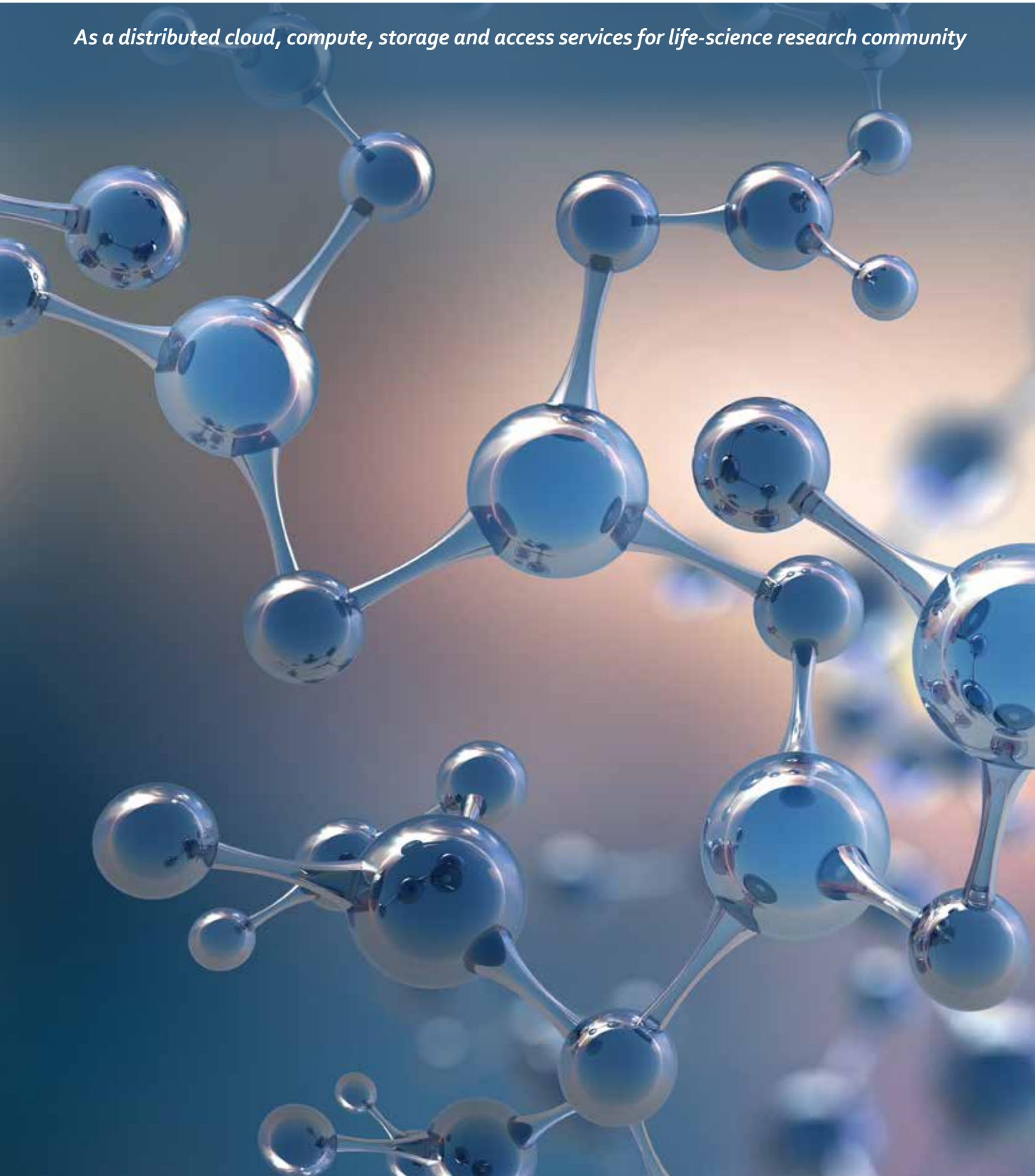
The ELIXIR Compute Platform

As a distributed cloud, compute, storage and access services for life-science research community

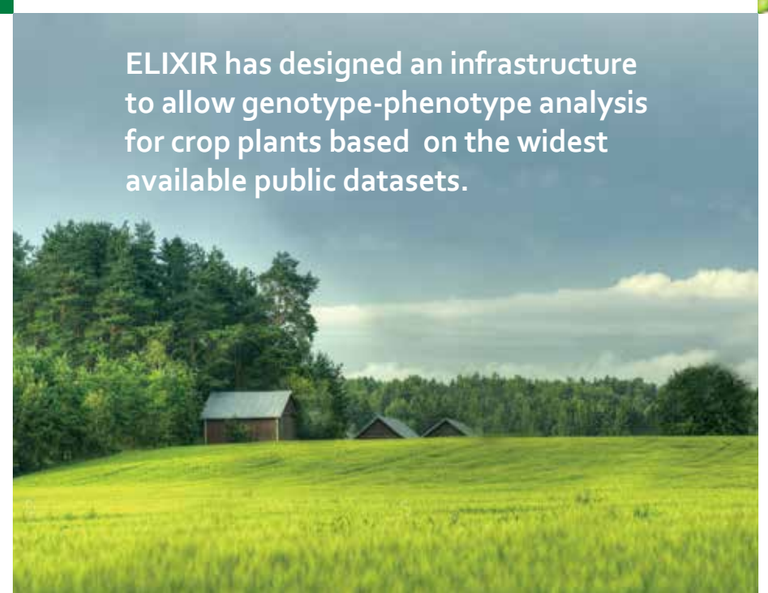


ELIXIR has developed and published a catalogue of rare disease resources. The catalogue is accessible through ELIXIR bio.tools API.

