In the past decades, biological and medical research has turned data-intensive. Today, thousands of science laboratories across the world generate massive amounts of data every day.

INTEGRATING ELIXIR INFRASTRUCTURE FOR RARE DISEASE RESEARCH
According to EURORDIS about 30 million people in the EU have a rare disease. ELIXIR Compute platform helps to link data from rare disease patients between different countries and health care institutions.

HELP FOR DIAGNOSTICS: HUMAN ACCESS-CONTROLLED DATA
ELIXIR creates a workflow that supports data submitters with data deposition.

MARINE METAGENOMICS
Microbial communities affect human and animal health and are critical components of all terrestrial and aquatic ecosystems.

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.
In the past decades biological and medical research has turned into a data-intensive science. Today, thousands of laboratories across the world generate massive amounts of data. Growing complexity is another major challenge – there are around 1,700 different biological databases, often with different formats, annotations and standards.

In this situation, the traditional ‘download-and-compare’ method is no longer viable – much of the data need to be managed as a federation. Within this federation, however, there has to be a robust technical infrastructure to access, transfer, exchange and analyse the data.

To address this challenge, the ELIXIR Compute Platform was established in 2015 to build distributed cloud, compute, storage and access services for the life-science research community. The platform is working closely with four scientific communities (ELIXIR Use Cases) to ensure the technical solutions fit their specific needs. Each of the scientific use cases will establish and manage standards for describing and accessing datasets, reporting data, matching and comparing content, and eventually building linkages between datasets. Those specifications then provide the basis for the work of the ELIXIR Compute Platform.

The objective is to combine all components of the ELIXIR Compute services into a seamless workflow: A researcher may 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.
The ELIXIR Nodes and their collaboration with European e-Infrastructures form the foundation of the ELIXIR Compute Platform. Moving files (i.e. data) between sites is one key capability of the ELIXIR Compute Platform. It defines an agreed set of standards 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 strategy is to build on existing services. The ELIXIR Compute Platform has already established a geographically distributed Authentication & Authorisation Infrastructure (AAI). It has also linked Cloud & Compute, Storage and File Transfer Services that are provided by the individual ELIXIR Nodes and which will be discoverable through ELIXIR.
**Authentication and Authorisation Infrastructure**

To facilitate access some of the ELIXIR resources (e.g. sensitive human data), a reliable electronic identification of users is needed to make sure the services are used only for legitimate research. The ELIXIR Authentication and Authorisation provides a single sign-on for multiple ELIXIR services and allows users to access these services using their academic, corporate or their ORCID account.

The ELIXIR service providers connected to ELIXIR AAI will benefit from centralised user identity and access management services. This will enable service providers to establish different levels of trust for different types of electronic user identity. When an institutional affiliation of the user changes, the access rights coupled to the institutional status of the user will be 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 services are possible and ideal for users that require high performance, high security and certified environments for sensitive human data handling, for instance.

**Storage and Data Transfer**

Data transfers are essential across all scientific Use Cases. The Compute Platform has investigated various mechanisms to organise data transfers between core biological data resources. ELIXIR’s Storage and Data Transfer services will include:

- Data replication and data submission to or from ELIXIR Data Resources
- Services to pull relevant datasets from ELIXIR 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. The aim is to decrease network overload for ELIXIR Nodes hosting large datasets avoid ad hoc data transfer and storage

**Infrastructure Services Registry**

An ELIXIR Infrastructure Service Registry will be deployed to provide a live picture of the technical capabilities of the ELIXIR Compute Platform. This registry will complement other registries in ELIXIR such as the ELIXIR Tools and Service Registry bio.tools. The information in the Infrastructure Service Registry will present all information needed to find and select a specific technical capacity needed for life science research: basic data such as contact URL or physical capacity, as well as technical data about the service capability (e.g. free CPUs, free storage). Information in the registry can be used to select which specific technical capability to use.

*See pages 26–27 for more details.*
Microbial communities affect human and animal health and are critical components of all terrestrial and aquatic ecosystems. They can be exploited for purposes such as environmental monitoring, identifying novel biocatalysts for production of fuels or chemicals (bioprospecting), and making functional feed for aquaculture species. ELIXIR brings together relevant databases and tools for metagenomics methodologies. This research community uses the ELIXIR Compute for identification of researchers, and receive improved data analysis and submission capabilities for their scientific pipelines. These pipelines give structure to data e.g. about bacterial life.
ELIXIR will use some of the high coverage sequence outputs from the TaraOceans and Ocean Sampling Day projects to build marine-specific reference databases. All datasets will be checked with respect to quality, consistency, and interoperability, and in compliance with standards developed in the project. These knowledge-enhanced databases will be the cornerstone for sustainable analysis of marine metagenomics sequence data. 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.

Metagenomics methodologies need to overcome a number of challenges related to standardisation, development of relevant databases and bioinformatics tools. New and emerging sequencing technologies and integration of metadata impose an extra burden on the development of future databases and tools. Due to the data biases of existing reference databases, only about one quarter of sequences have been annotated, and this fraction will diminish further when more diverse samples such as soil and marine are analysed. ELIXIR will construct 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 the European Nucleotide Archive (ENA) as part of the International Nucleotide Sequence Database Collaboration, UniProt and other publicly available datasets. This work will be executed by three ELIXIR Nodes.

