

Sensitive data infrastructure

Sharing biomedical data collected of humans is a prerequisite for disease prevention and treatment in the modern world. The Finnish ELIXIR node CSC is building an infrastructure in which human data obtained from Finland's biobanks and research organisations has been pre-processed and described and saved in a secure way. The parties responsible for sharing the data can automate their authorisation process with the CSC platform. This improves licensed availability of data for research and healthcare purposes.



Personalised drug treatments are only possible if patient data is available and it has been stored and pre-processed correctly. In a project funded by the Academy of Finland, an infrastructure is created that meets the requirements for storing and using sensitive data. The data consists of clinical register data, genomic data and material related to bioimaging. The project is participated in by not only CSC but also the bioimaging infrastructure Euro-Biolmaging, THL Biobank and the Institute for Molecular Medicine Finland (FIMM).

The project creates solutions to facilitate quick and easy access to various data for by researchers. The data can be stored in CSC's sensitive data infrastructure. Researchers are allocated a space in which the data and computing power are in the same place. Researcher can only access data to which authorisation has been obtained by the data owner. The project also makes use

of federated data management developed by CSC. ELIXIR AAI and REMS are applications developed by CSC for the managing users in the ELIXIR infrastructure.

Secure transfer of data will revolutionise healthcare in the next decades. The project supports researchers developing artificial intelligence algorithms by offering them computing services, more advanced research use of health data, and data management technologies. The data material's compatibility with international standards is also verified.

Secure pre-processing of genomic data security

The sequencing capacity of the Institute for Molecular Medicine Finland and Helsinki University Central Hospital is improved with a direct connection to CSC's computing and data services. Genomic data is transferred to CSC on a superfast and secure optic cable.

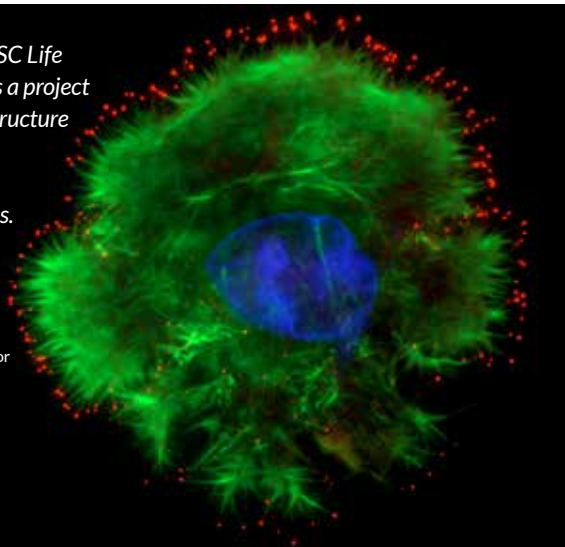
Data pre-processing and quality assurance are fast, because the data is located at CSC.

As the sequence data is physically closer to the computing services, the pre-processed data will be available to the researcher more quickly. The capacity can be used to sequence exomes, genomes and transcriptomes efficiently.

Combining genetic and clinical data still requires a lot of data storage and computing capacity. The European HPC Center of Excellence for Personalised Medicine (PerMedCoE), a joint project by CSC and the Barcelona Supercomputing Center, brought the data analysis methods of personalised medicine into the supercomputer environment. Algorithms developed in the project can significantly reduce the computing time required for analysis. Analysis of genetic and protein data is becoming faster, facilitating and speeding up disease diagnosis and identification of the appropriate treatments.

Breast cancer cell visualised. EOSC Life (European Open Science Cloud) is a project coordinated by the ELIXIR infrastructure with the objective of offering all European researchers a wide selection of bioindustry IT services. Its purpose is to integrate various federal infrastructures and data services.

Picture: Guillaume Jacquemet, Turku Centre for Biotechnology, Ivaska Laboratory



Disease diagnosis by utilising molecular biology can in future be done within hours or days.

Bioimaging material and artificial intelligence algorithm

CSC, together with the Finnish biobanks, the National Institute for Health and Welfare and Euro-Biolmaging, which operates from Turku, are developing an artificial intelligence algorithm for mining medical data. Euro-Biolmaging Finland offers image storing services and data services, such as image collections. Terabytes of images have been stored in the collections, and these can be used as reference data, for example. The material ranges from plankton imaging to cancer cells.

Euro-Biolmaging Finland also offers medical imaging material. Free access to imaging services is provided by six universities and three university hospitals in Finland. These use Open Microscopy Environment (OMERO) services, enabling researchers to view, organise, analyse and share material from anywhere with internet access.

"Turku already has two new OMERO services in production use for image data, one for research and the other for teaching purposes. Both also serve, to a limited extent, the entire country. Now would be a good time to plan how these could be linked with CSC services," says Pasi Kankaanpää, Senior Scientific Manager at Euro-Biolmaging.

Kankaanpää has submitted articles to the Nature Methods publication concerning recommendations for image data management and its metadata.

"This increases cooperation and also emphasises the importance of managing sensitive data. Data management and processing are key aspects at Euro-Biolmaging Finland – and indeed what this project funded by the Academy of Finland emphasises," says Kankaanpää.