Realisation:
Ari Turunen and Paula Winter
Up-to-Point Ltd.

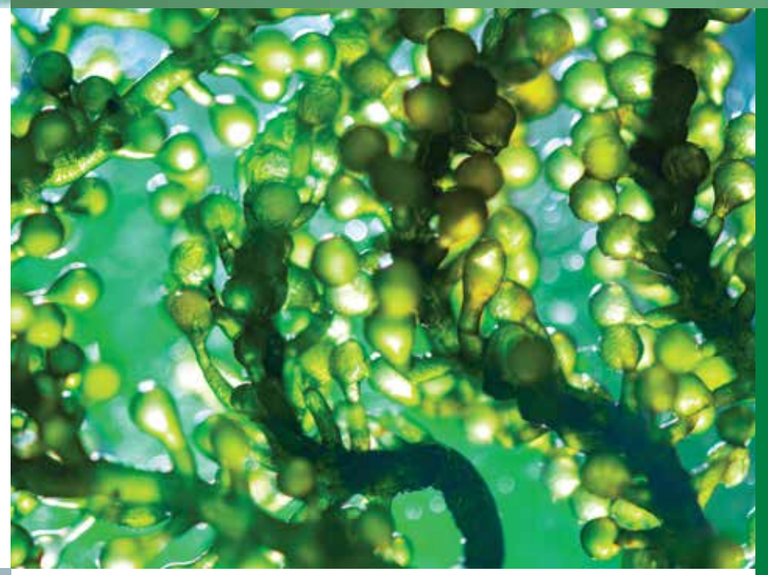
www.elixir-europe.org

ELIXIR constructed sustainable public data resources to improve the characterisation of marine metagenomic samples.



ELIXIR has designed an infrastructure to allow genotype-phenotype analysis for crop plants based on the widest available public datasets.

ELIXIR Compute Platform and Interactions with ELIXIR Scientic Community Use Cases



Researchers can access and process sensitive data from EGA (European Genome-phenome Archive). ELIXIR created a workflow that supports data submitters with data deposition.

INDEX

ELIXIR has built researchers a versatile computing platform that includes a number of important services. Once authorised to use the platform, a researcher can make use not only of the computing features, but also various data sources, in addition to storing, transferring and analysing data. All services are combined into a seamless workflow.

The ELIXIR Compute Platform (ECP) was established in 2015–2019 as a distributed cloud, compute, storage and access services for life-science research community. The ECP flagship infrastructure service is ELIXIR identity and access management ELIXIR AAI, which allows building of an access federation for software, storage, compute and data transfer and availability. In 2019 ELIXIR Compute platform consisted of 80,000 compute cores, 50,000 TB storage and 3,100 users. This represents a doubling of the available storage capacity and a 33% increase in compute capacity over 2 years.

In the 2019-23 ELIXIR will bring ELIXIR data resources to compute services across Europe, and allow scientist to run versatile software analysis workflows with enabling technologies such as container orchestration. This requires a step-up in the service and data access federation of ELIXIR. For instance, all compute resources provided by Nodes are typically used in the Nodes. A pan-European cloud service access for life science envisioned in projects such as European Open Science Cloud EOSC requires a cross-border joint undertaking. The intent of the ECP is to be able to orchestrate access to adequate e-Infrastructure capacities for the life-science research community, in order for Europe to be able to respond to data management challenges such as applying genomics for human health, and supporting molecular biological analysis of the state of the environment.

The work of the ECP brings Nodes and European e-Infrastructures (such as EOSC and Euro-HPC) closer together in the context of ELIXIR scientific communities and training provided by the ELIXIR Nodes. The ECP experts ensure that the technical solutions delivered fit to their specific scientific needs. Each of the scientific communities establishes and manages their standards for describing and accessing datasets, reporting data, matching and comparing content, and eventually building linkages between datasets and their analysis. Those specifications then provide basis for defining the responsibilities of the ECP in the overall research process.

The Communities driving ELIXIR Compute Platform forward are:

Marine Metagenomics

International Transfer of Human Access-controlled Data

Integrating Genomic and Phenotypic Data for Crop and Forest Plants

Integrating ELIXIR Infrastructure for Rare Disease Research

Training

ELIXIR Compute Platform Services

The overall objective is to combine all components of ELIXIR into a seamless workflow. Researchers can securely create a scientific software analysis environment and use this environment to access large biological data resources stored on a cloud. The vision is that ELIXIR Nodes and their collaboration with European e-Infrastructures are the long-term foundation of the European ELIXIR Compute Platform for life science, in compliance with regulations and global technical standards. Moving files (i.e. data) between sites is a one key capability for the ELIXIR Compute Platform. It defines a minimal 'neck' of an hourglass that ELIXIR researchers and application developers can build upon and which ELIXIR Nodes and other Infrastructure Service providers can deploy and support. This has led to a strategy of collaborating with existing initiatives and organisations rather than developing new services.

The ELIXIR Compute Platform offers a geographically distributed Authentication & Authorisation Infrastructure (AAI) as well as Cloud & Compute, Storage and File Transfer Services that are provided by the individual ELIXIR Nodes and are available through ELIXIR.

