELIXIR directory, Attribute self-management for users, Bona fide researcher management, Group/role management, Dataset authorisation management, Credential translation.

Partners: CZ, FI, EMBL-EBI, NL

Subtask 4.3.2: Cloud and Compute integration (41.5PM)

European and national compute centres service clusters and access to resource with open-source cloud technologies is growing. This task integrates the willing services to ELIXIR registry. The way how e.g. IaaS resources are consumed in the research community typically takes place on science-specific platforms and workflows. As a priority WP4 secures resources to support the scientific software workflows for the Use Cases WP6 to 9 and WP11 using the software environment workflows they have chosen for their data analysis framework (e.g. supporting provision of Galaxy as a

service for marine metagenomics pipeline).

(a) ELIXIR Cloud accounts: Integrate willing providers (e.g. Embassy Cloud), national level (e.g. CSC, SURFsara, Nordic Secure Cloud, MetaCentrum and CERIT-SC) and regional level (e.g. GÉANT, Helix Nebula and EGI cloud resources) and in the commercial sector (e.g. Amazon, Microsoft, Google). Mechanisms are needed to calculate virtual access costs that can be passed on to projects or funding agencies. Key target is to make accounts to provide resources for WP6 to 9 and WP11 activities.

(b) Enable SME access to ELIXIR cloud resources. We will support billing models such as monthly fee for service subscription or allocation-based costs when free (pre-paid) access is not available. Cost models will be developed with WP12.

Partners: NL, CZ, FR, EMBL-EBI, FI

Subtask 4.3.3: Storage and data transfer (21PM)

Data push and pull is needed in WP6 to 9 supported with commonly agreed technical tools and interfaces. Various transport mechanisms (e.g. GridFTP, http, Aspera, UDPipe, iRods) can be used to move the data to or from Data Resources (WP3). The managed access integration will be piloted in the ELIXIR-EXCELERATE Use Case WP9. Collaboration with GÉANT (e.g. bandwidth-on services) could be used to provide dedicated network links (e.g. lightpaths) for regular or large data transfer activities between the Nodes. Three common uses will drive WP4 storage and data transfer activities:

(a) Data replication (an updated dataset being moved to multiple remote locations) and data submission (where a dataset is made available for subsequent retrieval and remote analysis). In the former the data source triggers data movement to data sink(s) (e.g. using Globus Transfer) using a replication policy around the data and updates any relevant data catalogues (e.g. B2FIND).

(b) Service to pull relevant datasets for detailed analysis (e.g. Galaxy running on ELIXIR-affiliated cloud resource during training event). The retrieved dataset may be discarded after processing and just the results are retained based on the assumption that the original data will remain accessible for re-analysis.

(c) Data location services will be used to manage and discover data replicas within ELIXIR sites (using technologies such as B2FIND or the EGI Data Catalogue). AAI mechanisms and workflows (e.g. REMS) are needed for gaining approved access entitlements in collaboration with the responsible granting bodies such as data access committees (e.g. EGA).

Partners: SE, EMBL-EBI, ES, CZ, FI

Subtask 4.3.4: Service Registry (10PM)

Integrate with WP1 and WP3 service registry and existing e-Infrastructure registries to enable a wide range of ELIXIR services and resources (e.g. cloud, storage, datasets) so that they become discoverable entities. The service registry provides a ‘gateway’ by which service providers can advertise and the users consume services. The service registry needs to provide a ‘service discovery’ function for consumers, but also written advice and requirements on how service providers can advertise their services.

Partners: EMBL-EBI, CZ

7. Appendix 1: The ELIXIR Compute Platform: A Technical Services Roadmap for supporting Life Science Research in Europe

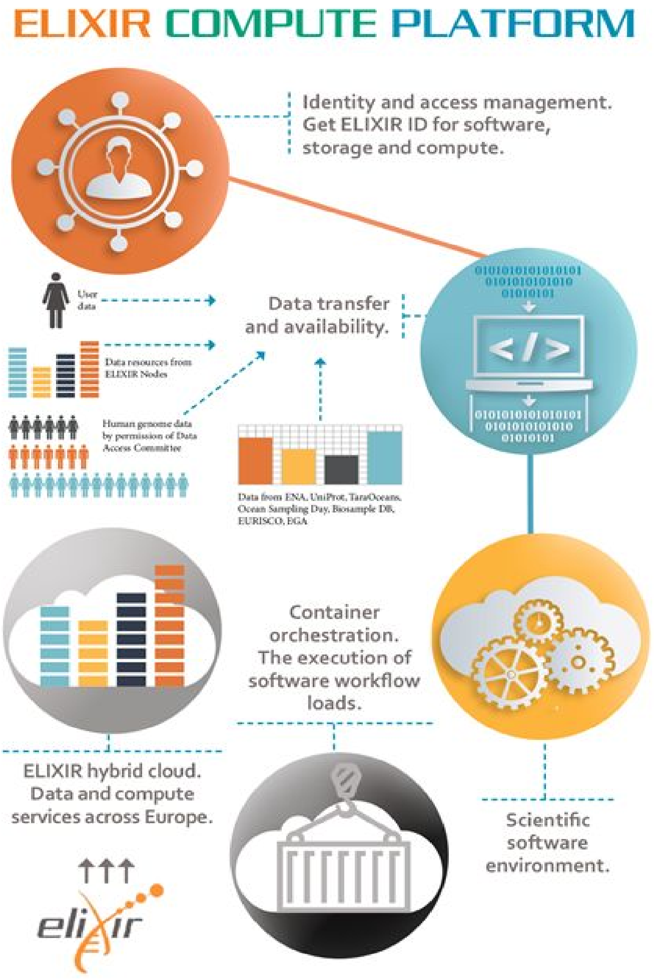

ELIXIR COMPUTE PLATFORM
Identity and access management. Get ELIXIR ID for software, storage and compute.
User data
Data resources from ELIXIR Nodes
Human genome data by permission of Data Access Committee
Data transfer and availability.
Data from ENA, UniProt, TaraOceans, Ocean Sampling Day, Biosample DB, EURISCO, EGA
Container orchestration. The execution of software workflow loads.
ELIXIR hybrid cloud. Data and compute services across Europe.
Scientific software environment.
elixir

Figure 1. The ELIXIR Compute Platform.

