Expanded NFEGA service description

This document is intended to aid in understanding terms and operations related to the Norwegian Federated EGA (NFEGA) service.

The European Genome-phenome Archive (EGA) is a resource for long term secure archiving and sharing of all types of potentially identifiable human genetic and phenotypic data. EGA was launched by the European Bioinformatics Institute (EMBL-EBI) in 2008, and is now established as a prioritized Core Data Resource in the pan-European ELIXIR data infrastructure. ELIXIR Europe is the European bioinformatics science research infrastructure (an ESFRI infrastructure). The main purposes of EGA and Federated EGA resources are:

- Provide controlled access to sensitive data
- Make sensitive data Findable, Accessible, Interoperable and Reusable (FAIR)
- Increase visibility of datasets, and facilitate secure sharing of these
- Put the dataset owner in full control of granting access to datasets, through a Data Access Committee (DAC) appointed for each dataset by the dataset owner
- All within the legal constraints of sharing sensitive data

Federation of EGA nodes

The European Union General Data Protection Regulation 2016/679 (GDPR) started to apply in 2018, and was implemented in Norway through the Personal Data Act (LOV-2018-12-20-116). ELIXIR Europe has since then collaborated to establish a federated network of archive nodes, where:

- No sensitive data needs to leave the country of origin when archived
- Only anonymous meta-data (data about the data) describing a dataset as a whole is shared between the nodes
- All information elements about individual samples are kept secure and encrypted in the country of origin
- Sensitive data is only made accessible to a Researcher requesting access after explicitly being granted access from the data controller (data owner), through an appointed Data Access Committee (DAC), as the acting arm of the data controller, when the DAC has reached satisfactory terms for how the Researcher can use and process the data further. This will require contracts to be signed between the parties. The DAC will typically consist of the Project PI, and a couple of designated other relevant parties to the project.

The Norwegian Federated EGA node

ELIXIR Norway, the Norwegian node of ELIXIR Europe, has established a Norwegian node in the Federated EGA network, the NFEGA resource. NFEGA utilizes a common set of software modules, implementing international standards from the Global alliance for genomics and health (GA4GH) for processing and storing sensitive data, that has been jointly developed by the Nordic ELIXIR nodes since 2018. The NFEGA resource is hosted by the University of Oslo, as a partner in the ELIXIR Norway consortium.

The **security of the data** stored in NFEGA and the conformity of NFEGA procedures to Norwegian law, is guaranteed by the following:

- The Sensitive Data Archive is fully deployed inside the TSD infrastructure, Tjenester for Sensitive Data.
- All data, both genomic and sample descriptions, are stored with strong encryption, and combined provides an additional layer of security to what is provided from TSD itself.
- The NFEGA procedures and Risk and Vulnerability analysis (in Norwegian: "ROS analyse") has been developed, reviewed and approved in close collaboration with the University of Oslo's legal representative, Data Protection Officer, Data Security Officer and the hosting institute.

Current service: NFEGA version milestone 1

The service offers, in its first incarnation, a facility for archiving of sensitive data through a portal-based system of deposition, aided by ELIXIR Norway personnel. In later milestones, more advanced geo-replication of data across several sites in Norway is planned to be developed, as well as more automation to replace manual procedures. Regulatory documents for the service will be amended accordingly when the changes related to later milestones come into operation.

Technical details on aspects of the NFEGA milestone 1 implementation

Secure Internet connectivity

The technical implementation of NFEGA is based on all data residing encrypted inside a dedicated TSD project (p969), where only NFEGA team members have login access. All communication with internet-connected servers outside TSD occurs through standard TSD APIs, which are approved, secured and operated by TSD staff. As part of the NFEGA resource, a set of micro-services and a web-portal is hosted on a proxy server outside of TSD that utilizes the available standard TSD API interfaces. The communication channels are all based on https encryption, and sensitive data is encrypted itself before being communicated over this encrypted channel. The proxy server is operated jointly by TSD staff and ELIXIR Norway, and is needed to enable communication to the central European EGA service for metadata management, as well as submission of data to be archived, and to perform approved delivery of data to DAC approved requesters.

Data protection at rest in the archive, and management of encryption keys

Crypt4GH is the Global Alliance 4 Genomics and Health (GA4GH)-recommended standard specification for encrypting sensitive data, both for protection in transit and for storage at rest. NFEGA services and tools implement the Crypt4GH standard for all storage and transit of sensitive data.

A Crypt4GH-encrypted dataset submitted to NFEGA is split in two before storage: 1) the header of the dataset contains an encryption key to decrypt the body of the dataset. This header is split from the body, and re-encrypted with the NFEGA master vault password, before it is stored encrypted at rest in an internal postgres database hosted by the TSD database hotel services. 2) The body of the dataset is stored as is encrypted at rest on the p969 project file system inside TSD. The master vault password is stored in the home directory of p969 for an NFEGA project user (a separate single account in the specific NFEGA user project, to which only a very limited number of NFEGA personnel has access (normally 2-3 people among operational staff)).