Authentication and Authorisation Infrastructure

Reliable electronic identification of users (ELIXIR ID) is needed to access the services. ELIXIR Authentication and Authorisation services allow Users to continue using their federated academic, corporate or social media identity by linking it to a personal ELIXIR ID. The ELIXIR service providers connected to ELIXIR AAI will benefit from a centralised user identity and access management services, for example it is possible to establish what level of trust can be applied to the electronic user identity. Another example is when an institutional affiliation of the user changes. The access right coupled to the institutional status of the user are automatically suspended, unless the resource owner decides otherwise.

Cloud & Compute

A cloud and local compute infrastructure is needed to undertake the 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. In 2019 ELIXIR Cloud & Compute 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 All the resources are fully used, and user access in decided by the operating ELIXIR Node.

Storage and Data Transfer

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 to deter ad hoc data transfer and storage

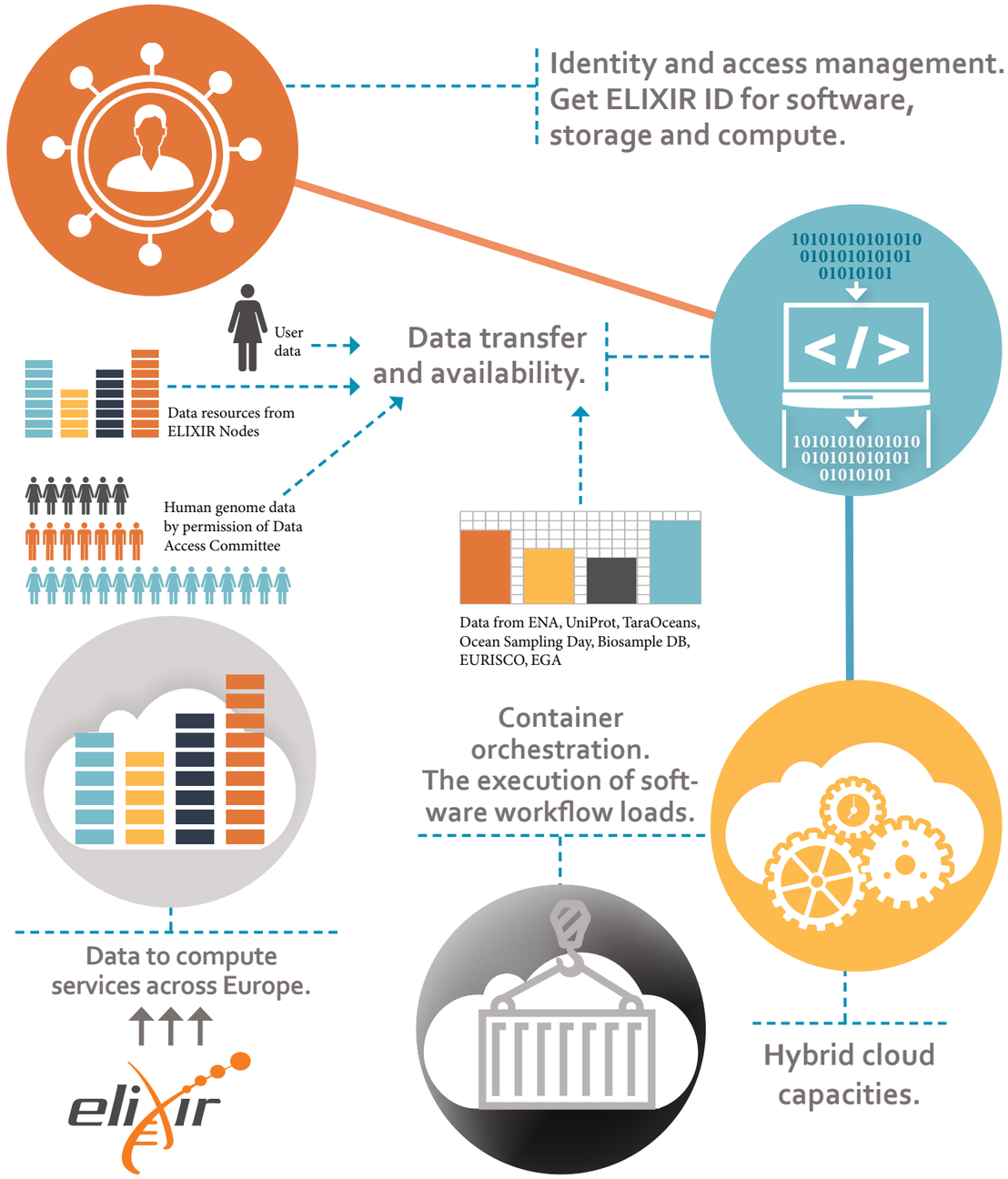
Infrastructure Services Registry

ELIXIR Compute is registering the following cloud resources within the EOSC marketplace for its Compute platform

- EMBL-EBI Embassy Cloud
- ELIXIR CZ MetaCentrum Cloud
- ELIXIR FI CSC ePouta and CSC cPouta

ELIXIR COMPUTE PLATFORM

ELIXIR Compute Platform includes ELIXIR's interlinked user authentication & authorisation infrastructure, storage and data transfer, cloud & computing resources, and Infrastructure Services Registry. For example, a researcher may use the ELIXIR Compute Platform to discover life science friendly compute services, and use their home organisation identity to provision a software analysis environment with access provided from the European Open Science Cloud.