7.1 Background

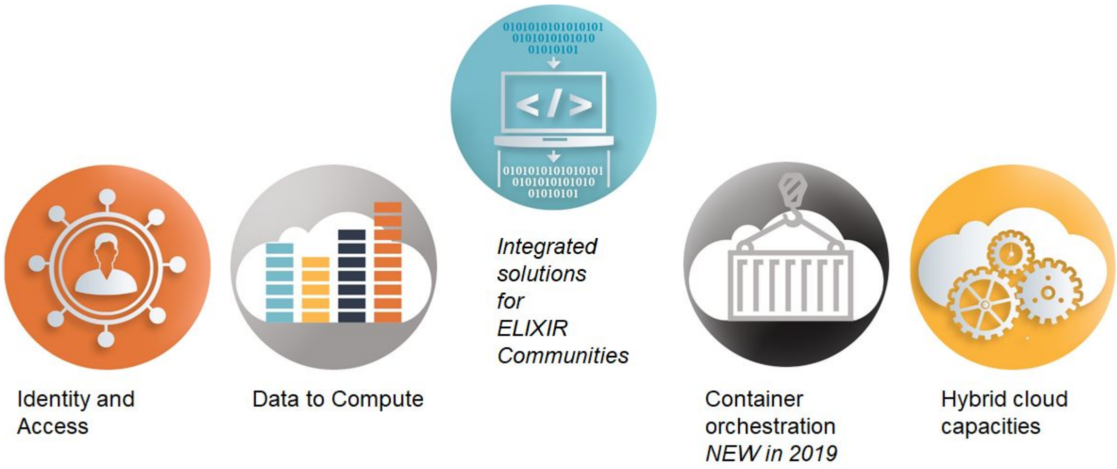
The ELIXIR Compute Platform (ECP) was established, with the support of the ELIXIR-EXCELERATE project, to support the ELIXIR Scientific Programme for 2014-18 and is now strongly integrated into the forthcoming 2019-23 programme. The ECP coordinates geographically distributed ELIXIR Authentication & Authorisation Infrastructure (AAI) as well as links to Cloud & Compute, Storage and File Transfer Services that are provided by the individual ELIXIR Nodes and are made available through ELIXIR and the emerging European Open Science Cloud. The ELIXIR Nodes and the ECP as a whole, collaborate with the European e-Infrastructures to increase the capabilities on offer to the European life science research community. The ECP defines a minimal 'neck' of an hourglass that Researchers and Application Developers can build upon and which ELIXIR Nodes and other Infrastructure Service providers can deploy and support. The service building strategy is to collaborate with existing initiatives and organisations (e.g. GÉANT, EGI.eu, EOSC, EU 1 million+ genomes, Biomedical science ESFRIs, GA4GH.org, IMI2) rather than developing new services only within ELIXIR.

This deliverable is synchronised with the work planned for the ELIXIR 2019-23 programme to ensure sustainability of the EXCELERATE outcomes. This report provides an update to the ELIXIR Technical Services Roadmap (D4.1[8], D4.2[9], and D4.3[10]) that were written at the end of PY1, PY2 and PY3 of the ELIXIR-EXCELERATE project. The ECP has also produced a public high-level description of the roadmap[11] based upon previous versions of this deliverable. The ELIXIR Technical Services Roadmap has been revised and updated to reflect the work and experiences that happened in PY4 and to help inform Researchers and Application Developers about the ECP expertise and services that will be available to them in the coming years. For ELIXIR Nodes and associated Infrastructure Service providers, this report identifies the technologies that should be deployed that will enable ELIXIR to provide a consistent set of Infrastructure Services to support life-science research in Europe.

It is important to note that the role of ELIXIR Compute experts is not to undertake middleware development within the ELIXIR-EXCELERATE project. Instead the focus is on leveraging the investment that has already been made nationally, by EC projects or commercially in services that can be integrated to serve the European research needs in biological information, and to influence on-going and future development priorities of European e-Infrastructures.

7.2 Services

The ECP is scoped around providing a set of key capabilities through a set of integrated services.



[8] https://docs.google.com/document/d/1chQOZwnaEQRQbow4L5Vq3PVCiXP6E_q5Qflu74inCFU
[9] https://drive.google.com/drive/folders/0B2GBqcAfmBgSNnN3Q3FKUHYzNWM
[10] https://drive.google.com/open?id=1hpb78hY_Lpt4h1qAVtBung76ZaH5jlbZ
[11] http://drive.google.com/file/d/0B0KXZdVao0kqUE9BbXVrc3ZLY1E/view

Figure 2. Key functionalities of the ELIXIR Compute Platform at the end of the ELIXIR-EXCELERATE project. Container Orchestration Task was included to the platform in the 2019-23 ELIXIR programme.

7.2.1 Authentication and Authorisation Infrastructure (Task 4.3.1)

Reliable electronic identification of users is needed to access the distributed services. ELIXIR Authentication and Authorisation services allow Users to use their federated academic, corporate or social media identity and link them to a unique ELIXIR ID. The scientific service providers connected to ELIXIR AAI receive securely communicated and managed user information using a common user identity. This allows for service integrations in the federated and distributed infrastructure. So access management services can be built on top of the federated identity services with appropriate contracts and policies, enabling building pan-European or global organisational trust frameworks. This has many benefits for the establishment of data security and privacy. For example, when the institutional affiliation of a user changes, the access rights coupled to the institutional status of the user are automatically suspended or revoked, unless the data resource access authority decides otherwise.

7.2.2 Cloud & Compute (Task 4.3.2)

A cloud and local compute infrastructure is needed to undertake data analysis. Cloud services need to be federated to provide uniform operation and secure access to storage. Private network solutions to access service are possible and ideal for users that require high performance, high security and certified environments for e.g. sensitive human data handling.

The ECP cloud/compute surveys conducted in 2017 and 2019 indicate that nodes operate considerable Cloud & Compute e-Infrastructure services for Life Science. In 2019 this consisted of 80,000 compute cores, 50,000TB storage and 3,100 users. This represents a doubling of the available storage capacity and a 33% increase in compute capacity over 2 years. These services are fully utilised and growing when the local and national funding situation allows growth. European users can gain access by e.g. complying with national or organisational access policy, being a member of a specific international collaboration project, or payment.

7.2.3 Storage and Data Transfer (Task 4.3.3)

Data transfers are needed across all scientific use cases and various data transport mechanisms have been investigated to organise data transfers between core biological data resources. Storage and Data Transfer include:

- Data replication and data submission to or from ELIXIR Data Resources
- Services to pull relevant datasets from Data Resources or their replicas to cloud or compute services for detailed local analysis
- Data location services to manage and discover data replicas within ELIXIR established to decrease network overload for ELIXIR nodes hosting large data sets and deter ad hoc data transfer and storage.

7.2.4 Infrastructure Services Registry (Task 4.3.4)

It was always expected that ELIXIR would need an Infrastructure Service Registry to be able to register cloud services and for these to be discovered by users. It was soon realised that this would be a core capability within EOSC and will provide the capability that ELIXIR would be looking for. The ECP is now registering cloud resources within the EOSC marketplace, including:

- EMBL-EBI Embassy Cloud[12]
- MetaCentrum Cloud[13]
- CSC ePouta[14] (description update submitted in July 2019 to include CSC cPouta[15])

7.3 European e-Infrastructures

The ECP has continued to establish and strengthen collaborations with the European e-Infrastructure community through the EOSCpilot, EOSCHub, AARC2, and in previous years the EGI-Engage, EUDAT2020 and AARC projects. Much of this work will now continue in the EOSC-Life project where the ECP will lead on the Life Science AAI and Cloud. EOSC - the European Open Science Cloud - represents a major new European initiative to integrate distributed resources to support Open Science. The EOSCpilot project was a community activity to help define what EOSC might be, and the EOSC-Hub project is tasked with implementing the initial phase of EOSC. The recently started EOSC-Life project brings together common requirements of the life science ESFRIs so that the European e-Infrastructure can support Life Science Research infrastructures more effectively. The role of the ECP is to help support and drive this process by acting as an interface to the life science community into EOSC.

Even with the emergence of EOSC, the role of the ECP remains the same: 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 services to provide. The requirements defining the ‘neck’ of the hourglass have been broken down into ‘generic’ Technical Use Cases (TUCs) that need to be delivered by the ECP. The focus of the ECP is to identify the services that need to be deployed in the ELIXIR Nodes or drawn from European e-Infrastructures to meet these established needs.

