E3.3. Informe Final de un Nodo EGA v 1.0

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES
Instituto de Salud Carlos III

IMPaCT

Infraestructura de Medicina de Precisión asociada a la Ciencia y la Tecnología

Program	IMPacT: Infraestructura de Medicina de Precisión asociada a la Ciencia y la Tecnología		
Project Name	IMPaCT-Data: Programa de Ciencia de Datos de IMPaCT		
Expedient	IMP/00019		
Duration	January 2021 – December 2024		
Work Package	WP3–Genomics		
Task	T3.1 - Implementación de sistemas de almacenamiento de información genómica primaria mediante instancias de local EGAs		
Deliverable	E3.3 Informe Final de un Nodo EGA		
Version	v0.1		
Due Date	July 31, 2024	***Approval Date***	September 30, 2024

Authors		
Organization	**Name**	**Role**
Acronym	**Name and Surname**	**Coordination / Author / Reviewer**
CRG	Jordi Rambla	Author
CRG	Amy Curwin	Author
CRG	Sabela de la Torre	Author
BSC	Albert Hornos	Reviewer
FPS	Javier Perez-Florido	Reviewer

Versions History			
N.	**Date**	**Description**	**Author**
v0.1	01-06-2024	first draft	Amy Curwin
v0.2	01-07-2024	sections elaborated further	All
v0.3	17-07-2024	final draft prepared by authors	Amy Curwin, Sabela de la Torre, Jordi Rambla
v0.9	19-07-2024	v0.9 sent to reviewers	Amy Curwin
v0.9	24-07-2024	reviewed	Albert Hornos
v0.9.1	26-07-2024	incorporated suggestions from review (e.g. clarify FEGA ecosystem, new Figure 2, updated concluding remarks)	Amy Curwin
v0.9.2	29-07-2024	reviewed	Javier Pérez-Florido
v0.9.3	30-07-2024	minor edits based on review	Amy Curwin
v1.0	01-08-2024	v1.0 finalised and delivered	Amy Curwin

Content

Executive Summary

The European Genome-phenome Archive (EGA) is a service for law-compliant storing and sharing of all types of human genomic and phenoclinic data. The EGA has functioned as a centralised service, co-managed by the EBI and the CRG, however in 2022 the Federated EGA (FEGA) was officially launched. FEGA is a network to support data management requirements in different countries where national laws do not permit personal human data to be submitted outside their borders. FEGA nodes offer EGA services to researchers within their jurisdiction and share public metadata with the Central EGA (CEGA), so that it is discoverable via the EGA website. A FEGA Community node, on the other hand, can be an individual institution or initiative, such as those within IMPaCT, with human genetic data to share (under specific conditions), but also face legal and/or ethical restrictions, and therefore cannot, or prefer not to, submit their data to a central repository. The technical and governance challenges faced are very similar, therefore the solutions to these are often interchangeable. Herein, we describe the technical, legal and human requirements to operate a FEGA node and, by extension, a FEGA Community node.

Introduction

Audience

This deliverable is envisioned as a useful document for those institutes who are interested in establishing a Community instance of FEGA to allow sharing and discoverability of human genomic and phenoclinic data.

Topic

This deliverable describes technical, legal and human requirements to establish a FEGA Community node, that can enable entities such as hospitals, research centres, etc. to participate in the management, analysis and sharing of genomic data (and associated metadata). This data is the basis for the development of personalised medicine and therefore, empowering the entities with the tools, knowledge and network to manage this type of data, is instrumental to the final mission of the IMPaCT-Data project. The current model is for nodes to minimise moving sensitive data from the node, while public metadata have no restrictions. Files will be stored encrypted in the nodes located at different institutions, while public metadata goes to Central EGA.

Relation to other Deliverables

Related to E3.1 “Requisitos de un Nodo Local EGA[1]” and E3.2 “Descripción de Interfaces de Instancias EGA Comunidad[2]”.

Deliverable Structure

1. Overview of the Federated EGA network
2. Technical requirements
3. Legal requirements
4. Human requirements
5. Concluding remarks

[1] https://b2drop.bsc.es/index.php/apps/files/?dir=/IMPaCT-Data/Entregables&openfile=624698
[2] https://b2drop.bsc.es/index.php/apps/files/?dir=/IMPaCT-Data/Entregables&openfile=2772397

