ELIXIR Compute Platform for life and health sciences

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 built for biomedical needs between 2015 and 2019. ECP is a geographically distributed platform in which ELIXIR centres act in concert to provide services for the management of biological data. The centres operate independently, but are brought together by the Authentication and Authorisation Infrastructure (AAI) with which cloud services, and computing, storage and data transfer services can be coordinated. Researchers log into the system, which checks their electronic identity and allows the appropriate level of access to biomedical data. Researchers can then create a secure analysis environment for their software. Data is stored on the European compute cloud platform. The operating environment also helps groups of researchers to create scalable services.

Thousands of research laboratories create massive amounts of data. This data is also becoming more complex, which poses a major challenge. Data must be managed so that all users understand and handle it in the same way. Effective data management requires a federation that manages the infrastructure within which the user can transfer, exchange, process and analyse data. This is why the ECP was developed in cooperation with ELIXIR centres and European research infrastructures. Services designed for ECP researchers were jointly built by four scientific user communities studying marine microbes; cultivated and forest plants; human genes; and rare diseases.

Services within the ECP are offered by various ELIXIR centres. ELIXIR's AAI service (Authentication and Authorisation Infrastructure) enables electronic user identification and the granting of access rights. Access to data is always decided by the owner of the data or computing service, but AAI will make access to data faster, and the data use policy and analysis are clear and straightforward.

A high-capacity network is used for data transfer, and software is used to build interfaces on top of it, like data pipelines. These handle data transfer, processing and analysis. The data flows are divided into smaller parts and are processed in parallel to increase computing power, enabling the transfer to occur without bottlenecks and delays. Analyses can be performed in a distributed manner. If the data is sensitive, data security federation is required.

In 2019, the ECP had a storage capacity of 50,000 terabytes. It provided 80,000 separate computing cores, that is, processing units. Between 2017 and 2019, its storage capacity doubled and the distributed computing resources increased by 33%. In 2019, the ECP had 3,100 users.

Marine metagenomics

Microbe communities affect the lives of humans and animals and play an important role in various ecosystems. However, only a small proportion of microbes have been categorised and analysed. Study of the genetics of microbe communities has created a new field in biosciences, metagenomics. A group of genes collected from the environment and then sequenced can be analysed in the same way as the genome of an individual species.

The oceans are the world's largest single ecosystem. Plankton is at least as important to the world's climate as the rainforests are. However, only a small number of the organisms that create this ecosystem have been categorised and analysed. Ecosystems formed by plankton contain a huge amount of life: there are more than 10 billion organisms in each litre of ocean water, consisting of viruses, procaryotes, single-cell eucaryotes and cnidarians. These unique organisms contain bioactive compounds that are useful in the pharmaceutical industry, as food, in cosmetics, and in bioenergy and nanotechnology. Between 2009 and 2013, an international research expedition called Tara Oceans collected 35,000 biological samples from 210 locations in the world's oceans. It is the largest plankton collection in existence.

ELIXIR built a permanent, public data resource to identify and chart metagenomics samples obtained from the sea. The tools needed for identification and the pipelines for data processing were made available for transfer to different platforms. This could result in the introduction of new biochemical materials, such as enzymes and medicinal molecules. The tools and data pipelines can be used through various ELIXIR centres (Norway, EMBL-EBI, Finland, Czech Republic and France).

Controlled cross-border transfer of human data

The European Genome-phenome Archive (EGA) is one of the world's most extensive public data resources and holds patient data gathered from biomedical projects. The archive contains various data resources from different data producers. The EGA collects human genome and phenome data on the basis of the consent of the persons involved. ELIXIR Compute Platform enables the transfer of confidential human data via the EGA to individual users authorised to access such data.

Through the ECP, researchers can access the EGA's sensitive data collections. First the user is identified electronically, and access is either granted or rejected on the basis of the

information on the application form. If the service requires multi-factor authentication, the user is redirected to an identification service, which performs an extra authentication by means of another security feature.

After this, researchers have access to EGA data resources and can process sensitive data. Through the ECP, researchers can also store data in the EGA archive. The ECP can ensure data description, access to data and compatibility. In order to transfer data securely, an architecture was created that uses two protocols. Oauth.2.0 and OpenID Connect (OIDC) are user identification protocols used in the industry.

Data integration of genome and phenome data of cultivated and forest plants

According to the FAO, plant diseases cause an annual loss of 20–40% in global food production. Massive sequencing of cultivated and forest plants enables the causes of plant diseases to be studied. Plant sequencing and genotyping, including pathogens and diseases, generate large amounts of genetic variation data. EURISCO (European Search

