Requisitos de un Nodo Local EGA

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN
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			
Period	Enero 2021 – Diciembre 2023			
Work Package	WP3 – Genomics			
Task	T3.1 - Implementación de sistemas de almacenamiento de información genómica primaria mediante instancias de local EGAs			
Deliverable	E3.1. Requisitos de un Nodo Local EGA			
Version	1.1.1			
Due Date	31/12/2021		***Approval Date***	17/05/2023
Responsable	CRG			
Dissemination Level	X	PU	Public	
		CO-IMP	Confidential, only IMPaCT pillars members, including the evaluation commission from IMPaCT.	
		CO-DATA	Confidential, only IMPaCT-Data members, including the evaluation commission from IMPaCT.	

Authors		
Organization	***Name***	***Role***
BSC-CNS	Lidia López	Coordination
EGA-CRG	Teresa D'Altri	Author
EGA-CRG	Jordi Rambla	Author
BSC-CNS	Salvador Capella-Gutierrez	Reviewer
FPS	Joaquín Dopazo	Reviewer

Version History			
N.	***Date***	***Description***	***Author***
v 0.0	08/10/2021	Created	L.López (BSC-CNS)
v 0.1	05/11/2021	First sections + scheme shared with reviewers	Teresa D'Altri (EGA-CRG)
v 0.9	29/11/2021	Full text added by authors	Teresa D'Altri (EGA-CRG)
v 1.0	24/12/2021	Full definitive version	Teresa D'Altri (EGA-CRG)
v 1.1	17/05/2023	Visibility changed to public and approved	Comité Directivo
v 1.1.1	14/06/2023	Cambio de formato para publicar en la Web de IMPaCT	David Velasco (ISCIII)

Content

Tables

Figures

Executive summary

The European Genome-phenome Archive (EGA) is a service for law-compliant storing and sharing of all types of genomic data and associated metadata. The EGA is currently a centralized service co-managed by the EBI and the CRG. In the last year, several countries have started working, under the coordination of the EGA, to establish a Federation. The Federated EGA (FEGA) is envisioned as a network to support national data management requirements in different European countries. The FEGA configuration is composed of Central EGA, Federated EGA nodes and Community EGA nodes:

- **Central EGA** offers international submissions and helpdesk support, currently EGA co-managed by EMBL-EBI and CRG.
- **Federated EGA nodes** offer EGA services to researchers within their national jurisdiction.
- **Community EGA nodes** are individual institutions or initiatives with human genetic data intended to be shared with the research community.

Countries, research centres, hospitals or any entity with the wish to join FEGA will have to face series of technical, human and legal requirements. Those are described in this document with the objective to provide a first-level information for the interested researchers. More detailed, experienced-backed and structured information will follow in the format of the next IMPaCT-Data deliverables.

Introduction

Audience

This deliverable is envisioned as a useful document for those institutes who would like to establish a Federated EGA node. We provide the basic information they need to evaluate the level of commitment they want to achieve, and the requirements they will face.

Topic

The mission of IMPaCT-Data project is to set the basis for the successful set up of the Spanish personalized medicine program; this deliverable is a tool for all those entities (hospitals, research centres, etc.) who will participate in the management, analysis and sharing of genomic data and associated metadata. This data is the basis for the development of personalized medicine end therefore, empowering the entities of our countries with tools, knowledge and network to manage this type of data is instrumental to the final mission of the IMPaCT-Data project.

Relation with other deliverables

This deliverable is the first of the WP3, and the work started in this document will be complemented with the next deliverables (E3.2 and E3.3) of this same WP. It is related to the deliverable E2.1 (due at the same M12 deadline) given they are both exploratory work for the implementation of genomic data sharing and analysis platforms. As described below, this deliverable is also connected with WP2 in relation to the infrastructure and WP5 in relation to Data discoverability.

Structure of the deliverable

The deliverable is structured in the following sections:

1. State of the Art regarding the European Genome-phenome Archive
2. Technical requirements. Where we describe the technical infrastructure needed to establish and maintain a Federated EGA node or Community.
3. Human requirements. Where we describe the human requirements to establish and maintain a Federated EGA node or Community.
4. Legal requirements. Where we describe the legal requirements to establish a Federated EGA node or Community.

1 State of the art regarding the European Genome-phenome Archive

The European Genome-phenome Archive (EGA) is a service for law-compliant storing and sharing of all types of genomic, phenotypic and clinical data and associated metadata describing every coherent data set.

How does the EGA work? Shortly, researches (in research or healthcare institutes) submit their data through an online process and such data is stored with an encryption algorithm (Crypt4GH, implemented together with the Global Alliance for Genomics and Health -–GA4GH –-) in the EGA archive. Submitted studies can be browsed on the EGA website, so that interested researchers can request access through an online application. The Data Access Committee (DAC) for each dataset is responsible of granting or denying access to their managed data. The decision and responsibility is therefore always retained by the data controllers, while the EGA facilitates the operations ensuring security and providing necessary documentation (like Data Processing Agreements, DPA and Data access agreements, DAA).