1 The Federated EGA Network

1.1 State of the art of the Federated EGA

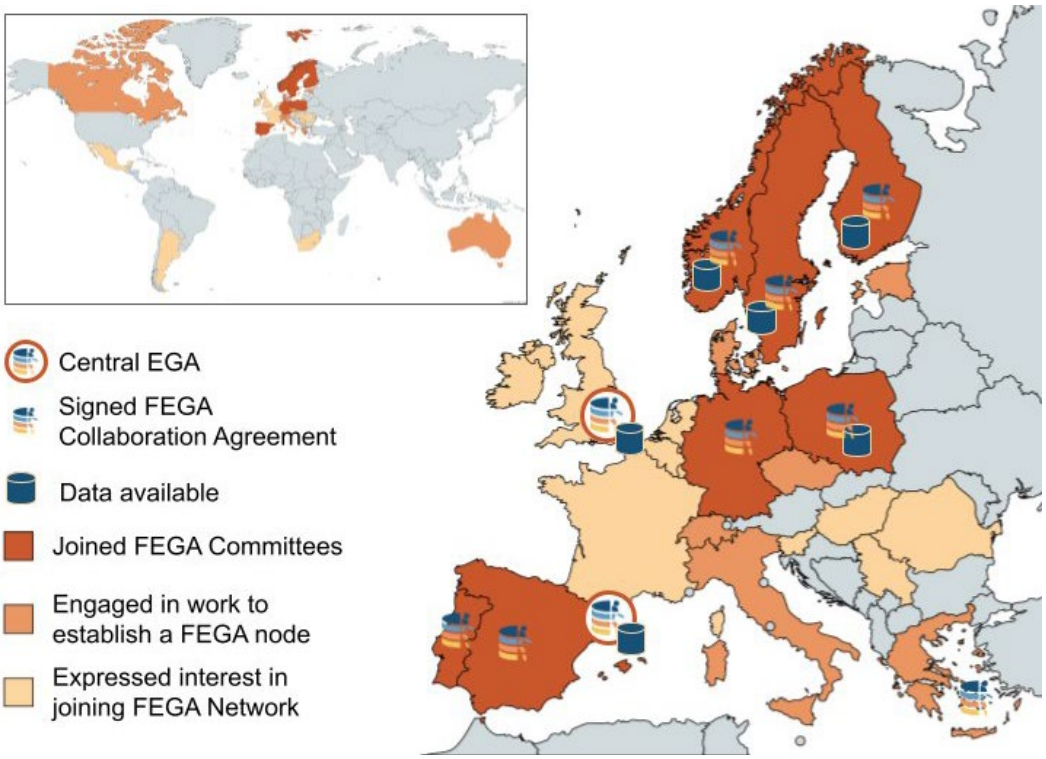

Figure 1. The current status of the Federated EGA Network

The European Genome-phenome Archive (EGA) has served as a resource for permanent archiving and sharing of personally identifiable genetic, phenotypic, and clinical data generated for biomedical research projects or in the context of research-focused healthcare systems, since 2008[3]. The increasing scale of data, along with introduction of national legislations which potentially restrict data movement posed significant challenges to this centralised data access model. In response to this, the Federated EGA (FEGA) officially launched in 2022 when the 5 inaugural nodes (Finland[4], Germany[5], Norway[6], Sweden[7], Spain[8]) signed the FEGA collaboration agreement. This was a huge milestone and the product

[3] *The European Genome-phenome Archive in 2021*.doi 10.1093/nar/gkab1059
[4] https://research.csc.fi/-/fega
[5] https://www.ghga.de/
[6] https://ega.elixir.no/
[7] https://fega.nbis.se/
[8] https://fega-test.bsc.es/docs/

of years of effort. Overcoming the legal and ethical issues indeed was very challenging. Since 2022, 2 more nodes (Poland[9] and Portugal[10]) have signed and officially joined, with 20+ numbers of others within Europe and worldwide at various stages of technical or legal readiness (Figure 1).

○ 1.2 Structure of the Federated EGA

The structure and characteristics of different FEGA nodes are depicted in Figure 2. Federated EGA nodes (referred to as FEGA nodes) commit locally, typically in a national jurisdiction, as is the case for the 7 current FEGA nodes described above, although this is not a strict requirement of a FEGA node. A FEGA Community node (also referred to as simply a Community node) also operates locally, but is committed only to their own data (e.g. through a specific project or consortium). Therefore, the main difference between a FEGA node and a Community node, is the acceptance of third-party data. In the case of the latter, this does not occur. The function of a Community node is to manage and distribute their own data, whereas, the function of a FEGA node is to manage and distribute data within their region, therefore, they do accept third-party data within their jurisdiction.

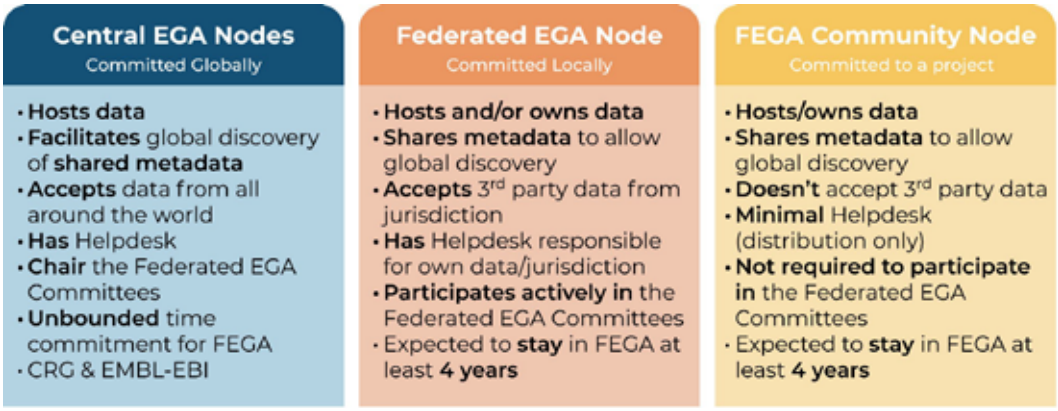

Figure 2. The Federated EGA structure at a glance

An example of a Community node could be a clinical research centre, a hospital, a consortium or project, like those within IMPaCT-Data, that wants the ability to share their data with approved secondary use, but for legal or technical reasons, institutional policy or otherwise, the data cannot be deposited outside the centre where it was generated. On the other hand, a FEGA node forms when national law or policy prevents deposition of data generated within their own nation outside their borders, but similarly they want to share the data for secondary use to the broader research community. For this reason, one major commonality, and indeed

