Massive data management project: Finns' heredity is collected and safeguarded

Over the past three years, over 500 TB of sequencing data from Finnish research samples was transferred over fibre optic cable from the United States to Finland. The licensed and protected data were transferred to Finnish biobanks and will significantly contribute to the study of genetic diseases.

Until 2015, there were no resources to return Finnish genomic data generated by international research projects back to Finland. As a result, the Academy of Finland funded a project in which the **Aarno Palotie**'s and **Samuli Ripatti**'s research teams from the University of Helsinki's Finnish Institute of Molecular Medicine (FIMM) and Finnish IT Center for Science, CSC, started transferring data back to Finland from the genome sequencing centres in St. Louis and Boston.

"We created a good process that included license tracking, data transfer, reliability and security. Not many have transferred such a large amount of material from the United States to Europe. Thanks to Finnish universities' core network, FUNET, the data transfer rate was sufficient. In addition, CSC has experience in recording massive data files, such as storing all Finnish TV and film production on tape," says CSC's **Ilkka Lappalainen**, Head of Service Development for Health and Life Sciences.

"Saimme luotua hyvän prosessin, johon kuuluivat lupakäytännöt, aineiston siirtäminen, luotettavuus ja tietoturva."

The eSISu project (e-Infrastructure for Sequencing Initiative Finland) securely stores the details in Finnish genomes, that is, gene variations. By analysing variations, new information on hereditary diseases can

“If we do not have data on our genetic composition, how can we investigate the genetic effects of different diseases?”

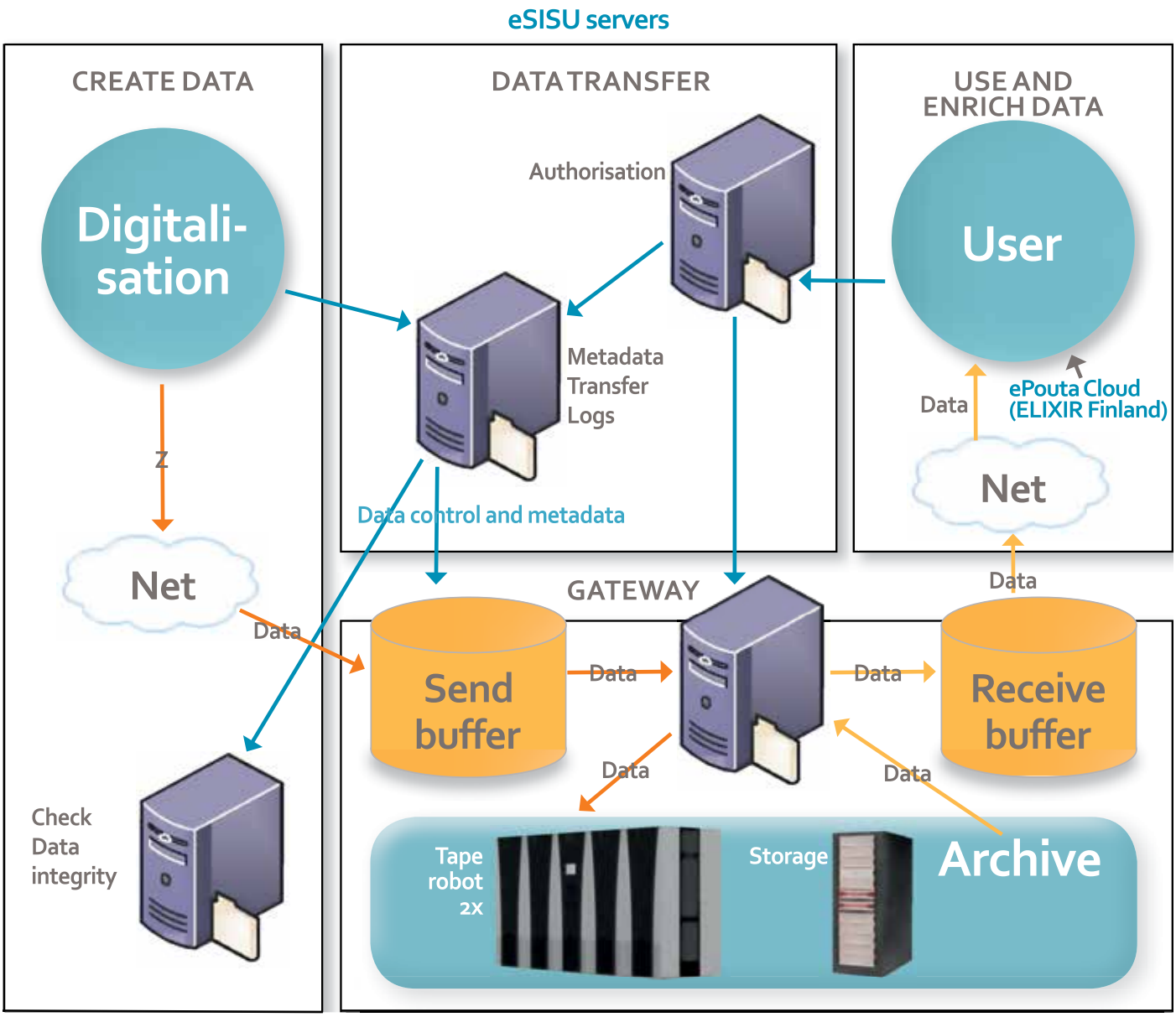
be discovered. The purpose of the SISu project (Sequencing Initiative Finland) is to aggregate genomics into a form that benefits best Finnish doctors and researchers. To date, the full genome of thousands of Finns and the protein-encoding parts of the genomes of almost 30,000 Finns have been determined.

