NeIC Heilsa: The Roadmap to Sensitive Data Archiving & Federated EGA

Stefan Negru
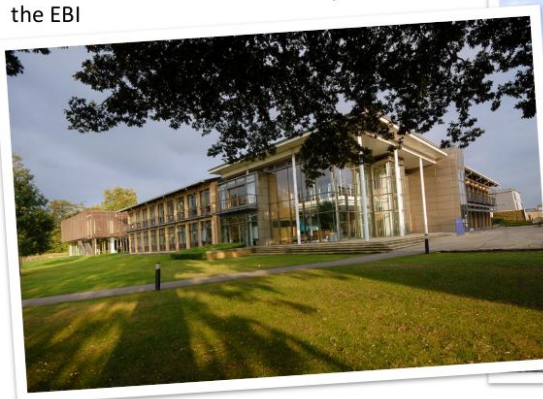### CSC – IT CENTER FOR SCIENCE

NeIC Heilsa WP1 co-lead

EGA brief history

- The European Genome-phenome Archive (EGA) is a **service for permanent archiving** and **sharing** of all types of p**ersonally identifiable genetic and phenotypic data** resulting from **biomedical research projects**.

How the EGA is managed

The EGA was launched in 2008 by the EBI

In 2012, EBI and CRG started working together to establish EGA as a joint venture, which has grown in the context of ELIXIR

EMBL-EBI

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EUROPEAN **GENOME-PHENOME** ARCHIVE

Heilsa

What is Federated EGA

"The **Federated European Genome-phenome Archive (EGA)** is a **distributed network of repositories** for **sharing human-omics data and phenotypes**.

Typically a **node** would be an **organization or project that hosts human genetic data so that the data can remain within a jurisdiction**."

https://ega-archive.org/federated (source of the text)
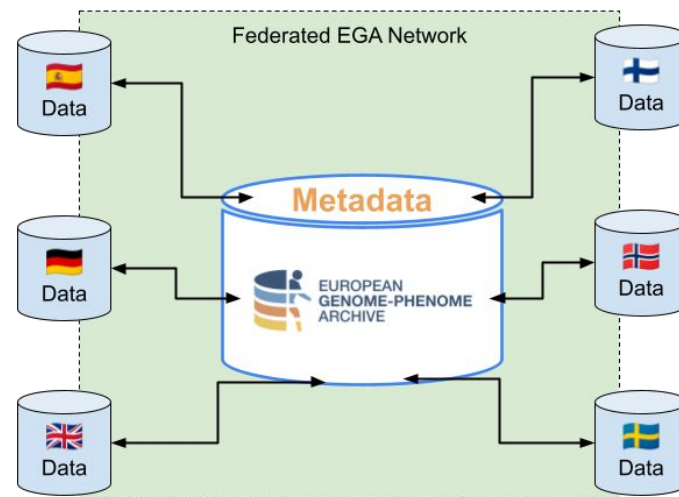
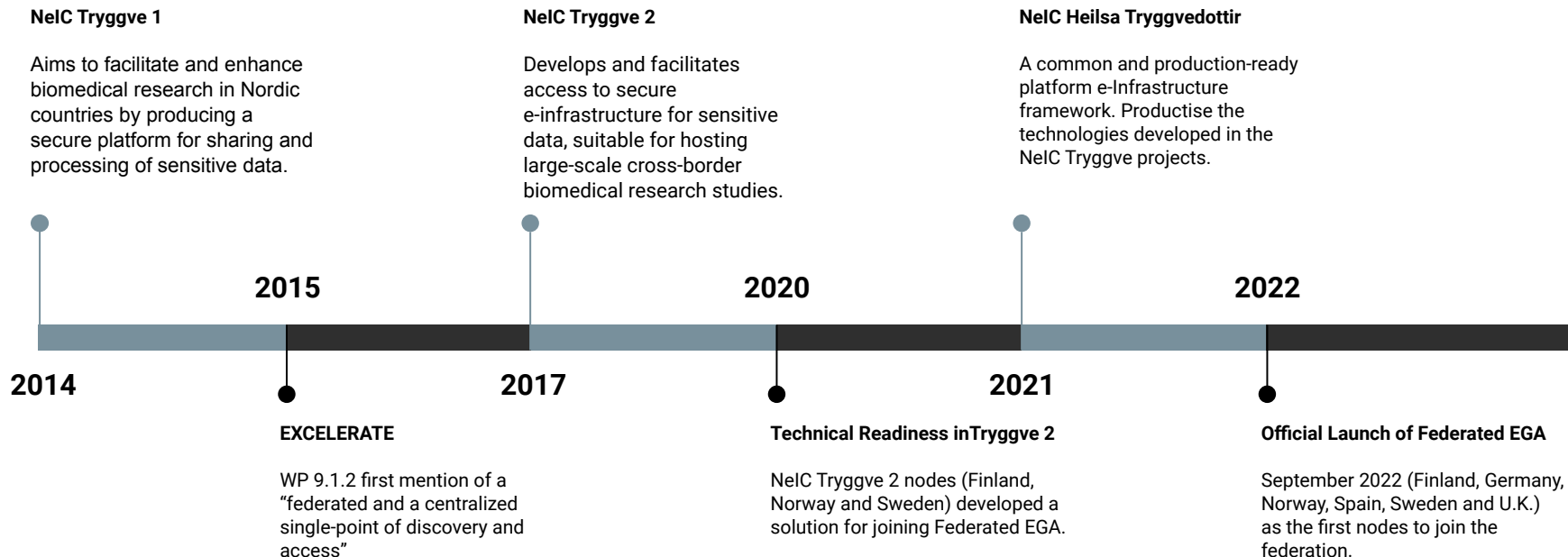# Benefits for research community