[9] https://www.fega.uni.lodz.pl/
[10] https://fegaportugal.biodata.pt/

a main purpose of such a federation, is that the public metadata is shared with CEGA and made discoverable worldwide via the EGA website[11] (Figure 3).

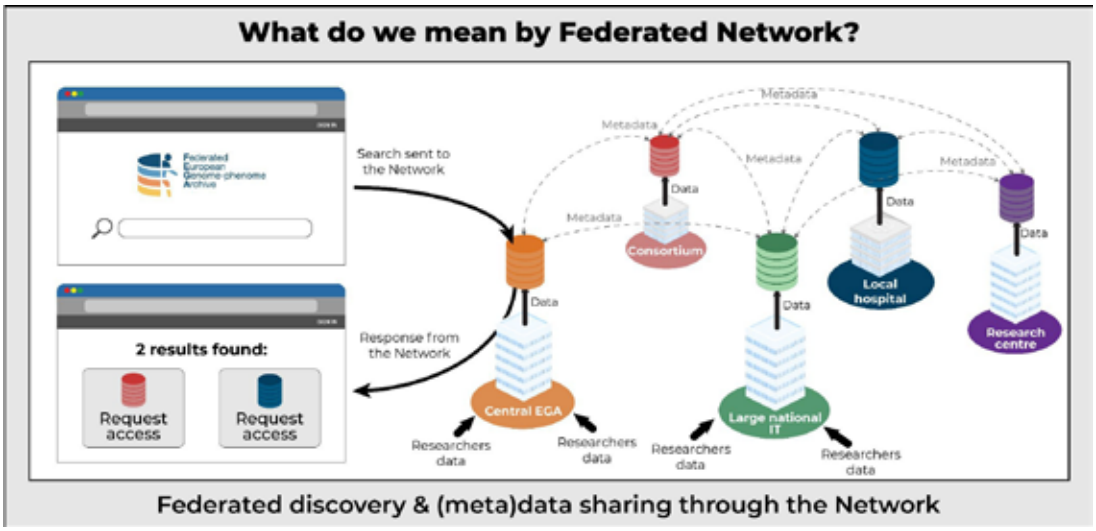

Figure 3. The Federated EGA network and federated discovery

○ 1.3 Benefits of joining the Federated EGA

- **Enhanced Discoverability**. Your (meta)data will be available in a global catalogue, maximising its discoverability.
- **Built with FAIR in mind**. FEGA is deeply committed to the FAIR principles[12]. By abbiding to FAIR you ensure that your data has the necessary Findability, Accessibility, Interoperability, and Reusability to meet requirements of funders.
- **Aligned with most relevant data initiatives**. Many health data related projects are now under way, in particular the Genomic Data Infrastructure (GDI)[13] and the European Health Data Space (EHDS)[14]. These projects use FEGA like architectures for data storage and distribution, once again proving the quality of this design.
- **Network of Experts**. By joining the FEGA ecosystem, you will be part of a network of experts, making sure to stay up to date with the State-of-Art on Data Sharing.
- **Globally Accepted Standards**. FEGA develops and follows interoperable global standards for human data access accepted by the Global Alliance for Genomics & Health (GA4GH)[15]. This commitment makes the whole service trustworthy for the most demanding data submitters, often concerned with data security and privacy.

[11] https://ega-archive.org/
[12] https://www.go-fair.org/fair-principles/
[13] https://gdi.onemilliongenomes.eu/
[14] https://www.european-health-data-space.com/
[15] https://www.ga4gh.org/

○ 1.4 The path to join the Federated EGA

Figure 4 depicts the necessary steps to join the FEGA network as a FEGA node. The top path in red are the legal and governance steps required, while the below path in yellow indicates the technical steps required. The two paths can be followed independently and converge in the Maturity Model self-assessment, that covers all aspects of operating and maintaining a FEGA node. Although this pathway and Maturity Model have been designed with a FEGA node in mind, some aspects are applicable to a Community node, especially on the technical side. The precise requirements to join FEGA as Community node are still under evaluation, but since the commitment is less, the path will be slightly different (e.g. presentation to Elixir Federated Human Data community or FEGA Strategic committee could be skipped, in lieu of another approval process).