Interactions with ELIXIR Scientific Community

The project 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.

Marine Metagenomics

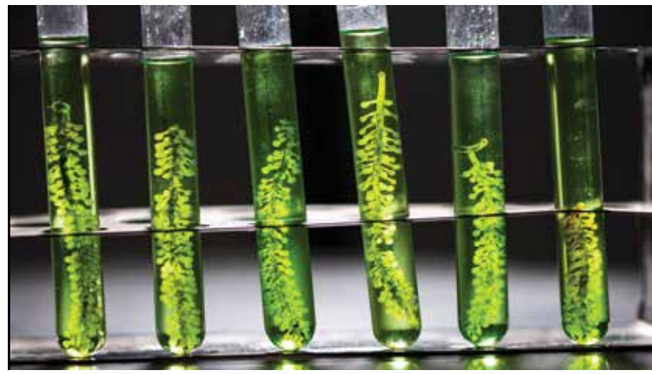
Microbial communities affect human and animal health and are critical components of all terrestrial and aquatic ecosystems. Communities can be exploited e.g. to identify novel biocatalysts for production of fuels or chemicals (bioprospecting), make functional feed for aquaculture species, and for environmental monitoring. The importance of plankton in maintaining the Earth's climate cannot be understated – their communities absorb an astonishing volume of CO_2 from the atmosphere and release oxygen in exchange. Yet only a small fraction of these life forms have been classified and analysed. ELIXIR contstructed sustainable public data resources to improve the characterisation of marine metagenomic samples.

This was achieved by establishing marine microbial databases including reference genomes, nucleotide and protein databases. The established databases, based on the standards developed in the project enhances 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. This work was executed by three ELIXIR Nodes.

Most microbes are very difficult to study outside the context of their communities. Scientists can't grow the microbes and analyse them in isolation. Instead, they must study the whole environment, looking at all the genetic material in a sample and trying to figure out what organisms it came from. Describing microbiomes in detail is difficult.

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 reveals 40 million novel genes and a raft of discoveries about life in the world's oceans. ELIXIR uses some of the high coverage sequence outputs from the TaraOceans to build marine-specific reference databases.

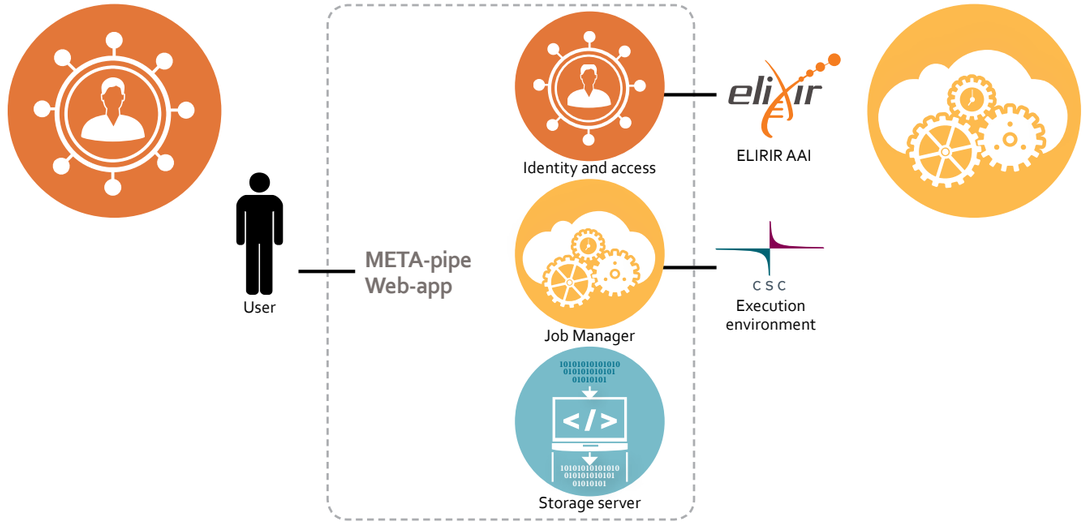
Metagenomics methodologies need to overcome a number of challenges related to standardization, development of relevant databases and bioinformatics tools. New and emerging sequencing technologies, integration of metadata gives an extra burden to the development of future databases and tools. Due to the 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. ELIXIR constructs sustainable public data resources to improve the characterisation of marine metagenomic samples.



ELIXIR uses some of the high coverage sequence outputs from the TaraOceans and Ocean Sampling Day projects to build marine specific reference databases. All datasets are checked with respect to quality, consistency, and interoperability, and in compliance with standards developed in the project. These knowledge-enhanced databases are the cornerstone for sustainable analysis of marine metagenomics sequence data. The databases were 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.

Initially a web based search engine was developed for the interrogation of marine metagenomics results available from the EBI Metagenomics Portal, based on combinations of queries to ELIXIR web services for the discovery of data through metadata, taxonomic and functional fields.

Technical demonstrator M4.1. Marine metagenomics community with the ELIXIR Compute Platform linked user portal tools 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).