- Unified accession identifiers and versioning scheme for identifying and referencing data conforming to the FAIR principles;
    - EGA accession identifiers can be used in scientific publications by researchers;

- Use of international standards (crypt4GH, beacon, HTSGET etc.);

- Unified brand for storage, discovery and access to human genetic and phenotype data.
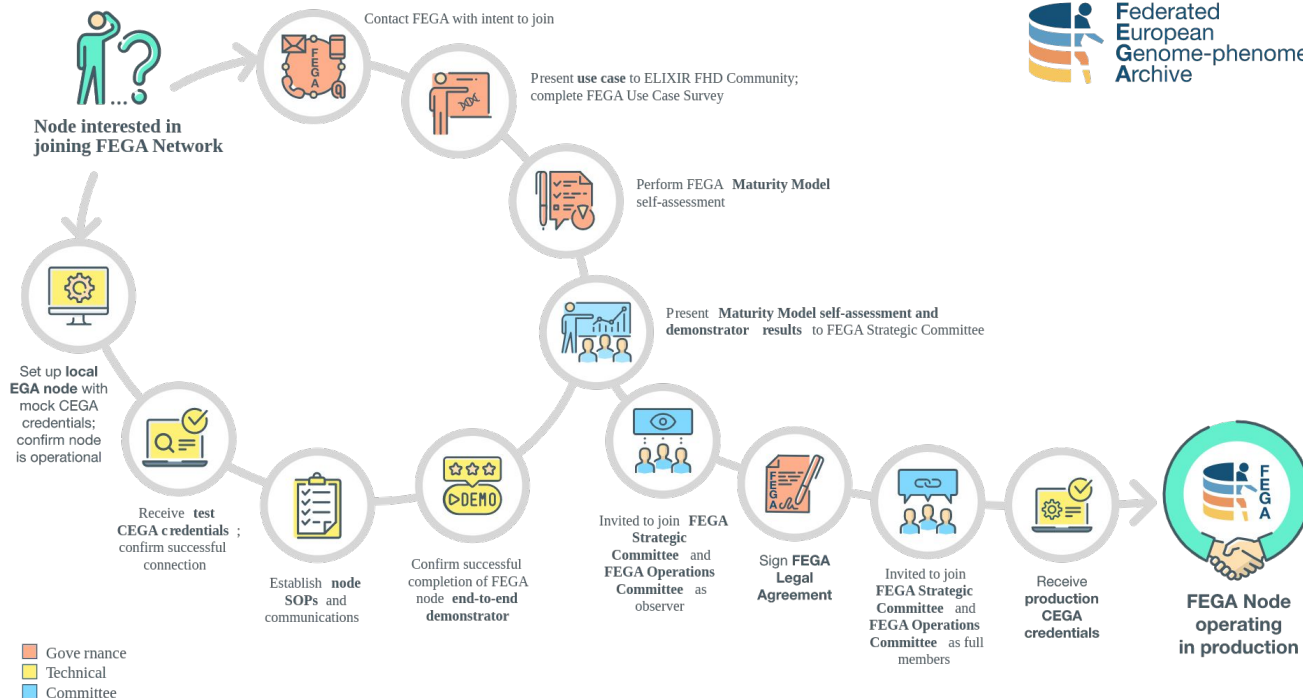
Federated EGA - History briefly