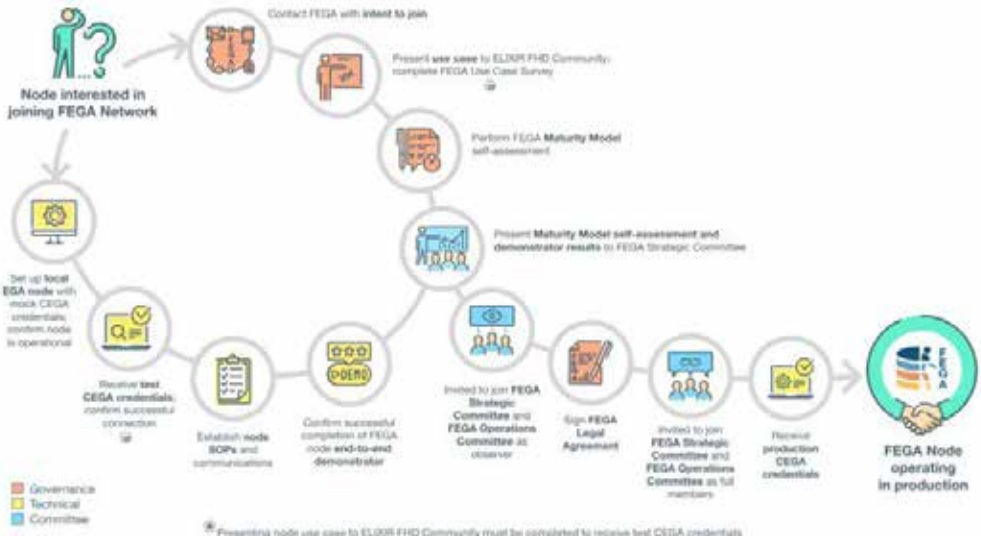

Figure 4. The path to join Federated EGA

The steps of the path to join FEGA will be referred to in more detail in the below sections and the differences for a Community node will be highlighted therein, as well. More details on the process of joining FEGA can be found in the FEGA Onboarding[16] and Maturity Model[17] websites. The overall requirements for operation of a FEGA node are described in the Node Operations Guide[18] produced by the FEGA committees and available on the main EGA website. Herein, we provide an overview of these requirements and how they would compare and differ with a Community node.

---

[16] https://ega-archive.github.io/FEGA-onboarding/
[17] https://inab.github.io/fega-mm/
[18] https://ega-archive.org/assets/files/Federated-EGA-Node-Operations-v2.0.pdf

2 Technical requirements

2.1 Overview and connection to Central EGA

The FEGA architecture encompasses the full data management life cycle, which includes: 1) the submission of data and metadata to CEGA or FEGA nodes, 2) secure storage, 3) metadata publication to facilitate data discovery, 4) handling researcher access requests, and 5) data distribution. These activities are conducted in secure environments with encrypted data, and user accounts are rigorously managed using advanced authentication and authorization systems. To ensure interoperability within the FEGA network, community standards such as those developed by GA4GH are crucial for data and metadata formats and external interfaces. FEGA nodes have the flexibility to implement or deploy any components of their choice, provided they adhere to the established standards. While FEGA nodes are not mandated to offer trusted research environments (TREs) or secure processing environments (SPEs), the data distribution components are designed for seamless integration with popular bioinformatics tools. This allows FEGA nodes to potentially combine their services with TREs or SPEs to offer data analysis services to researchers.

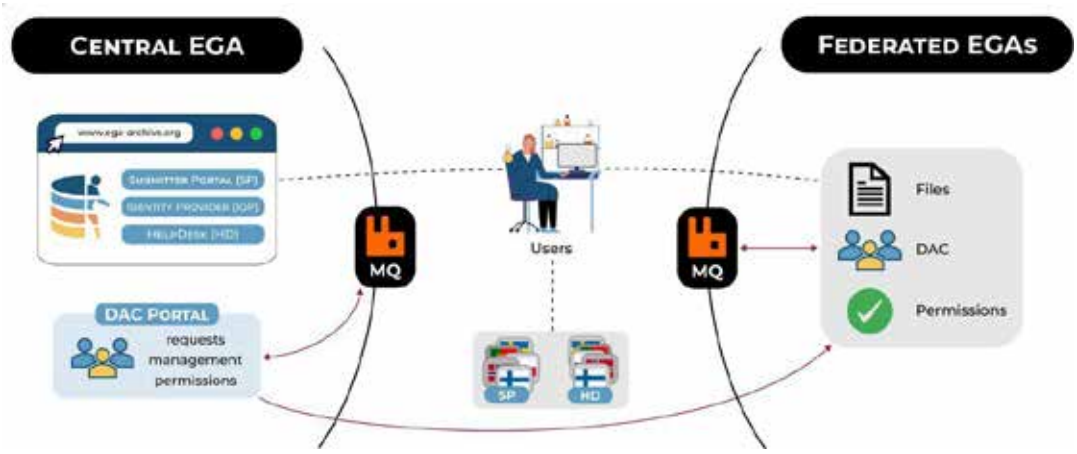

Figure 5. The technical connections between Central EGA and Federated nodes

Communication between the node and CEGA is done through RabbitMQ (Figure 5). CEGA also hosts a Data Access Committee (DAC) portal for managing DAC members, permissions and requests, therefore this information is also shared between CEGA and the FEGA nodes.