7.4 Assessment of Technical Use Cases

Each year the ELIXIR Technical Roadmap has undertaken an assessment of the maturity of the individual Technical Use Cases (i.e. capabilities) that were identified in PY1 during

[12] https://marketplace.eosc-portal.eu/services/embassy-cloud
[13] https://marketplace.eosc-portal.eu/services/metacentrum-cloud
[14] https://marketplace.eosc-portal.eu/services/csc-epouta
[15] https://research.csc.fi/cpouta

the analysis of the Scientific Use Cases[16]. The following table records the evolution of the maturity status of each TUC over the years using the following classification:

- PoC: Proof of Concept - A capability that is still being defined.
- ES: Emerging Services - A service that has a defined capability that is still being developed in conjunction with users.
- MS: Mature Services - A service that has a defined capability in production use.
- LS: Legacy Services - A service that has been identified for retirement as it is no longer needed.

Table 1. Assessment of the overall status of individual TUCs over ELIXIR-EXCELERATE

ID	Technical Use Case	PY4	PY3	PY2	PY1
1	Federated ID	**MS**	MS	MS	ES
2	Other ID	**MS**	MS	MS	ES
3	ELIXIR Identity	**MS**	MS	MS	ES
4	Cloud IaaS Services	**MS**	MS	MS	MS
7	Network File Storage	**ES**	PoC		
8	File Transfer	**MS**	ES	ES	PoC
9	Infrastructure Service Directory	**MS**	ES	ES	PoC
10	Credential Translation	**ES**	PoC	PoC	PoC
11	Service Access Management	**MS**	MS	MS	MS
12	Virtual Machine Library	**LS**	MS	ES	PoC
13	Container Library	**ES**	PoC	PoC	PoC
15	Data Set Replication	**ES**	PoC	PoC	PoC
16	Infrastructure Service Registry (from EOSC)	**MS**	ES	PoC	
17	Endorsed Personal Data or Compute Access Management	**ES**	ES	PoC	PoC
19	PID and Metadata Registry	**ES**	PoC		
20	Federated Cloud IaaS	**MS**	ES	ES	PoC
21	Operational Integration	**ES**	ES	ES	PoC
22	Resource Accounting	**ES**	ES	ES	PoC

[16] https://docs.google.com/document/d/1mLeaFk5jlYIKQVRC6Vu2X7AKZBVGywFtf28t_NUO1iU/edit#

The initial analysis of the Scientific Use Cases in PY1 identified TUCs that have not been implemented by the ECP as we have lacked user demand. These include a set of cluster related TUCs:

- HTC/HPC Cluster (TUC 5)
- PRACE Cluster (TUC 6)
- Module Library (TUC 14)
- Federated HTC/HPC Cluster (TUC 23)

When ELIXIR-EXCELERATE started, many researchers were considering the migration of research workloads from cluster to cloud resources. While clusters are still in active use, it is clear that new workloads are targeting cloud based, including provision of container environments, rather than clusters, so we don't expect an increase in interest for these cluster related TUCs. ECP commends application developers to consider portability (i.e. containerisation) early on in the development process, as this will technically enable leveraging distributed compute infrastructures if scaling up is needed.

The Cloud Storage (TUC 18) has effectively been implemented through the adoption of S3 (or equivalent) API to Object Stores which are supported by most ELIXIR, EOSC or commercial cloud providers.

7.5 ELIXIR-EXCELERATE Activities in Project Year 4

7.5.1 Background

During the fourth year of the ELIXIR-EXCELERATE project, WP4 has continued to build and develop the management, technical support, and technical structures needed for the ELIXIR Compute Platform (ECP) in response to the scientific and training use cases. Background for the work can be found in Deliverables D4.1, D4.2 and D4.3 as referenced previously in Section 2.1. A particular focus over this year has been planning the transition from ELIXIR-EXCELERATE which funded coordination and technical development activity across the ELIXIR community, to a mixed ecosystem where the ELIXIR Hub funds coordination of the ECP and other projects fund technical development activity. These other projects include the ELIXIR Implementation Study programme, EC funded projects such as EOSC-Life, EOSC-Hub, EOSCpilot, and CINECA, but also nationally funded activities coming from the individual ELIXIR Nodes.

7.5.2 Leadership

The management and leadership of the ECP continues through an Executive Committee that meets twice per month to coordinate internal and external activity. The ECP task leads meet once a month and once a month there is a general ECP community meeting. Furthermore, the ECP participates in the monthly cross-platform meeting between the major ELIXIR platforms and ELIXIR communities involved in ELIXIR-EXCELERATE. The ELIXIR-Hub provides a dedicated Platform Coordinator.

Bi-directional technical interactions have continued with European e-Infrastructures (such as EGI, EUDAT and GÉANT). Most of the effort during PY4 has focused on projects related to the European Open Science Cloud (EOSC). The ECP played a major role in EOSCpilot project which started in January 2017 and ended in March 2019, which through a series of three open calls supported 4 life-science related projects[17] including PanCancer cloud analysis, analysis of life-science data sets, Cryo3D workflows and Bioimaging. The EOSC-Hub project started in January 2018 and the EOSC-Life project (coordinated by the ELIXIR-Hub) started in March 2019 with ECP partners having lead roles in WP5 identity and access management and WP7 cloud deployment.

7.5.3 User Facing Support

User Facing Support with the users and relying parties of the ELIXIR AAI services has started. Direct interaction with the Scientific Use Cases and the Training activities coming from the ELIXIR Communities has triggered updates to the motivating use cases and the prioritised TUCs extracted from these use cases which are reported separately[18]. In PY4 there has been a focus on the rare diseases use cases with the ECP working closely with

[17] https://eoscpilot.eu/science-demonstrators

[18] https://docs.google.com/document/d/1mLeaFk5jlYlKQVRC6Vu2X7AKZBVGywFtf28t_NUO1iU/edit#

ELIXIR-EXCELERATE WP7 (Integrating Genomic and Phenotypic Data for Crop and Forest Plants) and WP9 (Human genetic Data) to support their use cases.

With the ECP now offering services to its users, the pilot support infrastructure needed to provide good user experiences so helpdesk, documentation, FAQs have been established. A central contact point for the ELIXIR AAI has been identified and a team (1st line) to triage and assign support requests to the relevant service specific support teams (2nd line) is now being established.

7.5.4 Technical Infrastructure Integration

This work has been driven by the continuing work supporting the ELIXIR-EXCELERATE WP6 (Marine Metagenomics) use case that started in PY1, and expanded with support in PY2 for Human Data and the plants use case in PY3. PY4 saw support being provided for the Rare Disease use case (WP8) The ECP now provides established services for transfer of large volumes of electronic, confidential, human data coupled to cloud compute access to the data with the ELIXIR AAI. More information about the accomplished technical integrations has been made publicly available (Milestone M4.1[19], Milestone M4.2 [20], Milestone M4.3[21], Milestone M4.4[22]).

7.5.4.1 Authentication and Authorisation Infrastructure

The AAI service allows a user to create a unique ELIXIR identity by linking pre-existing identities (e.g. Google, ORCID or the researcher's home university as an attribute and an identity provider). Individual users can enroll into an ELIXIR Virtual Organisation, and potentially into groups within this Virtual Organisation (using Perun[23]). Group structures allow support of actual international research groups or projects, enabling resource allocation decisions to be recorded in the user and group metadata. This organisational information has then been exposed through the ELIXIR Proxy IdP to identified relying parties to use this information for authentication and authorisation decisions. Step up authentication models through multi-factor authentication and bona fide researcher management service are now also provided.