The tools and the pipelines for the identification of gene products (e.g. enzymes and drug targets) will be implemented and made publicly available. Initially, a web-based search engine will be 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.
Massive sequencing and genotyping of crop and forest plants and their pathogens and pests generates large quantities of genomic variation data. ELIXIR is designing 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 understand the life of plants at the molecular level. ELIXIR Compute supports the community to track and bring these data together, which enriches data analysis capabilities – local data can be interpreted in the global context.
ELIXIR Nodes will analyse 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 understanding 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 annotation and interface standards. 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.

In order to establish a model applicable to all species, at least three exemplar species will be chosen, including one cereal species, one woody species, and one other crop species. The model will be 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 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.

Sequencing and genotyping efforts are likely to accelerate in the near future with the aim of cataloguing all genetic diversity present in global germplasm resources. However, the 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. Seven ELIXIR Nodes are working on a technical infrastructure and defining an open model for the publication and sharing of plant genotype-phenotype data.
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. ELIXIR will build a registry of data resources and analysis tools critical for the development of the rare disease research. This research community needs an access to molecular level information about rare diseases across many European institutions that host and collect these data. Secure access processes are provided, and data will be integrated to complex and evolving scientific software stacks hosted on secure cloud services.
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 will review current data resources and evaluate their usability and potential impact on the rare disease community. An important aspect of the evaluation will be the security of the data, which is a key aspect in rare disease research given the low frequency of the associated genomic variants in the population.

The International Rare Diseases Research Consortium established the ambitious goal of developing 200 new therapies by 2020. ELIXIR is aligned with this effort. This use case will address the data integration needs of the rare diseases community. The aim is to interface and empower ongoing and future rare disease research projects by addressing data interoperability and management bottlenecks. ELIXIR will create a customised portfolio of tools and services devoted to facilitating the development of new therapies. The portfolio includes the registry of data resources and analysis tools.

This use case is based on supporting research in rare (1 in 2000 people) chronic or genetic diseases that uses EGA as its data source – access to which is controlled. Patient-related metadata (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 that 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.
ELIXIR creates a workflow that supports data submitters with data deposition to the archive and data release to authorised individual users from the archive, and to partner downstream secure ELIXIR data analysis platforms. Data ownership and access is maintained in the hands of the original resource owner. Security is a priority for data transfers and target cloud infrastructures involving human data. ELIXIR Compute supports the European Genome-phenome Archive and the data custodians like biobanks to create a secure platform that can be trusted.

Help for Diagnostics: Human Access-controlled Data
Data ownership and access is maintained in the hands of the original resource owner who has acquired consent from study participants. This workflow will also allow 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.

The use case focusing on human access-controlled data will use the ELIXIR Framework for secure submission, archiving, dissemination and analysis of human access-controlled data. The work will extend and generalise 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 will allow authorised third-party services to programmatically check that a user is authorised to access data stored in ELIXIR-coordinated distributed repositories. This will also provide support for the dataset owners (e.g. usable technology, implementing policy, granting permissions).
The ELIXIR Compute Platform includes ELIXIR’s interlinked authentication and authorisation infrastructure, storage and data transfer, cloud and computing resources, and Infrastructure Services Registry. For example, a researcher may use the ELIXIR Compute Platform to discover cloud services from the Infrastructure Services Registry and use their ELIXIR Identity to provision a software analysis environment on a European Cloud Infrastructure for science (e.g. egi.eu).

**ELIXIR COMPUTE PLATFORM**

The ELIXIR Compute Platform includes ELIXIR's interlinked authentication and authorisation infrastructure, storage and data transfer, cloud and computing resources, and Infrastructure Services Registry. For example, a researcher may use the ELIXIR Compute Platform to discover cloud services from the Infrastructure Services Registry and use their ELIXIR Identity to provision a software analysis environment on a European Cloud Infrastructure for science (e.g. egi.eu).

**Get ELIXIR ID for software, storage and compute**

- Transfer data (from 4 scientific cases) to platform.
- Some ELIXIR data resources are available on the platform.

**Cloud & Compute provided by ELIXIR-nodes from 20 countries**

- Human genome data by permission of Data Access Committee
- Data from ENA, UniProt, TaraOceans, Ocean Sampling Day, Biosample DB, EURISCO, EGA
- ELIXIR Compute Platform includes ELIXIR's interlinked authentication and authorisation infrastructure, storage and data transfer, cloud and computing resources, and Infrastructure Services Registry.

**Scientific software environment**

- Virtual services and cloud storage
- 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 authorised users to access a specific service (e.g. scientific application review, phone number verification).

**TECHNICAL USE CASES MAKE ELIXIR COMPUTE PLATFORM HAPPEN**

**Services from ELIXIR nodes**

- 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-term grid 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 authorised users 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.

**Technical Use Cases (TUCs) for upcoming implementation**

- **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.

**Specific ELIXIR requirements**

- **23 scientific technology translations**

**23 scientific technology translations**

- **Services from ELIXIR nodes**

- **Specific ELIXIR requirements**

- **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.

**Technical Use Cases (TUCs) for upcoming implementation**

- **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.