2.2 The Local EGA solution

One available technical solution for the management of human sensitive data within the FEGA network is the Local EGA[19]. A Federated node can employ any technical solution as long as

[19] https://github.com/EGA-archive/LocalEGA

standards and interoperability are achieved. The Local EGA is designed to meet the requirements and is currently being used by several existing FEGA nodes. E3.2 described the technical components, but since then the software has been updated to version 2, with some new and updated features.

The components included in this implementation are:

- An inbox, to upload encrypted data files
- A long-term database to store files' headers, and payload
- A file storage interface to deposit files' payload
- A file ingestion pipeline
- A file distribution system
- A RabbitMQ server, to exchange messages with EGA

On the other hand, CEGA provides the following components:

- A Submitter Portal (SP) and API specific for each node, to register public metadata
- A Helpdesk Portal specific for each node, to manage registered metadata, and users.
- A DAC Portal and API (unique for the whole EGA), to register and manage data controllers, access requests and permissions.

○ 2.3 Technical details of the data life cycle

Figure 6 depicts the user journey for **metadata and file submission**. The user, in this case a FEGA submitter, needs to:

- Encrypt the data files with Crypt4GH and the public key of the node.
- Upload the files to the inbox of the node.
- Register the metadata in the SP of the node.

Ingestion of a file will be triggered as soon as it is used in the metadata. All files must be ingested and archived in order to complete the submission.

Figure 7 depicts the **ingestion** process, that consists of:

- Decrypting the file header with the service private key
- Decrypting the file payload with the data key
- Reencrypting the header with the master's (a.k.a. vault) public key
- Saving the header in the database.
- Saving the payload in the filesystem.
- Backing up the payload.