The ELIXIR AAI has been driven primarily from the ELIXIR FI and ELIXIR CZ nodes with strong support and feedback from the ECP. Successful collaborations have been established with the AARC2 project in terms of providing requirements and benefiting from some of their service prototyping.

In August 2019, at the end of ELIXIR EXCELERATE project, ELIXIR AAI enabled 721 institutions whose members can use ELIXIR AAI, integrated 63 production services and an additional 63 in testing, and had 2923 users organised in 503 groups.

[19] https://elixir-europe.org/events/webinar-access-to-sensitive-data-aai
[20] https://elixir-europe.org/events/webinar-access-to-sensitive-data-aai
[21] https://github.com/NBISweden/excelerate-demonstrator-4.3
[22] https://docs.google.com/document/d/1XXmgAD6zT9_Oa_2NF6l6MPUH3kAijx3mS_V41LUSQPA
[23] https://perun.cesnet.cz

7.5.4.2 Cloud & Compute

ELIXIR Nodes continue to provide either directly or through national partners a set of cloud and compute resources potentially available for use by the ELIXIR community. The potential capacity of 80,000 cores and 50,000 TB storage and 3,100 users is available for researchers to access under different conditions depending on the access conditions of each individual site. The lack of a consistent ELIXIR wide provisioning policy for cloud and storage resources is an issue. Indeed, each cloud service provider operates independently meeting the needs of its local user community using the mechanisms defined by its funding agency. These mechanisms may include open access to funded researchers, formal peer review of proposed research use, pay per use, etc. External users, such as those from ELIXIR Nodes, may not be able to use these resources directly, but have to apply through a local collaborator.

There is currently no clear mechanism as to how researchers who have no local cloud resources can gain access to cloud resources without paying for them directly, nor is there any mechanism likely to be put in place by which ELIXIR Nodes who might be able to provide such resources free at the point of delivery will be reimbursed. Currently the only foreseen mechanism is for usage to be reimbursed centrally for providing the service. However, the EOSC-Life project will establish a resource allocation process and has money to fund the use of cloud resources either from within the project (i.e. the BMS community) or from public cloud providers within a procurement framework (e.g. OCRE project or GEANT framework).

This is an issue that was a barrier to non-trivial resource use within EGI and similar issues were already occurring within the EOSCpilot - so ELIXIR is not alone in trying to establish a sustainable model and work will take place within the EOSC-Hub and EOSC-Life to explore and pilot different models, potentially learning from micro services based approaches used in finance and banking. The situation is complicated. The long-term aim of the ECP is to increase transparency of the details of infrastructure (resources) provisioning/allocation. With the support of e.g. EOSC-Life development, ECP expects to be able to better report international consumption of compute and storage by the Life Science users for the purposes of data analysis.

The work has been primarily led by EMBL-EBI and ELIXIR CZ and resulted in a strong collaboration with other ELIXIR sites, in particular with ELIXIR FI, and with e-Infrastructure activities (such as EGI, Helix Nebula, and the European Open Science Cloud).

7.5.4.3 Storage and Data Transfer

Data transfers are needed across all the scientific use cases and various data transport mechanisms have been investigated to organise data transfers between ELIXIR data centres. FTS3, a file transfer coordination service, has been integrated with the ELIXIR AAI as a potential service for the ECP. The Reference Data Set Distribution Service was initially developed with the support of the EUDAT2020 project and more recently with the support of ELIXIR-EXCELERATE. It is now an emerging service increasing usage and reliability.

The work has been primarily led by ELIXIR SE, EMBL-EBI and ELIXIR NL.

7.5.4.4 Infrastructure Services Registry

An Infrastructure Service Registry is needed to provide a live picture of the available technical capabilities of the ECP. The information in the Infrastructure Services Registry is composed of both static information (e.g. contact URL, physical capacity) and dynamic capability information (e.g. free CPUs, free storage). Similar work is being undertaken by the EOSC-Hub project that launched an EOSC Portal[24] and specifically by the eInfraCentral project that provides a catalogue of the European services for research[25]. In order to minimize duplication of work ECP supports registration of ELIXIR services into the EOSC portal (e.g. CSC ePouta[26], EBI Embassy Cloud[27] or MetaCentrum Cloud from CZ[28]).

This work is being led by EMBL-EBI and ELIXIR CZ.

[24] https://www.eosc-portal.eu/
[25] https://www.einfracentral.eu/home
[26] https://marketplace.eosc-portal.eu/services/csc-epouta
[27] https://marketplace.eosc-portal.eu/services/embassy-cloud
[28] https://marketplace.eosc-portal.eu/services/metacentrum-cloud

7.6 Interactions with ELIXIR Scientific Community Use Cases

ELIXIR Compute Platform (ECP) aims to ensure that distributed cloud, compute, storage and access services are increasingly available for the European life-science research community. In the ELIXIR-EXCELERATE project, the ECP worked closely with four scientific communities and training provided by the ELIXIR nodes to ensure the technical solutions fit their specific needs. Each of the scientific use cases establishes and manages their standards for describing and accessing datasets, reporting data, matching and comparing content, and eventually building linkages between datasets. Those specifications then provide basis for defining the responsibilities of the ECP in the overall research process.

The scientific use cases driving the establishment of the ECP in 2015-2019 were:

- Marine metagenomics data analysis
- Plant data coordination
- Rare diseases research and diagnosis
- Secure transfer and sharing of sensitive human genomics data
- Training

The objective for each use case was to combine relevant ECP services and expertise from the ELIXIR Nodes into a seamless workflow. For example, a researcher might use the ELIXIR Authorisation and Authentication services to securely create a scientific software analysis environment and use the environment to access large biological data resources stored in a cloud.

7.6.1 Marine Metagenomics data analysis support

Microbial communities affect human and animal health and are critical components of all terrestrial and aquatic ecosystems. Communities can be exploited e.g. to identify novel biocatalysts for production of fuels or chemicals (bioprospecting), make functional feed for aquaculture species,and for environmental monitoring. The importance of plankton in maintaining the Earth’s climate cannot be understated – their communities absorb a staggering volume of CO2 from the atmosphere and release oxygen in exchange. Yet only a small fraction of these life forms have been classified and analysed.