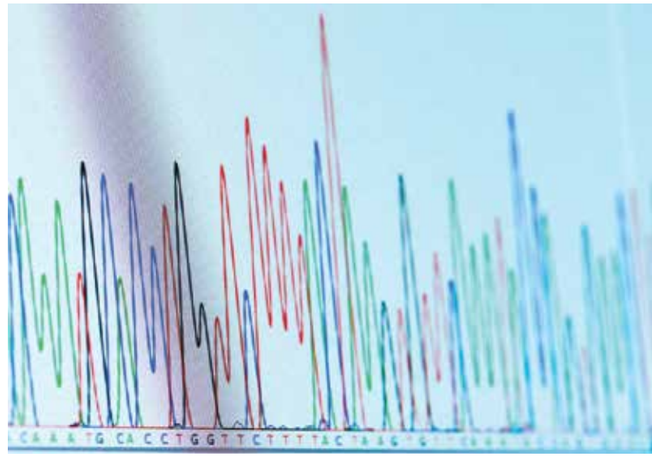
International Transfer of Human Access-controlled Data

ELIXIR supports transfer of large volumes of confidential, electronic, human data, while maintaining appropriate access rights. Researchers can access and process sensitive data from EGA (European Genome-phenome Archive). The archive allows exploration of datasets from genomic studies provided by a range of data providers.

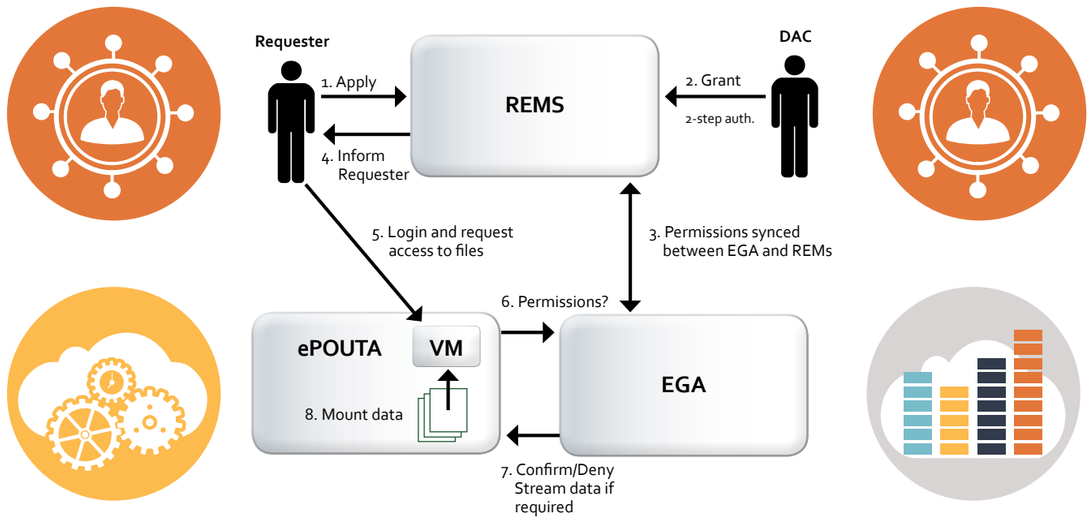
EGA is a repository for all types of sequence and genotype experiments, including case-control, population, and family studies. The archive allows exploration of datasets from genomic studies, provided by a range of data providers. The EGA serves as a permanent archive that preserves several levels of data including the raw data as well as the genotype calls provided by the submitters. A centrally federated human data service allows 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 also provides support for the dataset owners to automate their data access processes (e.g. usable technology, implementing policy, granting permissions.

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 ELIXIR Compute Platform. Researchers identified with ELIXIR AAI are given permissions to access datasets. Access control needs to be enforced on 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.

This use case took care of data deposition to the European Genome-phenome Archive (EGA), enabled data release to authorized individual users from the archive, and to partner downstream secure data analysis on selected compute platform.



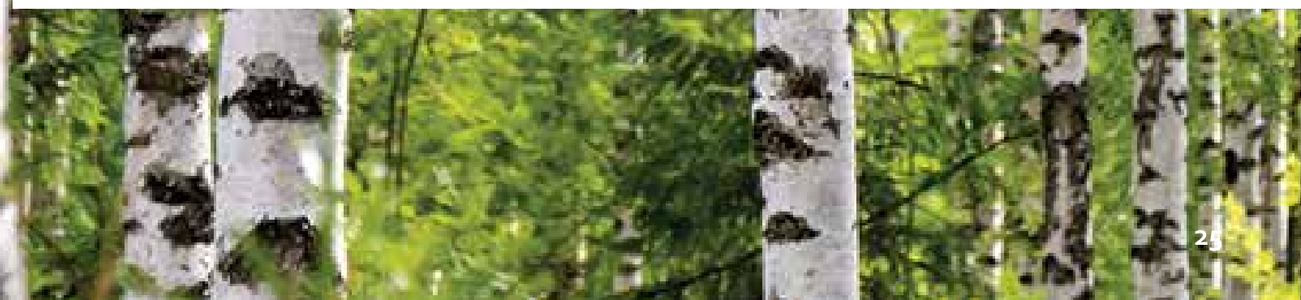
Data ownership and access is maintained in the hands of the original resource owner who has acquired consent from study participants. This workflow also allows resource owners to focus on their unique areas of data generation and analysis expertise while being able to rely on EGA and the ELIXIR infrastructure for their common big genomics data storage, coordination and distribution needs under appropriate legal and data security frameworks.



This case uses the ELIXIR framework for secure submission, archiving, dissemination and analysis of human access-controlled data. The work extends and generalises the system of access authorisation management and high volume secure data transfer developed in the EGA project to address the secure data access needs across ELIXIR resources. A centrally provided service allows authorised third-party services to programmatically check that a user is authorised to access data stored in ELIXIR-coordinated distributed repositories. This also provides support for the dataset owners (e.g. usable technology, implementing policy, granting permissions).