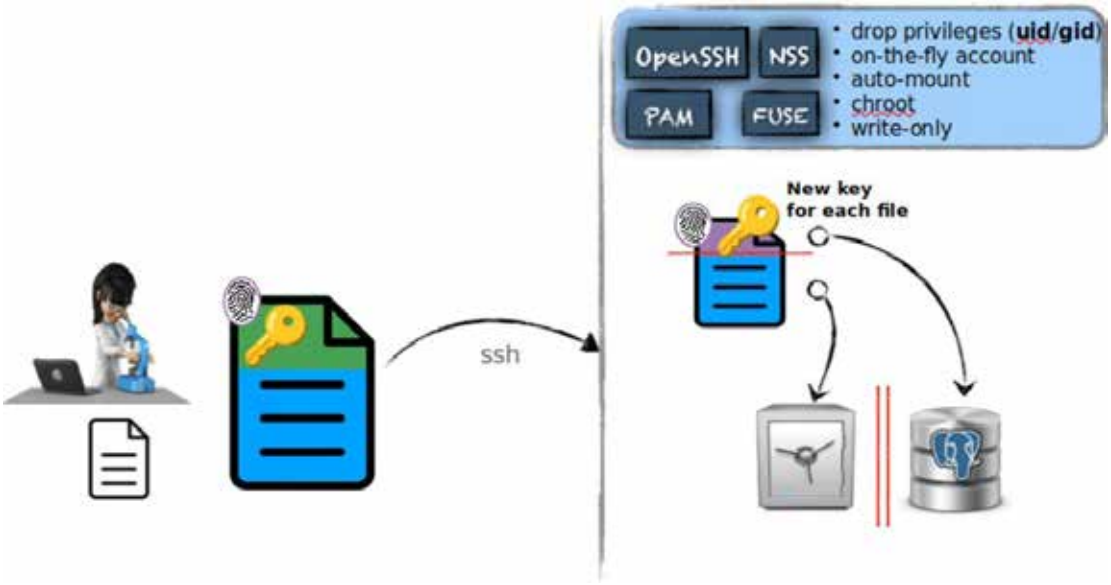

Figure 6. User journey for data submission

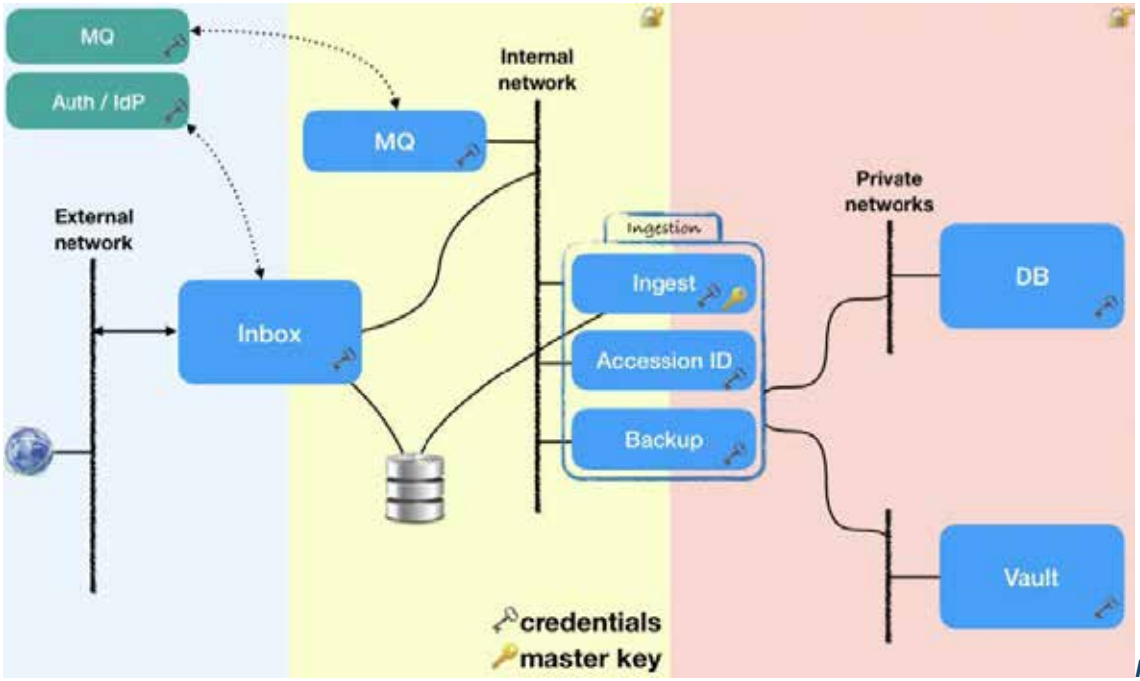

Figure 7. Data ingestion process with Local EGA

When the submission and ingestion are complete the following occurs:

- *Metadata release*: The submitter decides when metadata should be discoverable in the EGA website.
- *Data access request*: as soon as the release happens, any EGA user can request access to it.
- Access requests can be *granted* or *denied* in the DAC Portal.

When a request is granted, CEGA will send the node some information about the permission (which dataset) and the user (e.g. their public key), in the case where nodes use the DAC Portal to manage permissions. Nodes can also manage their own permissions and share this information back with CEGA through the DAC API.

The **distribution** process is depicted in Figure 8 and consists of:

- Recovering the file header from the database (DB).
- Decrypting it with the master's private key (DB).
- Reencrypting it with the user's public key (DB).
- Retrieving the payload from the file system (FS), stitching it together with the header (FUSE), and sending it through an encrypted channel (SFTP).
- The user can decrypt the file locally using their private key.

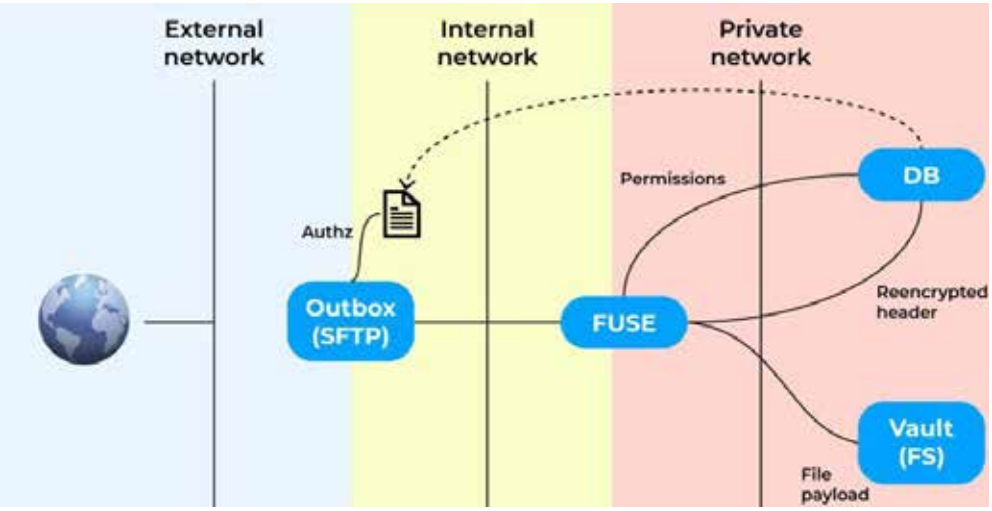

Figure 8. The distribution service in Local EGA

○ 2.4 Summary of the technical requirements

In summary, the technical steps for the data life cycle at a FEGA node have been described, with the example of the Local EGA solution. However, as mentioned above, nodes are free to employ any technical solutions they prefer, as long as the minimal level of services and interoperability are accomplished. Completing the end-to-end demonstration of technical capabilities is the last step in the technical branch of the path to join FEGA (refer to Figure 4). At the technical level, a Community node would need to meet the same requirements as a FEGA node, although the scale of operation would be considerably less. The Local EGA offers an out of the box solution to meet all the technical needs.

● 3 Legal requirements

○ 3.1 Governance structure

FEGA nodes typically operate using the same governance model as CEGA, where institutions submitting data are Data Controllers, usually represented by a DAC (Figure 9).