In the last 10 years, most human omics data has been generated in the context of research consortia while recently there has been an emergence of large cohorts of human data generated by healthcare initiatives. Many countries in Europe now have nascent personalised medicine programmes. In this context, EGA identified the need for the development of a federated network to enable secure sharing of data whilst enabling genetic data to remain within the jurisdiction in which it was generated. The Federated EGA (FEGA) is designed to support national data management requirements for genomic and clinical data collected from citizens as part of healthcare or biomedical research projects. It includes a secure authorised access mechanism to support research use of these data across Europe and worldwide. Central EGA has engaged with over 14 ELIXIR countries to develop the federation model. The EGA federated configuration is composed of Central EGA, Federated EGA nodes and Community EGA nodes:

- **Central EGA** offers international submissions and helpdesk support, currently EGA co-managed by EMBL-EBI and CRG.
- **Federated EGA nodes** offer EGA services to researchers within their national jurisdiction.
- **Community EGA nodes** are individual institutions or initiatives with human genetic data intended to be shared with the research community.

The specific requirements are summarized in Table 1 and described in the following sections of this document. Those requirements are also described in the official FEGA documents collected in the EGA website[1].

[1] https://ega-archive.org/federated

For IMPaCT-Data, the proposed architecture is:

- **Central EGA** (represented by the CRG), that would make data discoverable worldwide and provides the link with ELIXIR network and other international initiatives.
- **Spanish Federated EGA Node**, hosted and managed by the BSC
- **EGA Community nodes**, formed by current disease centred networks (e.g. CIBERER) or networks from areas or Comunidades Autónomas (CCAA)

Table 1. Specific requirements of the different FEGA entities

	Central EGA	Federated EGA node	Federated EGA community
Data submission	Offers international submission service	Offers submission service in a particular jurisdiction	Does not offer an external submissions service
Helpdesk support	Provides international helpdesk support	Provides helpdesk support for submitters in its jurisdiction and for approved users of data managed at its facilities	Provides helpdesk support only for approved users of data managed at its facilities
Data distribution	Manages worldwide data distribution for data hosted at central EGA	Manages worldwide distribution for data hosted at Federated EGA node	Distribution for data hosted at Community EGA node

2 Technical requirements

The local EGA documentation is publicly accessible at the Local EGA Read the Docs[2]. The reference code repository is publicly accessible at GitHub[3]

The technical requirements to establish a Federated EGA node are mainly represented by the infrastructure that would allow submission and distribution of the data. The technical solution should do such, while sharing metadata with the central EGA.

The Local EGA is a software which enables sharing sensitive genetic data and its associated metadata. As depicted in Figure 1, the distributed solution created works so that submitters upload encrypted files into a Local EGA inbox, located in the relevant jurisdiction. The ingestion pipeline moves the encrypted files from the inbox into the long-term storage, and saves information in the database. In the process, each ingested file obtains an Accession ID, which identifies it uniquely across the EGA. The distribution system allows requesters to securely access the encrypted files in long-term storage, using the accession id, if permissions are granted by a Data Access Committee. Files are encrypted whether in transit or at rest. Files are stored using the Crypt4GH4 file format. The metadata of the encrypted files and the permissions to access them are located at Central EGA and at the Federated EGA Node.

[2] https://localega.readthedocs.io/en/latest/

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

[4] http://samtools.github.io/hts-specs/crypt4gh.pdf

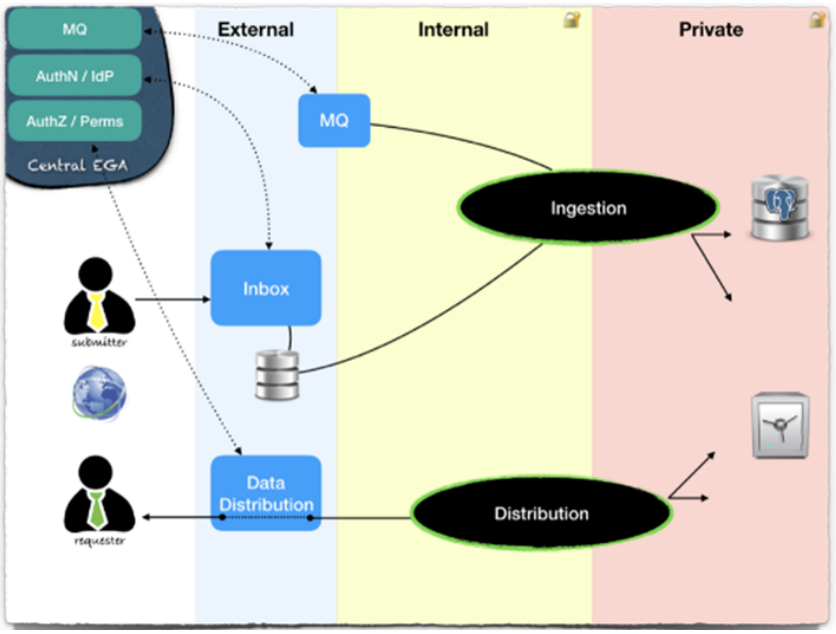

Figure 1. Scheme of the Local EGA technical infrastructure created by the developers of the EGA-CRG.