Technical demonstrator M4.2 presents a solution to enforce access to sensitive datasets in a distributed infrastructure setting with secure infrastructure as a service (IaaS) compute clouds.

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.



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. Data is scattered across the laboratories seeking to describe and understand the life of plants.

The plant science community is supported to track and bring these data together, data transfers from geographically distributed sites onto the ELIXIR Compute Platform. Organisations can set up a cloud resource using Elixir ID, deploying storage endpoint virtual machines, and using the Elixir Data Transfer Service to move a set of files to the cloud.

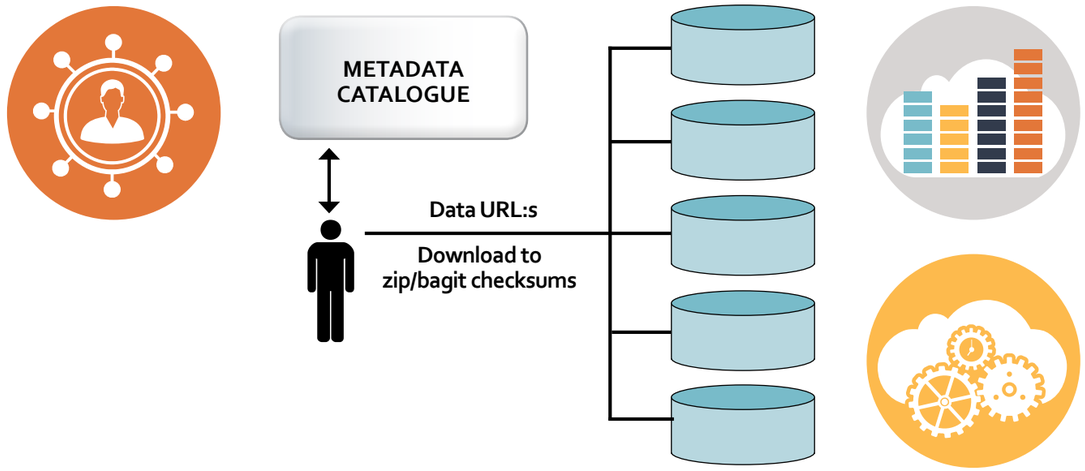
Sequencing and genotyping efforts are likely to accelerate in the near future aiming to catalogue all genetic diversity present in global germplasm resources. However, structural variation in most crop plants is enormous – more so than in humans. Phenotypic characterisation of data is often inaccessible, diverse and non-standard. 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.

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.

ELIXIR Nodes establish common guidelines for ontology usage when annotating crop and forest species. Sample identification will be handled through the BioSample DB at 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 develop a common API for data query and retrieval.

Working on exemplar species, ELIXIR Nodes established a sustainable model for the interaction of distributed phenotypic repositories with defined genomic and sample reference data. The model is scalable, distributed, and transparently integrated through the development and use of common vocabularies and search technologies. This was done by using established repositories for genomic data and sample metadata. This accelerates research and plant breeding through the exploitation of an interoperable commons of public data. ELIXIR Nodes also work on establishing common guidelines for ontology usage when annotating crop and forest species

Technical demonstrator M4.3. supports the plant science community to track data, and data transfers from geographically distributed sites onto a scalable ELIXIR Compute Platform server. This scientific use case integrates genomic and phenotypic data of crop and forest plants from a variety of open access and open data sources. These data sources 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 need to achieve this functionality include cloud & compute, catalogue of dataset and other (persistent) identifiers, file transfer and ELIXIR AAI. Also a distributable storage endpoint virtual machine was developed. It uses ELIXIR Compute Platform Data Transfers technologies to move a set of files to the target cloud service.



SEARCH

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 use case addressed the data integration needs of the rare diseases community.

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.

ELIXIR has developed and published a catalogue of rare disease resources. The catalogue is accessible through ELIXIR bio.tools API. Rare disease researchers' submit their raw data, run the mapping and obtain unannotated gvcf files (genomic variant call format) for analysis through ELIXIR bio.tools API.

This use case is based around supporting research around rare (1 in 2000 people) chronic or genetic diseases that uses EGA as its data sources – access to which is controlled. The metadata around a patient (i.e. their illness, treatments, outcomes), patient samples stored in a biobank, and any sequenced material stored in EGA is searchable through a central portal which can only be accessed by authorised users.

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.

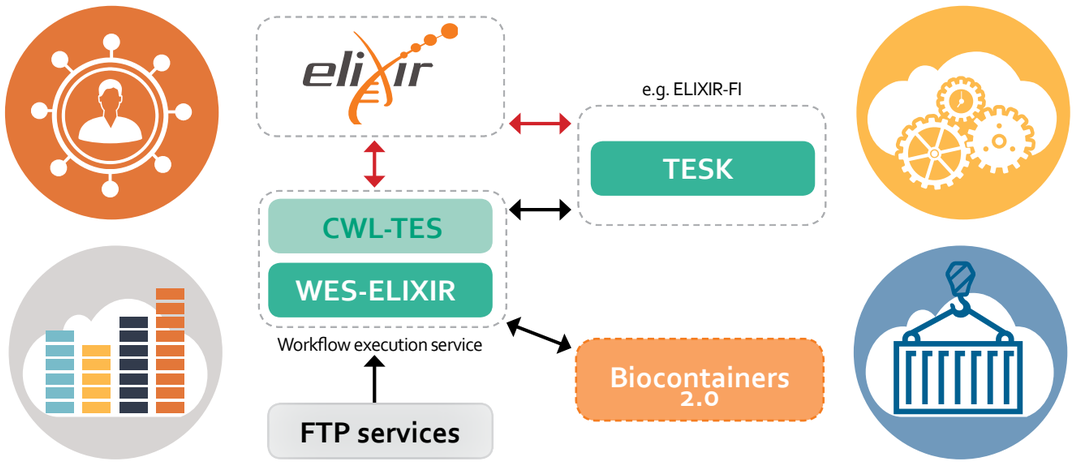
This use case addressed the data integration needs of the rare diseases community. ELIXIR created a customized portfolio of tools and services devoted to assist in the development of new therapies. Portfolio includes the registry of data resources and analysis tools.