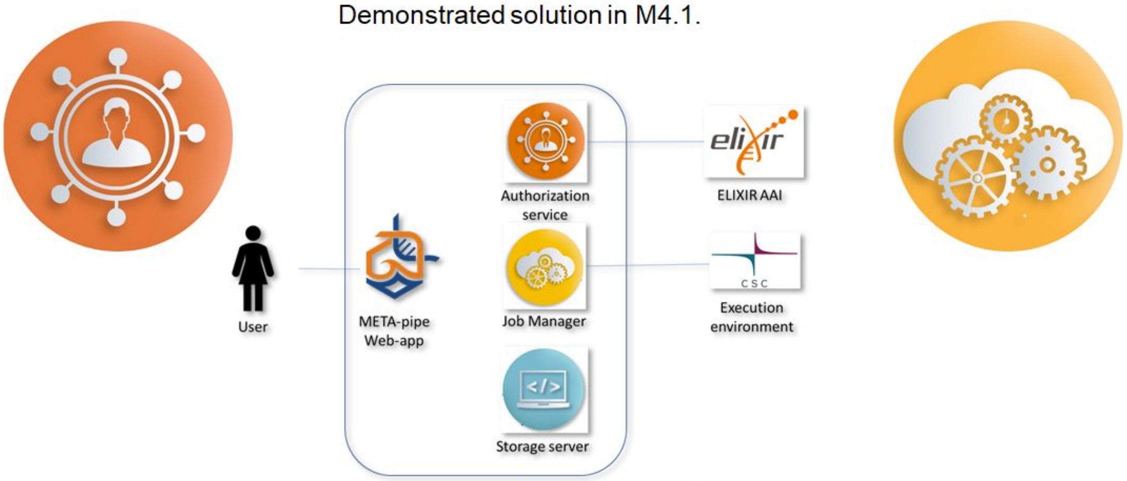
Marine metagenomics community aims to build sustainable public data resources to improve the characterisation of marine metagenomic samples. This will be achieved by establishing marine microbial databases including reference genomes, nucleotide and protein databases. The established databases, based on the standards developed in the project will enhance the precision and accuracy of biodiversity and function analysis. The reference databases will be non-redundant datasets generated from sequences acquired from European Nucleotide Archive (ENA) as part of the International Nucleotide Sequence Database Collaboration, UniProt and other publicly available datasets. For example, Tara Oceans expeditions resulted in 35,000 samples of seawater, each of which contained millions of small organisms. The samples were sequenced at Genoscope in

France, generating over 7000 datasets. This revealed 40 million novel genes and a raft of discoveries about life in the world's oceans. ELIXIR will use some of the high coverage sequence outputs from the TaraOceans to build marine-specific reference databases

Metagenomics methodologies need to overcome a number of challenges related to standardization, development of relevant databases and bioinformatics tools. New and emerging sequencing technologies, integration of metadata gives an extra burden to the development of future databases and tools. Due to the data biases of existing reference databases, only about one quarter of sequences are annotated, and this fraction diminishes further when more diverse samples such as soil and marine are analysed. The scientific goal of this work is to construct sustainable public data resources to improve the characterisation of marine metagenomic samples. This work is executed mostly in three ELIXIR Nodes (EMBL-EBI, NO, FR). The databases will be developed in collaboration with members of the ESFRI infrastructures European Marine Biological Resource Centre (EMBRC) and Microbiological Resource Research Infrastructure (MIRRI) and made publicly available through ELIXIR.

ELIXIR-EXCELERATE technical demonstrator M4.1[29]. with Marine metagenomics community (WP6) with the ELIXIR Compute Platform (WP4) linked user portal Tools (WP1 & WP2) integrating ELIXIR AAI to identify the end-users across the nodes providing the services. The technical integration needed by the community included federated AAI, cloud and compute services, file transfer services and storage availability in the distributed sites and a way to account for the overall usage. The tools and pipelines for the identification of gene products developed (e.g. enzymes and drug targets) were made portable by the creators of metagenomics data analysis experts and available for users from several ELIXIR nodes (NO, EMBL-EBI, FI, CZ, FR).

Demonstrated solution in M4.1.



Link (more information)

Figure 3. Demonstrator with Marine metagenomics WP6 linked user portal Tools (WP1 & WP2) to ELIXIR Compute (WP4) integrating ELIXIR AAI.

[29] https://elixir-europe.org/events/elixir-webinar-elixir-compute-platform-roadmap

7.6.2 International Transfer of Human Access-controlled Data

The use case around human access-controlled data uses the ELIXIR framework for secure submission, archiving, dissemination and analysis of human access-controlled data.

ELIXIR Human data community work extends and generalise the system of access authorisation management and high volume secure data transfer to address growing needs by major initiatives such as the EU 1 million + genome declaration[30] by 21 signatory countries. The European Genome-phenome Archive (EGA) is designed to be a repository for all types of sequence and genotype experiments, including case-control, population, and family studies. The archive allows exploration of datasets from genomic studies, provided by a range of data providers. The EGA will serve as a permanent archive that will preserve several levels of data including the raw data as well as the genotype calls provided by the submitters. A centrally federated human data service will allow authorised third-party services to programmatically check that a user is authorised to access data stored all the distributed repositories that belong to the federation. This will also provide support for the dataset owners to automate their data access processes (e.g. usable technology, implementing policy, granting permissions).

Following data deposition to the EGA, the essential following infrastructure function in the workflow is data release to authorized individual users from the archive, and to partner with downstream secure data analysis on a trusted compute platform. Data ownership and access decisions need to be maintained in the hands of the original resource owner who has acquired consent from the study participants. The overall workflow will also allow resource owners to focus on their unique areas of data generation and analysis expertise while being able to rely on the EGA and the ELIXIR infrastructure for sustaining their common big genomics data storage, coordination and distribution needs under appropriate legal and data security frameworks.

ELIXIR-EXCELERATE technical demonstrator M4.2[31] presents a solution to enforce access to sensitive datasets in a distributed infrastructure setting with secure infrastructure as a service (IaaS) compute clouds. The workflow supports distributed data collection and release to authorized individual users from the EGA, and allows the data service to partner in downstream secure data analysis with external clouds with the support of the ECP. Researchers identified with ELIXIR AAI will be given permissions to access datasets, which are managed in a central database (like EGA). Access control needs to be enforced on a trusted clouds at the infrastructure level. Technical standards such as OAuth2-based and OIDC-based architecture have developed during 2017 and 2019, and meet the requirements to achieve data security needed to manage research data from human subjects in collaboration with the data access committees..

[30] https://ec.europa.eu/digital-single-market/en/european-1-million-genomes-initiative
[31] https://elixir-europe.org/events/webinar-access-to-sensitive-data-aai

The researcher first discovers data through EGA website, and is then directed to use ELIXIR Authentication and Authorisation Infrastructure (ELIXIR AAI) to apply for the data access. To store and analyse the data, the researcher can use local storage and computing resources made available from a distributed ELIXIR cloud service: The researcher requests the ELIXIR Node offering the cloud service to transfer the dataset from the EGA data resource to the cloud service. Transferred data are then validated and moved to a secure location accessible to the cloud virtual machines. The ELIXIR cloud service notifies the researcher when the data are ready to be used. The researcher will then gain access to the data by authenticating with ELIXIR AAI. Technical integrations included ELIXIR AAI including both identity and access entitlement management, secure cloud & compute, secure file transfer and streaming, file storage, virtual machine & container library, and persistent dataset identifier services.