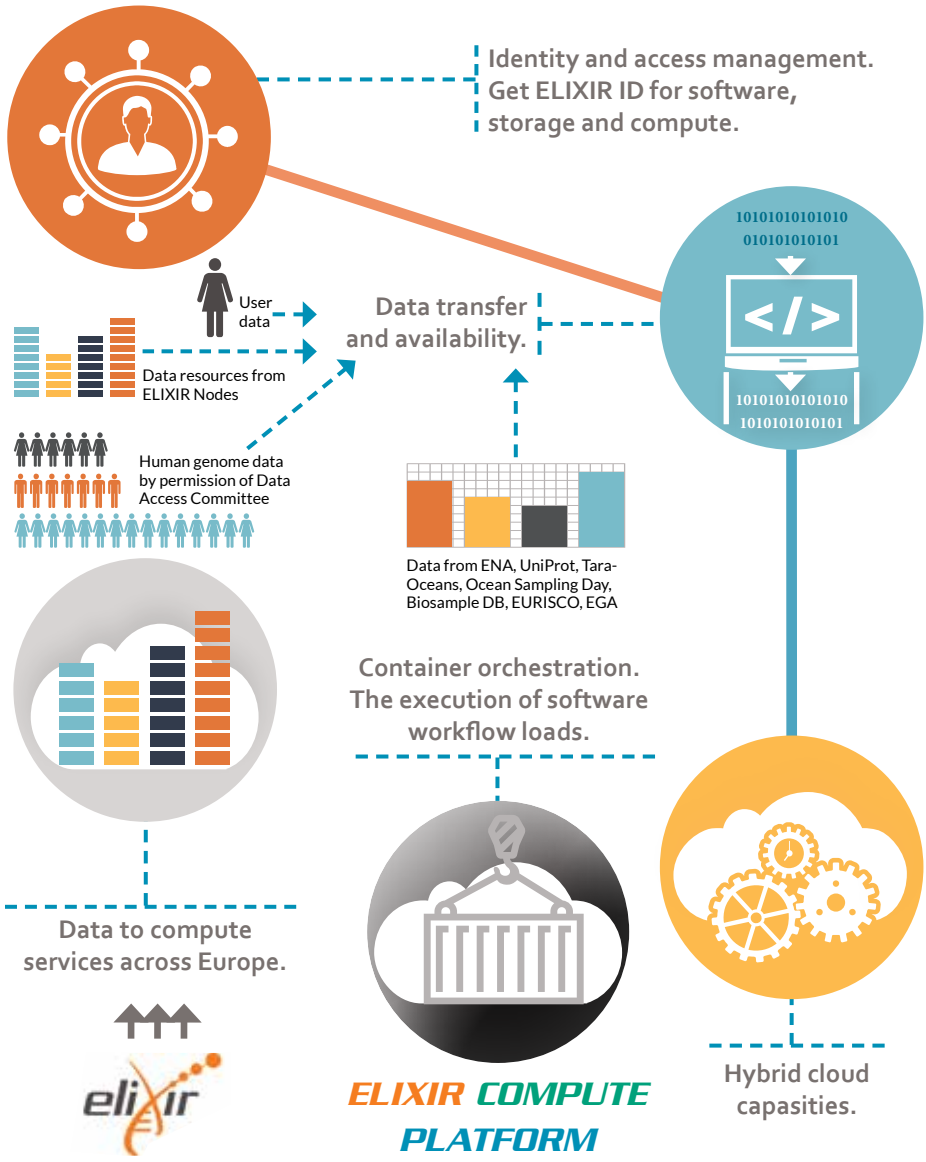
Catalogue for Plant Genetic Resources) contains information about 1.9 million cultivated plants and their wild relatives. The samples have been collected by almost 400 organisations.

The ECP enables genotype-phenotype analysis of cultivated plants, based on the widest available public data resources. This data has been assembled from geographically separate research institutions. The key function is a search robot that receives searches from users, to whom it transfers integrated search results obtained from various data sources. Users can transfer the selected data into the cloud infrastructure for analysis.

Integration of ELIXIR infrastructure for the study of rare diseases

The European Organisation of Rare Diseases (EURORDIS) estimates that some 30 million people in 25 EU countries have a rare disease. This translates into 6–8% of all

Transfer of data files between locations is one of the key functions of the ECP. There are currently 1,800 different biological databases that store data in various formats and standards, and describe and categorise stored data in various ways. Data obtained from various sources is presented in the ECP accordingly. A high-capacity network is used for data transfer, and software is used to build interfaces (data pipelines) on top of this. These handle data transfer, processing and analysis.



people in the EU. The International Rare Diseases Research Consortium has set itself the target of developing 200 new forms of treatment for rare diseases by 2020.

ELIXIR has published a customised collection of tools and services to help in the development of new treatments. The collection is available through the ELIXIR biotools service (bio.tools), which is part of the ECP platform. Researchers of rare diseases can deposit raw data, run a gene mapping and pick gvcf files (genomic variant call format) for analysis. This defines the text file used in bioinformatics when gene sequence variations are stored.

The patient's metadata (illness, treatment, treatment results), patient samples in biobanks and all EGA can be obtained through the ECP.

Ari Turunen

ADDITIONAL INFORMATION:

CSC – IT Center for Science
is a non-profit, state-owned company administered by the Ministry of Education and Culture. CSC maintains and develops the state-owned, centralised IT infrastructure.
http://www.csc.fi
https://research.csc.fi/cloud-computing

ELIXIR
builds infrastructure in support of the biological sector. It brings together the leading organisations of 21 European countries and the EMBL European Molecular Biology Laboratory to form a common infrastructure for biological information. CSC – IT Center for Science is the Finnish centre within this infrastructure.
http://www.elixir-finland.org
http://www.elixir-europe.org

SUOMEN ELIXIR
Puh. +358 9 457 2821 ● e-mail: servicedesk@csc.fi
www.elixir-europe.org/about-us/who-we-are/nodes/finland

www.elixir-finland.org

ELIXIR PÄÄMAJA
EMBL-European Bioinformatics Institute
www.elixir-europe.org