There are a wide range of data resources and analysis methods provided by ELIXIR Nodes. For example the European Genome-Phenome archive (EGA) stores data from major research initiatives in rare diseases. ELIXIR reviewed current data resources and evaluated their usability and potential impact on the rare disease community. An important aspect of the evaluation was the security of the data that is a key aspect in rare disease research given the low frequency of the associated genomic variants in the population.

Technical demonstrator M4.4. with the rare diseases allowed a researcher to submit their local raw data files, process the raw genomic data using the RD-Connect software 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.

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 adapts 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.



Training

How ELIXIR bioinformatics courses could use cloud resources? A cloud provider survey was conducted, 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.

- Introduction to Metabolomics Analysis
- Good practice in high-throughput experiments
- Lipidomics MS Data processing
- Ligand-protein docking, and computer-aided drug design
- Python for Life Scientists
- etc.

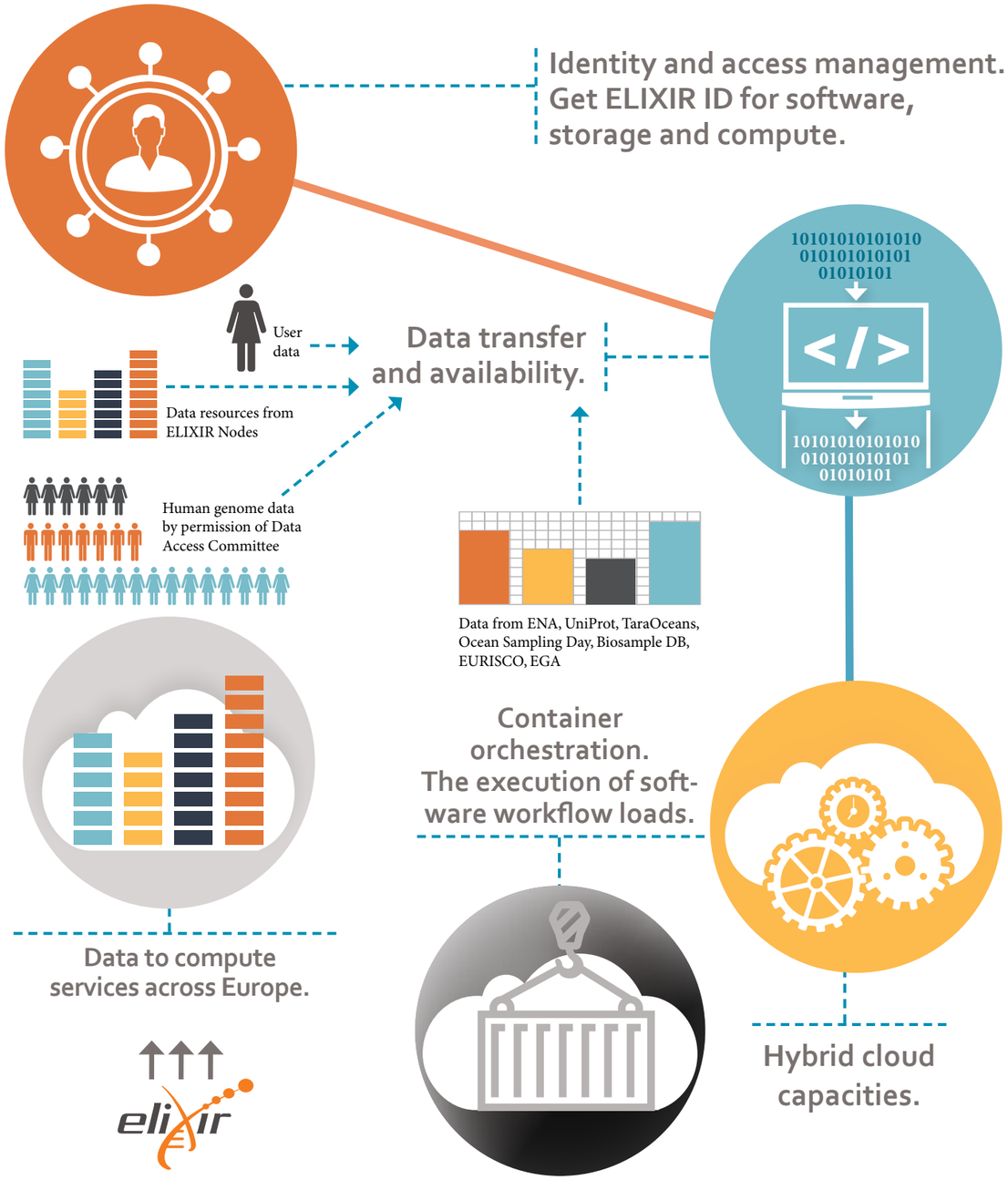
The bottlenecks to leverage distributed infrastructure of ELIXIR for Training can be eased in several ways. In general, evaluating and selecting suitable cloud providers, requires a technical expert. Some ELIXIR Node cloud and compute services already provide easy-to-use interfaces where participants 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. Trainers 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 the 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.