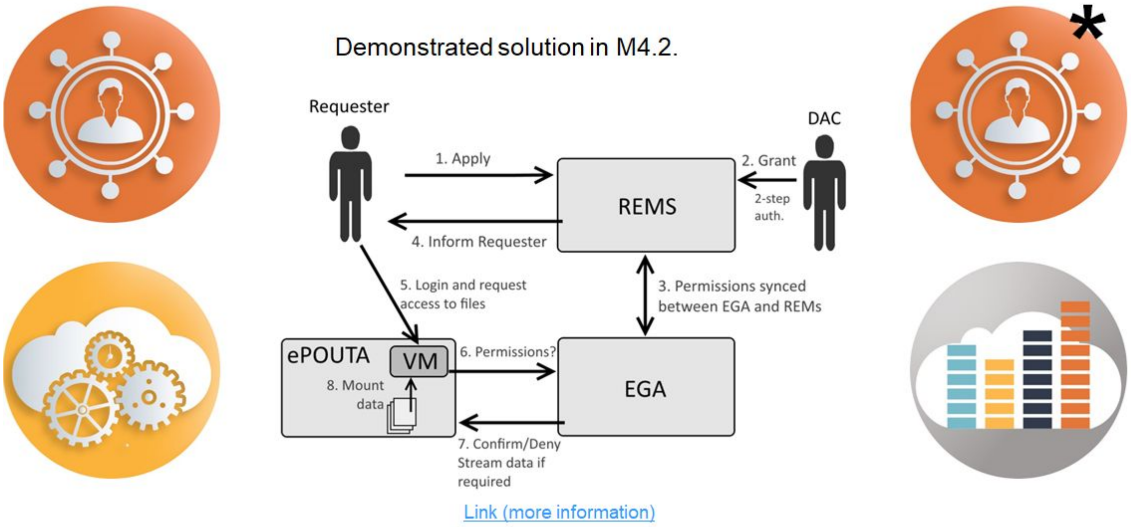

*
Demonstrated solution in M4.2.
Requester
DAC
1. Apply
2. Grant
2-step auth.
REMS
4. Inform Requester
5. Login and request access to files
3. Permissions synced between EGA and REMs
ePOUTA
VM
8. Mount data
6. Permissions?
EGA
7. Confirm/Deny Stream data if required
Link (more information)

Figure 4. ELIXIR-EXCELERATE demonstrator M4.2. with the Human Data community WP9 transferred sensitive human data from the EGA ELIXIR Core Data Resource (WP3) to users on a trusted cloud infrastructure (ePouta) according to Data Access Committee decisions.

7.6.3 Integrating Genomic and Phenotypic Data for Crop and Forest Plants

Massive sequencing and genotyping of crop and forest plants and their pathogens and pests generates large quantities of genomic variation data. Data is scattered across the laboratories seeking to describe and understand the life of plants at the molecular level. ELIXIR has designed an infrastructure to allow genotype-phenotype analysis for crop plants based on the widest available public datasets. Sequencing and genotyping efforts are likely to accelerate in the near future aiming to catalogue all genetic diversity present in global germplasm resources. However, structural variation in most crop plants is enormous - more so than in humans. Phenotypic characterisation of data is often inaccessible, diverse and non-standard. Furthermore, data lacks any route of unified access.

ELIXIR Plants community analyses many phenotypes against large panels of crop accessions through the aggregation of locally held data. This enables more powerful association analysis and opens the way to understand the candidate gene prioritisation in order to improve crop breeding. Working on exemplar species, ELIXIR Nodes will establish a sustainable model for the interaction of distributed phenotypic repositories with defined genomic and sample reference data. Organisations can expose data to the system through conformity with standards for annotation and interface. This allows the subsequent expansion of the approach to other species. It also provides resources in the form of standards, ontologies and models for annotation and collaboration for use within ongoing species-centric (e.g. the Wheat Initiative) and/or national endeavours.

The optimal model will be a scalable, distributed, and transparently integrated through the development and use of common vocabularies and search technologies. This will be done by using established repositories for genomic data and sample metadata. The expected impact will accelerate research and plant breeding through the exploitation of an interoperable commons of public data. ELIXIR Nodes will also work on establishing common guidelines for ontology usage when annotating crop and forest species. Sample identification will be handled through the BioSample DB at the EMBL-EBI, or, where the sample is an accession from a public gene bank, by cross-references to EURISCO, the European catalogue of plant collection data. The Nodes will develop a common API for data query and retrieval.

ELIXIR-EXCELERATE technical demonstrator M4.3[32] supports the plant science community to track and bring these data together, data transfers from geographically distributed sites onto a scalable Compute platform server. The scientific use case integrates genomic and phenotypic data of crop and forest plants from a variety of open access and local data sources. These data sources will need to conform to minimum standards set by the community. The central component is a search engine that receives search requests from the users and passes integrated search results retrieved from the distributed data sources back to the user. Based on these results users can select a cloud infrastructure that they have access to and transfer the selected data to that cloud resource to undertake their own analysis.

Technical service integrations needed to achieve this functionality include cloud & compute, catalogue of datasets and other (persistent) identifiers, file transfer and ELIXIR AAI. Demonstrator also developed a distributable storage endpoint virtual machine, which uses ECP Data Transfers technologies to move a set of files to the target cloud service.

[32] https://github.com/NBISweden/excelerate-demonstrator-4.3

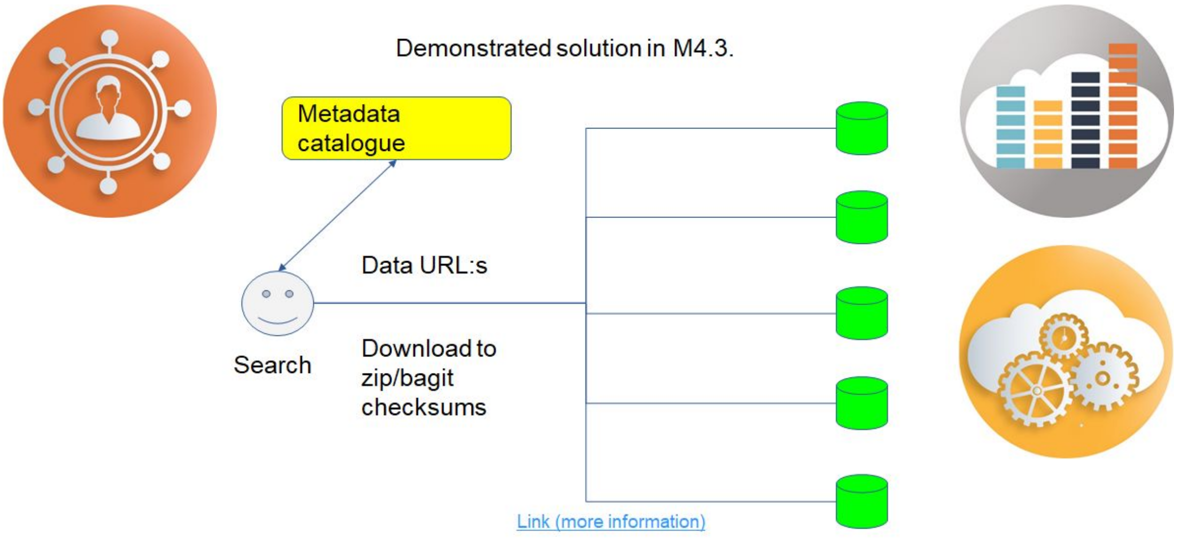

Demonstrated solution in M4.3.
Metadata catalogue
Data URL:s
Search
Download to zip/bagit checksums
Link (more information)

Figure 5. ELIXIR-EXCELERATE demonstrator M4.3 with the Plants community WP7 coordinated transfer of distributed (non-sensitive) user datasets to ELIXIR Compute.

7.6.4 Integrating ELIXIR Infrastructure for Rare Disease Research

According to EURORDIS (European Organisation of Rare Diseases) about 30 million people have a rare disease in the 25 EU countries, which means that 6% to 8% of the total EU population are rare disease patients. The International Rare Diseases Research Consortium established the ambitious goal of developing 200 new therapies by 2020.