Transfer of encryption keys between sender and receiver

The Crypt4GH standard is based on the principle of a private-public pair of keys. The data to be securely communicated are first encrypted by the sending party, with the public key of the receiving party (the public key can freely be distributed) together with a private key only known to the sender. The receiver is then able to decrypt the dataset later with only the receiver´s private key, which is never shared. This principle thus only requires sharing of public keys between the parties that need to exchange data securely.

In NFEGA, we use this mechanism both for submitters to deposit encrypted sensitive data for archival, as well as to deliver controlled access-encrypted sensitive data from NFEGA for download by requesters approved by the appropriate Data Access Committee. Detailed instructions for submission and reception of encrypted data are available at the NFEGA web pages (for submission of data; https://ega.elixir.no/submission.html, for download of data: https://ega.elixir.no/retrieval.html).

Data requesters and access to the TSD project itself

In NFEGA Milestone 1, a requester of data will never be granted TSD project access to download data themselves. NFEGA personnel with TSD project access will prepare for the requested download if duly approved by the DAC, according to the restrictions set forth in the signed data access agreement as defined by the data controller acting through the DAC. A re-encrypted version of the dataset to be ready for download is prepared according to the Crypt4GH GA4GH standard described in the previous section. This re-encrypted dataset, that can only be decrypted by the intended recipient data requester, is then made available through the TSD API, without the requester ever being given any direct project access. This is ensured through the existing functionality of the TSD API through provision of specific files to specified recipients.

Inside TSD, the need for manual access to the data by NFEGA personnel is minimized through the use of preprogrammed procedures for mobilizing data for export to the data requester.

Legal aspects and concerns

Role of a Data Access Committee

The Data Access Committee is appointed by the Data Controller upon submission of a dataset to the NFEGA resource. It will serve as the contact point for all requests regarding the submitted datasets, on behalf of the Data Controller organisation. Thus, any questions from NFEGA staff or data requesters will be routed to the appointed DAC of the given NFEGA dataset in question.

Agreements between parties

To facilitate the use of the services, a contract template has been made available as a suggested way to regulate the relationship between the data controller and the NFEGA service, in the form of a Data Processor Agreement (DPA). Another template, a Data Access Agreement (DAA), has also been developed as a suggestion for the agreement to be made between the Data Controller, through the DAC, and the Data Requester. The specific type of agreement to be made between the Data Controller and the Data Requester is fully up to the Data Controller to negotiate with the Data Requester, and a satisfactory agreement between the parties is the basis of granting controlled access for the Data Requester to the NFEGA hosted dataset in question.

Impartiality of Elixir Norway as the NFEGA Service Provider

While the NFEGA service is operated by ELIXIR Norway partners that actively perform research themselves, this will in no way influence priorities in the quest to provide unbiased high-quality services to the users of NFEGA. All users are given equal access and priority to the system. ELIXIR personnel operating NFEGA will not in any way access the data for research purposes.

Risk of re-identifying subjects from pseudonymized data

The risk of accidental re-identification of subjects from the data itself as part of the NFEGA archival operations is evaluated as minimal. There is no routine processing of the data in Milestone 1 that may lead to accidental findings like this. The main risk of re-identification will be from further analysis of the data by authorized requesters that the Data Access Committee has approved after entering the appropriate agreements with the requester. Such agreements must directly address the risk of re-identification and procedures for addressing accidental findings if that should arise. In the template text for such agreement that NFEGA provides as a courtesy to the parties, we have included the following section:

> *“The data processor agrees to preserve, at all times, the confidentiality of these Data. In particular, it undertakes not to use, or attempt to use these Data to compromise or otherwise infringe the confidentiality of information on Research Participants. Without prejudice to the generality of the foregoing, the data processor agrees to use at least the measures set out in Appendix I to protect these Data.*
> *The data processor agrees to protect the confidentiality of Research Participants in any research papers or publications that they prepare by taking all reasonable care to limit the possibility of identification.*

> *The data processor agrees not to link or combine these Data to other information or archived data available in a way that could re-identify the Research Participants, even if access to that data has been formally granted to the data processor or is freely available without restriction.”*

In the case of a need for updated consents

If the purpose of re-use of data intended by a data requester is not sufficiently covered by the current participant consents for an archived dataset, it is solely a matter for the Data Controller, through the appointed DAC, to see if this is possible and base any access decisions upon that. NFEGA will not be directly involved in such legal matters.

Procedures

For researchers that want to submit data to NFEGA, we are maintaining relevant procedure documentation and contact information at the main NFEGA website (https://ega.elixir.no). This includes templates for agreements mentioned in this text, a general description of how to involve relevant roles at their own organisation for different tasks, etc. We are also willing to include external pointers to relevant data of the submitter's own documentation and guidelines to use NFEGA.

In addition, NFEGA has internal procedures of relevance to a data controller for assessing the NFEGA services in their own security and data protection related work, such as ROS-analysis, DPIA etc. We commit to make these internal procedures, as specified in our Data Processing Agreement template, available upon request, but will not publicly distribute them in general.