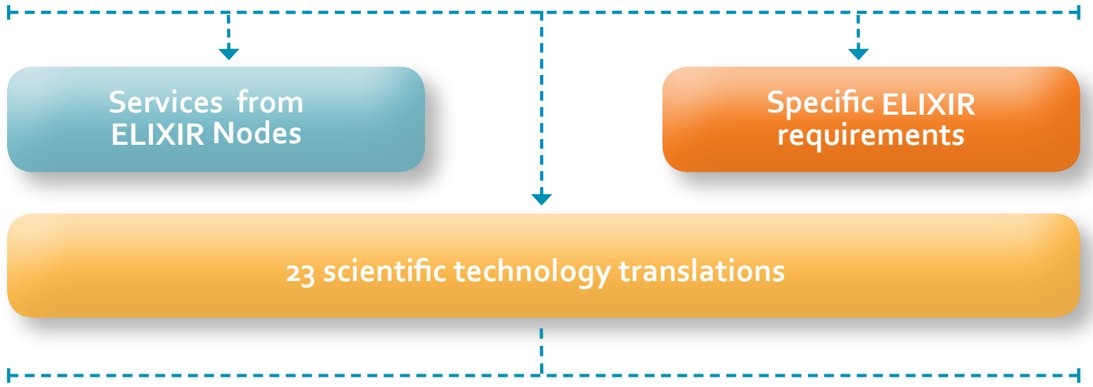
As a conclusion, recommendation was that the ELIXIR Hub could appoint a technical evaluation panel and provide an annual budget to get and support utilisation of distributed cloud resources for key ELIXIR bioinformatics training events.

ELIXIR COMPUTE PLATFORM

ELIXIR Compute Platform includes ELIXIR's interlinked user authentication & authorisation infrastructure, storage and data transfer, cloud & computing resources, and Infrastructure Services Registry. For example, a researcher may use the ELIXIR Compute Platform to discover life science friendly compute services, and use their home organisation identity to provision a software analysis environment with access provided from the European Open Science Cloud.



TECHNICAL USE CASES MAKE ELIXIR COMPUTE PLATFORM HAPPEN



Federated ID. Provides means for individuals to identify themselves with different levels of assurance using their institutional credentials linked with ELIXIR Identity.

Other ID. Use of internet identities (e.g. Google, Facebook, ORCID) with different levels of assurance linked to ELIXIR Identity to gain access to services.

ELIXIR ID is used as the basis for accessing ELIXIR services.

Cloud IaaS Services. Provides information needed by users to gain access to national or regional ELIXIR Cloud services.

File Transfer. Supports movement of files between authenticated locations by command line, web service or web page.

Infrastructure Service Directory. Provides human-readable and machine-accessible technical and contact details of ELIXIR compute services.

Credential Translation. Converts ELIXIR ID into a credential on demand. For instance, a federated identity could be converted into a short-termgrid proxy.

Service Access Management. Manages access rights for user groups, group membership and attributes for allocated resources. A principal user can, for example, create a group, add/remove members to the group and grant the group access to a specific service.

Virtual Machine Library. Virtual machine image library of software environments that are compatible with cloud services and typically updated by scientific software service experts.

Container Library. A source of containers of common (scientific) software components. Containers can be deployed to software environment on virtual machines or servers.

Data Set Replication. Replicates Data Sets between major centres upon data set release.

Endorsed Personal Data or Compute Access Management. A process to give entitlement to authorise user to access a specific service (e.g. scientific application review, phone number verification).

Federated Cloud IaaS. Standard where a user can gain access to multiple cooperating cloud services through a single access decision.

Operational Integration. Compute services (federated ID, cloud, storage, etc.) of ELIXIR and their dependencies are monitored as a whole to ensure service availability.

Resource Accounting. View of consumption of services (e.g. CPU time, service invocations, storage, data sets) by individual users, projects/groups across different services.

HTC/HPC Cluster. Provides information needed by users to gain access to an High Throughput Computing or High Performance Computing services.

PRACE Cluster. Links ELIXIR (e.g. data resources and users) with some of the PRACE services for Highest Performance Computing in Europe.

Network File Storage. Provides network accessible non-local storage space where an authenticated user can retrieve or store a file.

Module Library. A library of modules of common software components.

Infrastructure Service Registry. A registry of currently available infrastructure services available for use that matches the Infrastructure Service Directory.

Cloud Storage. Storage attached to virtual machines running in cloud services.

PID and Metadata Registry. Service that links a PID (Persistent Identifier) to metadata relating to a data file/set. The same data file/set may be registered with multiple physical locations under the same PID.

Federated HTC/HPC Cluster. Standard where a user can gain access to multiple cooperating compute cluster services through a single access decision.

ELIXIR receives funding from the European Commission within the Research Infrastructures Programme of Horizon 2020.

Contact us at **contact-compute@elixir-europe.org**

www.elixir-europe.org