This ELIXIR scientific community use case addresses the data integration needs of the rare diseases community. For example, the European Genome-Phenome archive (EGA) stores data from major research initiatives in rare diseases. The use case needs supporting research around rare (1 in 2000 people) chronic or genetic diseases and will use access controlled data sources such as the EGA. The metadata around a patient (i.e. their illness, treatments, outcomes), patient samples stored in a biobank, and any sequenced material stored in the EGA is searchable through a central portal which can only be accessed by authorised users. ELIXIR rare diseases community reviews current data resources and evaluates their usability and potential impact on the rare disease community. An important aspect of the evaluation is security of the data in rare disease research, given the low frequency of the associated genomic variants in the population.

The overall ELIXIR rare disease community aim is to interface and empower on-going and future rare disease research projects by addressing data interoperability, security and management bottlenecks. The portal queries the individual national search engines on behalf of the users. Selected datasets can then be downloaded into an EGA compatible cloud or cluster local to the researcher.

ELIXIR-EXCELERATE rare diseases demonstrator M4.4[33] allowed a researcher to submit their local raw data files, process the raw genomic data using the RD-Connect software

[33] *https://docs.google.com/document/d/1XXmgAD6zT9_Oa_2NF6l6MPUH3kAijx3mS_V41LUSQPA/*

tool pipeline, and map and obtain a genomic variant call file for further analysis via standard (GA4GH compatible) interfaces. RD-Connect is an existing integrated platform solution connecting databases, registries, biobanks and clinical bioinformatics for rare disease research purposes. In the demonstrator, files were processed using the RD-Connect pipeline on the distributed ELIXIR Compute nodes using GA4GH container cloud technologies. Importantly, all the software tools in the process are dockerised and made portable using similar technologies as in the marine metagenomic use case demonstrator (ELIXIR-EXCELERATE demonstrator M4.1.). Containers allow the creation of isolated environments that all share the same kernel. Workflows were submitted to different clusters in different places. Once processed, genomic variant call file were returned to RD-Connect for further annotation and inclusion. RD-Connect pipeline will adapt Common Workflow Language (CWL) which is a specification for describing analysis workflows and tools. CWL makes workflows and tools portable and scalable across a variety of software and hardware environments (workstations, cluster, cloud, and high performance computing).

Technical integrations needed by the demonstrator included ELIXIR AAI including identity and access entitlement management, cloud & compute, file transfer, file storage, virtual machine & container library, and persistent dataset identifier services.

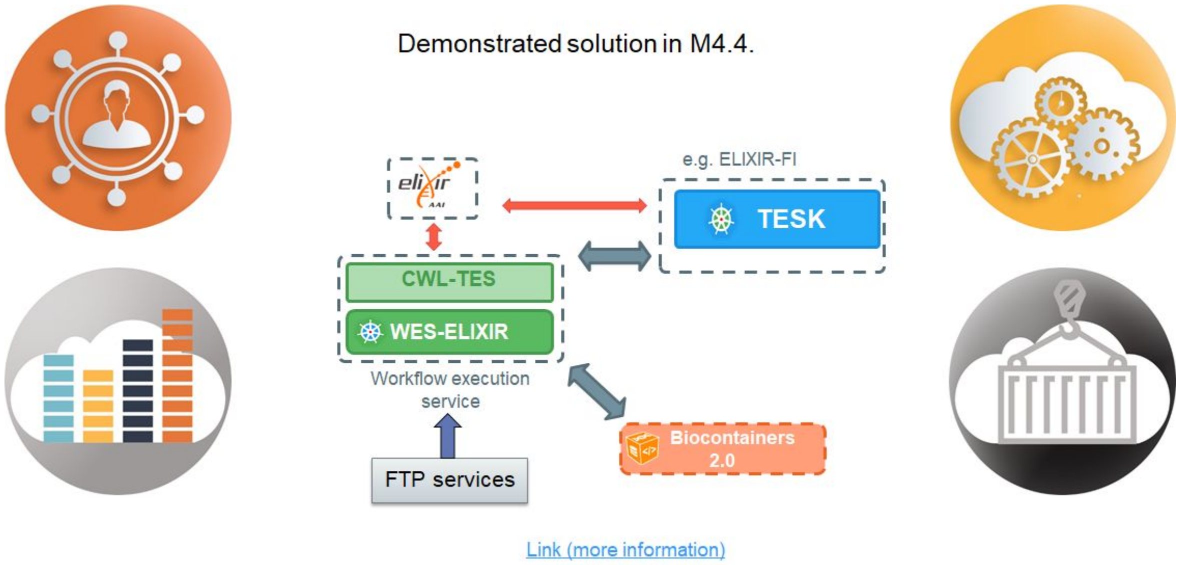


Figure 6. ELIXIR-EXCELERATE Demonstrator M4.4 with the Rare Diseases (WP8) created container workflow orchestration over distributed ELIXIR Compute (WP4) sites using GA4GH compatible architecture.

7.6.5 Training

The work of the ECP with the Training platform WP11 focused on the question: *How can ELIXIR bioinformatics trainers effectively use cloud resources as infrastructure for their courses?*

A cloud provider survey was conducted with the support of an ELIXIR Implementation Study funding, and usage experiences from six courses were collected. A clear message is that the cloud resource allocation process could be made clearer and faster. Trainers want to focus on preparing the course content.

The bottlenecks to leverage distributed infrastructure of ELIXIR for Training can be eased in several ways. In general, translating and evaluating cloud resource requests from training events and selecting suitable cloud providers requires a technical expert. Some ELIXIR node cloud and compute services already provide easy-to-use interfaces where participants can start e.g. virtual machines using ready-made images containing all the course software. ELIXIR also provides training for trainers on how to use cloud environments. Once resources for a course have been secured, the amount of technical support needed to set up training classroom environment to the servers for a particular bioinformatics course varies a lot. Trainer's have varying technical skills and time available for technical infrastructure work. Typically the needs for technical support exceeds the original estimates. ELIXIR considers setting up a technical support team for trainers, because this would remove the burden from each ELIXIR node to set this up individually, and opens possibilities to use any third-party cloud service providers.

During a course a lot of cloud resources are needed simultaneously, but for a limited amount of time. In order to ensure compute resource availability on the course day, the resources need to be reserved in advance. This means that those resources won't be available for other users meanwhile, and it adds to the computational service costs, because reserved idle resources consume billing units like active ones.

As a conclusion, recommendation is that the ELIXIR Hub could appoint a technical evaluation panel and provide an annual budget to get and support utilisation of distributed cloud resources for key ELIXIR bioinformatics training events.

7.6.6 Supporting ELIXIR Communities

Given the TUC prioritisation coming from the motivating use cases and the work that has been undertaken during ELIXIR-EXCELERATE by the ECP to implement the TUCs, the overall status at the end of the project is captured in the following table. It shows that the ECP is capable of supporting the motivating use cases with TUCs classed as 'Emerging' or 'Mature'.

Table 2. List of services offered to different use cases.