Use of national biodata for research

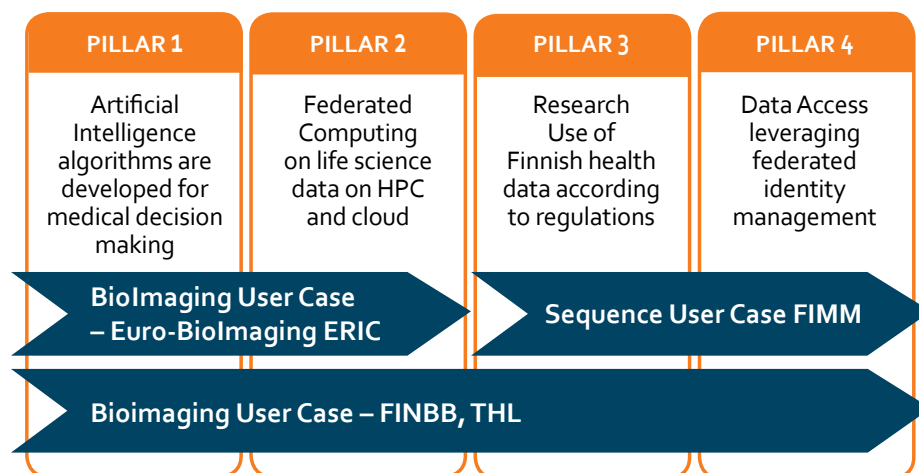
At the moment, the transfer and utilisation of genomic data does not work across borders. CSC is developing standards for genomic data technologies (such as GA4GH.org Passport, Cloud, Beacon), which are also relevant outside Europe, such as North America, Ja-

pan and Australia. The purpose of the ELIXIR infrastructure is to adopt global standards for the responsible sharing of genomic data. Europe also has a strong desire to create a federated data security infrastructure for sensitive genomic data. The plan is to create what is known as European Health Data Space (EHDS).

"ELIXIR has been developing good tools for a long time for researchers – improving usability by creating new tools. ELIXIR's cooperation with the Global Alliance for Genomic Health has created a fine vision on how this global cooperation could work, and also concrete tools and models," says THL Biobank's Director Sirpa Soini.

The aim is to make biobanks operate in a compatible, federated data infrastructure that transcends national borders. This is connected to the '1+million genomes' and 'Beyond million genomes' projects funded the EU member states and the Commission. In the 'Beyond million genomes' project, CSC is in charge of the technical infrastructure work.

THL Biobank's part in the project is to design management processes for national health data for research. The objective is to enable researchers and students easier access to material in Finnish biobanks. This would also mean that data could be transferred securely from biobanks to CSC's sensitive data environment and sharing it with those who have been authorised to access it.



Work in the project is divided into four pillars: artificial intelligence algorithms, computing services, research use of health data, and data management technology. Pillars are theme which, when combined, create solutions for the construction of services containing sensitive data. The success of development work is measured by means of three cases worked on in collaboration with ELIXIR, Finnish Biobank Cooperative (FINBB), the Finnish Institute for Health and Welfare and the Institute for Molecular Medicine Finland (FIMM).

Sirpa Soini is very well aware of the concerns and regulations concerning the use of sensitive data. She nevertheless feels that GDPR is too often blamed for any problems, although it is in fact many member states that themselves restrict the transfer of sensitive data in their own legislation or interpretations. Soini is also a lawyer by training and thinks that many issues can be solved provided there is enough political will.

“At the moment it seems that people are simply saying that various things cannot be done because of GDPR. But that’s not the real reason. It not the reason in Finland or elsewhere, and solutions are available.”

According to Soini, GDPR does not restrict data use, but in fact enables it, but in a responsible way, and taking account of the risks. National legislation is required to support certain use cases.

According to Soini, in secondary use of data it is difficult to predict subsequent use. But in cases like this, the premise should be that medical and applied research and product development is possible under GDPR, based on the law.

“This would mean that consent would not have to be obtained. Our law prescribes use for the general good, complete with the appropriate data protection and data security measures. You do not need full, detailed consent as such, although transparency should be promoted.”

She also says that there are no absolute legal obstacles for transferring data abroad. THL Biobank, for example, has made agree-



ments about data transfer to the United States and Australia.

“I suggested a cooperation agreement to the US and Australian lawyers, emphasising which responsibilities each partner has in terms of risk management. It is important that the agreements have precise restrictions and that the material is pseudonymised. It is also always specified where the data can be stored.”

One such place could be the European Genome-phenome Archive (EGA). To protect the identities of data providers, data made available for research has been pseudonymised. Only an authorised party, such as the Finnish Institute for Health and Welfare, may decrypt the pseudonymisation.

Soini speaks of a dream cloud in which the data itself would not be moving.

“Data could be stored securely in an international database. Direct searchers and identification would be possible within the Trust Federation Network, provided the datasets were ready. This would put the controller in control of its data, assessing requests to use the register. In an ideal situation, permits could cover several datasets around the world, in effect creating a type of federated solution: the data itself would not move anywhere, rather the researcher would be given access to a “dream cloud”. This could be accessed by researchers from various locations.”

Ari Turunen

MORE INFORMATION:

Institute for Molecular Medicine Finland (FIMM)

<https://www.fimm.fi/en/>

THL Biobank

thl.fi/en/web/thl-biobank

Euro-BioImaging

www.eurobioimaging.eu

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>

ELIXIR FINLAND

Tel. +358 9 457 2821s • e-mail: servicedesk@csc.fi
www.elixir-europe.org/about-us/who-we-are/nodes/finland

www.elixir-finland.org

ELIXIR HEAD OFFICE

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