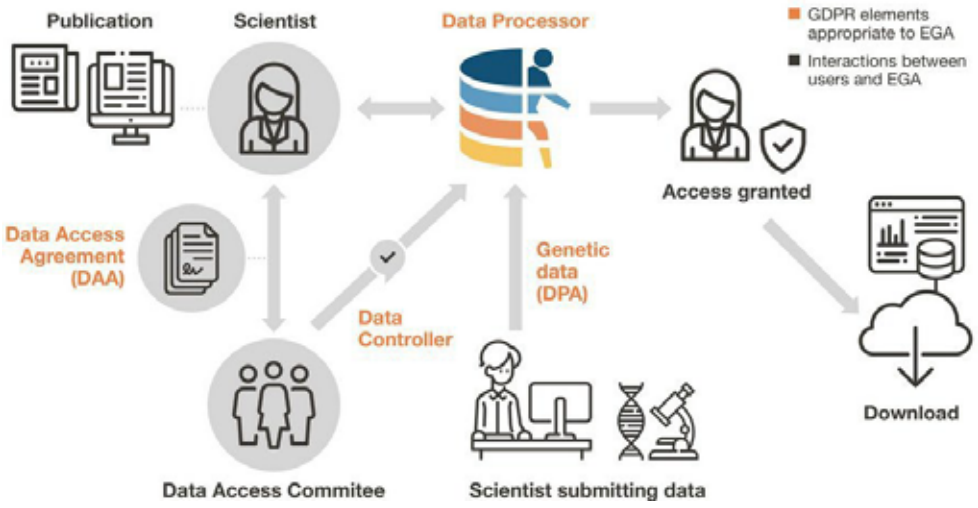

Figure 9. The EGA GDPR schema

A Data Processing Agreement (DPA) is signed between the submitting institution and the institution(s) hosting the FEGA node, which then acts as Data Processor, holding the rights to store the data and apply sanity and quality checks on them. Researchers interested in accessing data at any FEGA node must submit a request to the corresponding DAC for that data. The DAC then makes a decision to grant or deny the access request according to the appropriate law. If approved, researchers must sign a Data Access Agreement (DAA) with the

Data Controller before being given permission to access or download the corresponding data from the FEGA node where it was deposited.

Ethical, Legal and Social Implications (ELSI) for human genomic data sharing have been at the heart of FEGA's mission since its inception. The FEGA governance model supports compliance with the General Data Protection Regulation (GDPR) in the European Union (EU) and European Economic Area and other national and local legislation by ensuring that decisions to share data with third parties are always made by DACs. Respecting an individual's right to privacy is a key value of the FEGA Network and underlies many technical design choices, for example the choice of data encryption technique. Moreover, genetic data generated in a healthcare context is often subject to more stringent information governance than research data and must comply with national legislation.

○ 3.2 Legal aspects in more detail

When dealing with data protection subjects, specifically the EU's GDPR, it's important to understand the following key terms, while keeping in mind that the same terms may mean something different in other (non-EU) legal frameworks:

- **Data Controller**: This entity determines the *purpose and means* of processing personal data. It decides why and how personal data should be processed;
- **Data Processor**: A third-party entity that processes personal data on behalf of the data Controller. The Processor performs actions on the data, like storing, analysing, or transferring it, as instructed by the controller;
- **Joint Controllers**: When two or more entities jointly decide on the why and how of processing personal data, they are considered joint controllers. They share GDPR compliance responsibilities and must transparently define their respective obligations, usually through an agreement – a Joint Controller Agreement (JCA);
- **Data Processing Agreement (DPA)**: A legally binding contract between the data Controller and the data Processor. It outlines the scope of the processing, its nature and purpose, the rights and obligations of both parties, and specific measures for data privacy and protection. This document is vital for ensuring GDPR compliance as it mandatory per the Regulation;
- **Processing**: Under GDPR, processing refers to any operation or set of operations performed on personal data or sets of personal data. This includes a wide range of actions, such as *collecting, recording, organising, structuring, storing, adapting, retrieving, consulting, using, disclosing by transmission, disseminating, aligning, combining, restricting, erasing, or destroying* data. Essentially, processing encompasses anything that is done with personal data, from its initial collection to its final deletion.

○ 3.3 The DPA

As described at the beginning of this legal section, the DPA should be arranged between each FEGA node and their respective submitters. The EGA DPA[20] can be used as a template, but ultimately each node must create their own. In the case of Community nodes, given they will not be accepting third-party data and are most likely themselves the data controllers, a specific DPA may not be required, although this remains to be explored in more detail. Similar to a FEGA node, a Community node would need to evaluate what legal agreements should be in place between the data controllers and the Community node host institute.

○ 3.4 The FEGA Collaboration Agreement

To join the FEGA network as a FEGA node, the leading institute must sign the FEGA Collaboration Agreement[21], that was drafted and agreed upon by the five inaugural FEGA nodes. It outlines the legal commitments and responsibilities between the FEGA node and CEGA. FEGA nodes agree to a certain level of service and secure infrastructure for 4 years and agree to adhere to the terms of reference established by the FEGA Strategic[22] and Operations[23] Committees. In the path to join FEGA (Figure 4), the signing of the Collaboration Agreement occurs after the technical and governance paths merge, the maturity model self-assessment is complete, and the node has been approved by the FEGA Strategic Committee. Then, upon signing, they become a full FEGA node.