ID	Technical Use Case	Current Status	WP6	WP7	WP8+9	WP11
1	Federated ID	Mature	Y	Y		Y
2	Other ID	Mature				Y

3	ELIXIR ID	Mature			Y	Y
4	Cloud IaaS Services	Mature	Y	Y	Y	Y
5	HTC/HPC Cluster	Not in use	Y	Y	Y	Y
7	Network File Storage	Emerging	Y	Y	Y	Y
8	File Transfer	Emerging	Y	Y	Y	Y
9	Infrastructure Service Directory	Mature		Y		
12	Virtual Machine Library	Legacy	Y		Y	Y
16	Infrastructure Service Registry	Mature		Y	Y	
17	Endorsed Personal Data or Compute Access Management	Emerging			Y	
19	PID and Metadata Registry	Emerging		Y	Y	Y
22	Resource Accounting	Emerging	Y		Y	Y
23	Federated HTC/HPC Clusters	Not in use				Y

7.7 Future Plans

The future priorities for the ECP have been encapsulated in our plans for the ELIXIR Work Programme 2019-2023[34]. This work will focus around the coordination of activities across the ELIXIR Nodes in these key areas:

- AAI
- Moving data to compute
- Hybrid Clouds
- Containers

The coordination effort provided by ELIXIR will be used to:

- Integrate national efforts from ELIXIR nodes
- Identify opportunities for ad hoc projects using the ELIXIR Implementation Study model in collaboration with ELIXIR Communities and other ELIXIR Platforms.
- Collaborate with other projects (e.g. EOSC-Life) to continue the development of core capabilities that can then be deployed and sustained within the ECP.

The EOSC-Life project will be used to take forward key technologies that have been identified during the ELIXIR-EXCELERATE project. These areas include:

- Establishing the Life Science Identity based upon the model adopted within the ELIXIR AAI to be able to federate trans-national service access.
- Driving the emergence of container based workloads within the ELIXIR communities to increase scientific software portability.
- Using GA4GH standards to communicate researcher identity and access between platforms globally, and provide European platform for executing secure cloud based research on human data.
- Supporting establishing a fair trans-national resource allocation process for international life science research by implementing a process to funding analysis activities on academic or commercial clouds available through EOSC.

[34] *https://elixir-europe.org/about-us/what-we-do/elixir-programme*

7.8 Appendix A: Glossary of Key Terms

AAI = Authentication and Authorisation Infrastructure. Processes to verify person who they claim to be and permit to do what they want to do.

AARC = Authentication and Authorisation for Research and Collaboration H2020 project. More information at https://aarc-project.eu/

Ansible = A tool for software remote management. It can be used for software installation, configuration and other necessary actions found on its playbooks.

Availability = Availability is the ratio of time a system or component is functional to the total time it is required or expected to function. This can be expressed as a direct proportion (for example, 9/10 or 0.9) or as a percentage (for example, 90%).

Container = A container middleware is a virtualization layer between the application and the operating system. The containers isolate the runtime environment and allow distribution in the containers. The containers are more lightweight with less overhead than the virtual machine images as they do not include operating system.

CWL = Common Workflow Language: open standards for the description of computational data analysis workflows in a portable and executable manner.

Data Provider = the individual researcher or investigator or body of researchers or investigators that makes data available or submits data for access and use in the context of an ELIXIR Service.

DevOps, Development and Operations = Agile working method to develop eServices. Close cooperation on development and production.

ECP = ELIXIR Compute Platform.

eduGAIN = GÉANT's service that enables trustworthy exchange of information related to identity, authentication and authorisation (AAI).

EGA = European Genome-phenome Archive. The EGA provides a service for the permanent archiving and distribution of personally identifiable genetic and phenotypic data resulting from biomedical research projects. Data Access Committees (DACs) control the access policies. More information at https://www.ebi.ac.uk/ega/home

EGI = A federation of shared computing, storage and data services from national and intergovernmental service providers that delivers sustainable, integrated and secure distributed computing services to European researchers and their international partners. More information at https://www.egi.eu/

ELIXIR Service(s) = refers to ELIXIR Services as defined in the Node Collaboration Agreements, i.e. Node-funded Services or Commissioned Services.

EOSC = European Open Science Cloud. An initiative being driven by the European Commission through the work initially of a High-Level Expert group[35] which has helped to define the scope and direction of the work. The vision is being refined within the community through the EOSCpilot project[36] and will be implemented through the EOSC-Hub project due to start in January 2018. The EOSC-Life project supporting science demonstrators from across the BMS RIs is due to start in March 2019.

Federation = different computing services and/or infrastructures adhering to a certain standard of operation in a collective manner to facilitate its communication and interoperability.

Galaxy = open source, web-based platform for data intensive biomedical research. More information at https://usegalaxy.org/

GÉANT = Interconnects NRENs in Europe. Various services such as identity federation interconnection service eduGAIN. More information at http://www.geant.org/

GitHub = Git is a version control system and GitHub is a service for git based projects. It allows public and private repositories (license costs). GitLab can be used to run a private instance.

GoCDB = EGI's Grid Configuration Database. Contains general information about the sites participating to the production Grid.

GridFTP = High-performance data transfer protocol. Integrated to Grid Security Infrastructure.

HPC/HTC = High Performance Computing, High Throughput Computing. In HTC the tasks are loosely-coupled when an HPC task requires low latency and high performance environment.

IaaS = Infrastructure as a Service, infrastructure level cloud service where the user administer their virtualised hardware such as virtual machines and their network and storage.

IdP = Identity Provider. In addition to the identifier of the user also other user information may be delivered for the Service Provider (SP).

[35] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
[36] https://eoscpilot.eu/

Image = A Virtual Machine image contains operating system and possible other software readily installed. An image is a file with specific format such as raw or qcow2. A conversion might be possible.

IS = Implementation Study. ELIXIR Implementation Studies provide a mechanism by which funds contributed by the ELIXIR Nodes can be used to explore issues around implementing a service. IS are put forward by an ELIXIR Platform to the Hub for consideration for 12-18 months funding. They are a lightweight precursor to possibly build up to a full ELIXIR Commissioned Service application.

Metadata = Metadata contains descriptive, contextual and provenance assertions about the properties of a Digital Object. Makes data findable, usable and documented. Minimally the PID.

NREN = A National Research and Education Network. Provides various level network services.

OpenStack = A cloud middleware to manage the virtualised hardware.

ORCID = Persistent digital identifier for researchers. More information at http://orcid.org/

PaaS = Platform as a Service, readily installed software such as application server to run or develop the applications.

Perun = Identity and access management system developed and run by CESNET. More information at https://perun.cesnet.cz/

PID = A persistent identifier is a long-lasting ID represented by a string that uniquely points to a digital object and that is intended to be persistently resolvable. Used in search, linking and identifying.

Pipelines = A set of data processing elements that are connected in series, where the output of one element is the input of the next one.The elements of a pipeline are often executed in parallel, so several processes happen at the same time and the final result is obtained combining the results of the different processes or stages.

PRACE = Partnership for Advanced Computing in Europe. More information at http://www.prace-ri.eu/

PY = Project Year.

Reliability = refers to the ability of a computer-related hardware or software component to consistently perform according to its specifications. In theory, a reliable product is totally free of technical errors.

Relying parties = refers to a server providing access to a secure software application, or a web site or other entity on the Internet that uses an identity provider to authenticate a user that wants to log-in.

SaaS = Software as a Service, service such as Google Docs. No need to install or administrator any software by the end user.

TUC = Technical Use Case has been defined by the ELIXIR Compute Platform to capture a technical capability that may be repeated (in slightly modified forms) across a number of Scientific Use Cases.

Virtualisation = Layer on top of the physical hardware to allow multiple users to utilise the hardware in a secure manner.

Virtual Machine = Server on top of the virtualisation layer with (guest) operating system which the owner of the virtual machine administrates.