- "Portable submission toolkit" or "mini-EGA in-a-box" mentioned in the EXCELERATE grant (2015-2019) https://elixir-europe.org/about-us/how-funded/eu-projects/excelerate
  - *Subtask 9.1.2: Portable submission toolkit. This task is composed of data format definitions and software components, a "mini-EGA in-a-box" will allow increased local control and coordination of data collection, and allow early validation of standardized data and metadata formats. This implementation provides the practical means for distributed projects to collect access-controlled human biomedical data in a manner that maintains a coordinated data model and dataset registry, enabling federated and a centralized single-point of discovery and access.*
- In March 2016 (AHM Barcelona) when polling the ELIXIR Nodes, the need *for *not* moving the sensitive files from "local jurisdictions"* arise as the main driver for it
  - … but still willing to share data for research
- Known over the years as *"mini-EGA in-a-box", "Federated EGA DROPBOX", "Local EGA", "EGA sensitive data archive", "FEGA", "LEGA"* etc.
- As the ELIXIR funding for the task was minimal, some Elixir/Nordic Nodes step in to move the project forward

Federated EGA - Timeline*

NeIC Tryggve 1

Aims to facilitate and enhance biomedical research in Nordic countries by producing a secure platform for sharing and processing of sensitive data.

NeIC Tryggve 2

Develops and facilitates access to secure e-infrastructure for sensitive data, suitable for hosting large-scale cross-border biomedical research studies.

NeIC Heilsa Tryggvedottir

A common and production-ready platform e-Infrastructure framework. Productise the technologies developed in the NeIC Tryggve projects.

2015

2020

2022

2014

2017

2021

EXCELERATE

WP 9.1.2 first mention of a "federated and a centralized single-point of discovery and access"

Technical Readiness inTryggve 2

NeIC Tryggve 2 nodes (Finland, Norway and Sweden) developed a solution for joining Federated EGA.

Official Launch of Federated EGA

September 2022 (Finland, Germany, Norway, Spain, Sweden and U.K.) as the first nodes to join the federation.

* from a NeIC involvement perspective

Federated EGA - Journey*

Node interested in joining FEGA Network

Contact FEGA with intent to join

Present **use case** to ELIXIR FHD Community; complete FEGA Use Case Survey

Perform FEGA **Maturity Model** self-assessment

Present **Maturity Model self-assessment and demonstrator results** to FEGA Strategic Committee

Set up **local EGA node** with mock CEGA credentials; confirm node is operational

Receive **test CEGA credentials**; confirm successful connection

Establish **node SOPs** and communications

Confirm successful completion of FEGA node **end-to-end demonstrator**

Invited to join **FEGA Strategic Committee** and **FEGA Operations Committee** as observer

Sign **FEGA Legal Agreement**

Invited to join **FEGA Strategic Committee** and **FEGA Operations Committee** as full members

Receive **production CEGA credentials**

FEGA Node operating in production

- Governance
- Technical
- Committee

Federated European Genome-phenome Archive

* as illustrated by https://ega-archive.github.io/FEGA-onboarding

NeIC Heilsa Tryggvedottir - Work Packages

- WP1: Federated cross-border analysis, sensitive data archives, and roadmaps

- WP2: Creating and sharing genome/phenome test datasets

NeIC Heilsa Tryggvedottir - joining Federated EGA

- A code base for archiving of sensitive data
  - Documentation available at: https://neic-sda.readthedocs.io/en/latest/
  - Packaged to be convenient for setting up Federated EGA, or stand-alone instances for archiving of sensitive data
  - Technical questions about the solution https://gitter.im/neicnordic/sda-pipeline
- Common operating procedures for (Federated EGA) nodes
  - Documentation
  - Processes
  - Roles and responsibilities
  - Security

NeIC Heilsa Tryggvedottir - Sensitive Data Archive Components

NeIC Heilsa Tryggvedottir - Outreach

- The Nordic ELIXIR nodes are engaged in multiple international projects and collaborations, such as:
  - 1+MG https://digital-strategy.ec.europa.eu/en/policies/1-million-genomes
  - B1MG https://b1mg-project.eu/
  - ELIXIR Federated human data community https://elixir-europe.org/communities/human-data
  - ELIXIR Converge https://elixir-europe.org/about-us/how-funded/eu-projects/converge
  - EOSC-Nordic https://www.eosc-nordic.eu/about-eosc-nordic/
  - EuroHPC - Lumi https://www.lumi-supercomputer.eu/eurohpcju/
  - GA4GH https://www.ga4gh.org/

Thank you!