The data collected from Finns is largely similar to that of other European countries, but certain parts of the Finnish genome have either been refined for this northern condition or they only existed in a few families that inhabited small villages in the north.

“For this reason, certain genetic variations occur in Finns, affecting, for example, cardiovascular diseases. If we do not have data on our genetic composition, how can we investigate the genetic effects of different diseases?”, asks Ilkka Lappalainen.

Genomic data is part of an integrative entity that combines lifestyles, medications, treatments and individual health data. It will allow for accurate statistical studies where the combined health effects of individual differences in genetic makeup and varying response to medication can be assessed.

“In certain cases, e.g. cancer treatments at the Helsinki University Hospital (HUS), this is already in use. Huge amounts of statistical data are being collected from cancer causing genes to redefine procedures and medical therapies and recommendations.If standardised data can be obtained from the entire population of Finland, people can, if necessary, be called in for cancer screening and decisions can be made about the appropriate treatment. Future forms of treatment are not possible with the data only collected from Finns. Cancer treatments are largely developed through international cooperation.”



The hope is that this knowledge will be passed on to general health care that notifies the person who gave the sample. It is then up each person to decide if they want to know more about their risks.

One of the most important research topics in bioinformatics is to understand the mechanisms causing diseases. Part of the data collected during the eSISu project came from migraine patients. In summer 2016, the project reached a major milestone when the first set of data for the migraine genome was transferred to Finland. The data transfer was executed without any technical or security problems.

New information about migraine and coronary heart disease

FIMM's scientists have been using SISu data to verify that inherited susceptibility to migraine does exist, and showed that genetic starting points for migraine sensitivity can be traced to 38 regions in the genome. The finding is a first step in understanding the mechanisms of migraines and opens up the possibility of more accurate diagnostics and treatments.

Finnish researchers have also found new genetics variations that affect the susceptibility to coronary heart disease. Thanks to this new information, risk groups for coronary heart disease can identified earlier and instructed to start preventive measures that might include lifestyle changes or preventive medication.

There is still plenty work left to do with the collected eSISu data. There are cases where several samples have been taken from one individual and sent to different places for analysis. Now these all will be traced back to the source.

"We are now working with metadata to find out what was collected earlier and add value to future research projects."

According to Lappalainen, a lot of valuable experience in data management was

gained through the project. This will be useful for the new FinnGen project.

Started in December 2017, the FinnGen project will store the genomes of half a million Finns. The project utilises samples collected by Finnish bio banks. The data from the genome is combined with the information in national health registers. This makes it possible to better understand the mechanisms by which diseases are born and to then develop new treatment methods.

Good governance of metadata opens up opportunities for data integration for research. For example, there are about 5.4 million people in Finland and almost all medical prescriptions end up in the archives. The bio bank law in Finland guarantees the responsible research use of genomic data.

SISu has already been identified as a significant data resource for ELIXIR and BBMRI infrastructures. In the next phase, the organisation and management of data will move to a scalable and secure platform (ePouta cloud service). This will make data computationally available. Finnish biobanks, such as the THL biobank, will continue to oversee the use the material and grant permissions for the use of it.

"Currently, data transfer is being tested and it will continue to work once the metadata has been updated."

eSISu is creating Finnish capability to manage controlled access to genome data between the Finnish ELIXIR Centre and other ELIXIR Centres. It enables CSC, with consent from the data owners, to integrate these data with other registers and databases in Finland.

"This way, we can combine Finnish data with European EGA (European Genome-phenome Archive) data."

The European Genomic Archive is one of the world's largest public data repositories with patient data from biomedical research projects. EGA shares human genetic and phenotypic data through a consent process that allows the reuse of data for research purposes. Thanks to EGA, many ELIXIR research projects are possible.

"www.sisuproject.fi is a search service where you can find more about genetic variants of the Finnish population."

www.sisuproject.fi is a search service where you can find more about genetic variants of the Finnish population. The KITE search engine, on the other hand, searches the data on the basis of the metadata. These are examples of services that are being developed for both national and international use. Data management and licensing practices are handled using the new ELIXIR REMS (Resource Entitlement Management System) software.

"Technically, the data management works well. A significant part of SiSu's material will be available during 2018."

Ari Turunen

MORE 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

Tommi Nyrönen
Head of Node, ELIXIR Finland
tommi.nyronen@csc.fi
+358-50-3819511