Such a collaboration agreement has yet to be put in place for the case of a Community node, although it is currently being drafted. The level of commitment and responsibility would be less than for that of a FEGA node, but some agreement to a certain level of security and a data distribution service would also need to be in place for a Community node.

● 4 Human requirements

Here we describe the human requirements for the establishment and maintenance of a FEGA node and how they compare with a Community node's need. They can be categorised in the following groups:

- Leading force
- Management and coordination
- IT development
- Helpdesk (HD) service

---

[20] https://ega-archive.org/assets/files/EGA_Data_Processing_Agreement_v1.4-unsigned.pdf
[21] https://drive.google.com/file/d/1QAKABRB5ZxAJlQSwFHaJ3e6TdtwZaCYs/view
[22] https://ega-archive.org/assets/files/ToR-Federated-EGA-Strategic-Committee-v1.1.pdf
[23] https://ega-archive.org/assets/files/ToR-Federated-EGA-Operations-Committee-v1.1.pdf

Community nodes would require less human force for establishment and maintenance (e.g. the same person could play different roles or different people would invest only some time on the activities related to the Community Node).

○ 4.1 Leading force

Each FEGA node would need one or multiple persons in charge of pushing the project ahead. They would own, shape and update the project's vision, in the frame of the FEGA mission. They would be responsible for:

- The node's short and long-term sustainability plan. That means that they have engaged proper stakeholders to ensure that the node can count with enough funding to provide the right services to the users.
- Overseeing the overall functioning of the node, making sure they comply with the quality level of services they have committed.

In the case of a Community node, this leading force could overlap with management and coordination, depending on the scale of the Community node, and whether or not there are significant legal or ethical challenges to face.

○ 4.2 Management and coordination

FEGA is a complex international network that requires consistent coordination efforts. Each node would need one or multiple persons in charge of their node's internal management and participation in federation coordination activities. They would be responsible for:

- Ensuring the smooth participation of the node to the Federation committees and working groups, making sure that the persons with the right expertise are representing the node in the right places.
- Envisioning and implementing the node's communication strategy.
- Ensuring the coordinated communication between the node and the other members of the federation, including timely sharing of relevant documents and SOPs.

In the current model, Community nodes are not expected to participate in the FEGA committees. Depending on the scale or the nature of the node, communication strategies could be in place or not, but they would not be expected as in the case of a FEGA node.

○ 4.3 IT team

FEGA and Community nodes will have the need for a technical infrastructure to function. Therefore, the node would necessarily need human power responsible for:

- Ensuring the smooth operations of the node's Local EGA system or other interoperable solution, including data security measures

- Troubleshooting with the help of CEGA system administrators, in case of issues
- Implement Federated EGA standards
- Share relevant information, acquired knowledge and SOPs (Standard Operating Procedures) with the partners

○ 4.4 Helpdesk

Each FEGA node needs to provide help desk service to their users. Therefore, the nodes need to allocate human resources to assist the users in need of:

- Help to submit or download data
- Information regarding the services, the documentations and the legal compliance

In the case of a Community node, a very minimal HD service would be required, only to support users wanting to access the Community Node specific datasets. The CEGA HD would support Community node internal submitters in the way general submitters are supported now.

○ 4.5 Summary of the human requirements

The human requirements for establishing and operating a FEGA node can certainly vary depending on the needs of the node, the scale of activities (e.g number of submissions and data access requests), the legal framework or other bureaucratic challenges specific to the country/node, and more. The needs during the establishment phase would also be different from the operational phase, with the latter requiring more technical staff, such as IT and HD, as operations increase, while the former would require more strategy, funding acquisition, long-term sustainability planning, and so on.

In contrast, a Community node would require less human force and would minimally need to agree to distribute data as requests are granted. Some IT expertise could be required when setting up infrastructure, but this would depend on the technical solution employed by each Community node.

● 5 Concluding remarks

Upcycling of data is crucial for empowering the scientific community to achieve the best possible advances for the benefit of society as a whole and instrumental in implementation of personalised medicine programs. The FEGA ecosystem is designed to meet these challenges, providing the means for worldwide discovery and sharing of genomic and phenoclinic data. WIthin IMPaCT, researchers are able to join the FEGA network through any of the three node options, depending on their needs and requirements.

The current 7 FEGA nodes are already serving the needs of researchers across Europe, with other national nodes outside Europe also very close to joining, further broadening the network.

Establishment of Community nodes in the FEGA network would enable projects, institutes, research hospitals, etc., such as those within IMPaCT, to also share data that they otherwise might be unable to share, increasing the value and the impact of their research and of the FEGA network.

Acronyms and Abbreviations

API	Application Program Interface
CEGA	Central EGA
DAA	Data Access Agreement
DAC	Data Access Committee
DB	Database
DPA	Data Processing Agreement
EGA	European Genome-phenome archive
EU	European Union
FEGA	Federated EGA
FHD	Federated Human Data
GDPR	General Data Protection Regulation
HD	Helpdesk
JCA	Joint Controller Agreement
SOP	Standard Operating Procedure
SP	Submitter Portal
WP	Work Package