Description of the components of the Local EGA:

- *Inbox*. the data submitter first logs onto the Local EGA's inbox and uploads the encrypted files. Login credentials are provided by Central EGA.
- *Ingestion pipeline*. For every uploaded file, Central EGA receives a notification that the file has landed. The data submitter then prepares a submission and Central EGA sends an ingestion trigger to the connected Local EGA, and the files are moved securely into the long-term storage.
- *Metadata release*. After a file is successfully ingested (including a backup confirmation), has an accession id, and the metadata is marked as released, the file becomes available for download. If a file access has been granted by a DAC, the file can be served in Crypt4GH format to the requester.
- *Data access*. Ownership of the data and related rights are retained by the data submitters and Data access control is delegated to a Data Access Committee (DAC). Central EGA and Local EGAs are never to be considered the files' owner. Therefore, permissions are granted by the DACs (and not Central EGA). Once the permission has been granted, the data is made available to the data requesters.
- *Data distribution*. If a file access has been granted by a DAC to a data requester, the file can be served in Crypt4GH format.

We have developed extensive documentation for the implementation of the Local EGA in the federated nodes and communities: Local EGA — Local EGA. The documentation describe all needed information regarding:

- the components of the Local EGA

- Message interface (API) Central EGA-Local EGA
- Connection settings to Central EGA
- Ingestion
- Distribution over HTTPs
- An Inbox implementation, with centralized NSS and file updates notifications
- The encryption used across the sensitive storage

We have also made available a reference implementation: This repository contains the necessary code and instructions to set up a Local EGA.

The above described functionalities of authentication, authorisation and distribution could be used and integrated to archive the objectives of WP2 of this same IMPaCT-Data project.

The Federated EGA has an organ, the Operations Committee, formed by members of Central EGA and the federated nodes, responsible to coordinate all the activities related to the implementation of the Local EGA. Developers of the EGA-CRG team, who created the Local EGA provide support to the nodes implementing it and ensure an efficient knowledge exchange.

The EGA Community concept includes a variety of projects and organizations, therefore, there is also a heterogeneity of capacities in terms of technical solutions for data managing and sharing. The solutions range from having no infrastructure at all to having a fully functional and mature solution in place. By contrast, candidates to be Federated EGA Nodes usually are organizations or institutions whose goal is to provide services to their respective communities, therefore, they are ready for building and deploying advanced solutions for data management and sharing.

A solution that covers the whole lifecycle of data and its uses must include components for Data Discovery, Data Management and Data Sharing (Figure 2). EGA Community nodes could leverage the Local EGA software as Data sharing platform, in case they do not have one in place already.

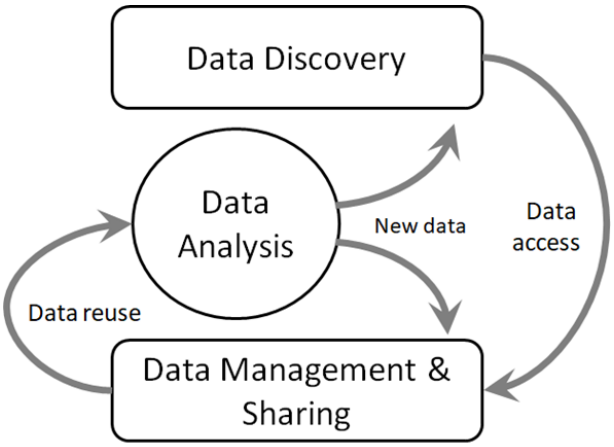

Figure 2. Scheme illustrating the Data Discovery, Data Analysis and Data Sharing components.

The systems architecture of the solution should rely on the existing rules and protocols established at the institution that deploys a Federated or EGA Community node, but, all of them

are strongly recommended to follow the standard model for secure setups that are accessible from Internet.

The actual technical infrastructure dimensioning would depend enormously on the volume of data, number of users and complexity of the analysis (processing) to be done. A rough estimation, to be used as a guide, could be based on the fact that a whole human genome sequence occupies between 30 and 200 GB, depending on the sequencing coverage.

The components of Data analysis and Data Discoverability will be developed in alignment with the work developed by the WP5.

3 Human requirements

The human requirements for the establishment and maintenance of a Federated EGA node are multiple and diverse. They can be categorized in the following groups:

- Leading force
- Management and coordination
- IT development
- Help desk service

Leading force. Each 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 Federated EGA 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.

Management and coordination. Being the Federated EGA a complex international network, it requires consistent coordination efforts. Each node would need one or multiple persons in charge of their node's internal management and participation to 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 of 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.

IT team. As detailed in Section 1, each Federated node 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
- Troubleshooting with the help of Central EGA system administrators, in case of issues
- Implement Federated EGA standards
- Share relevant information, acquired knowledge and SOPs (standard Operating Procedures) with the partners

Help desk service. Each 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

4 Legal requirements

An institute or entity wishing to establish a Federated EGA node should take in account a number of legal requirements. These requirements are enclosed and detailed in the legal agreement that must be signed to formalize the entrance into the Federation.

In summary, the future node should consider the need to:

- Have Data protection polices established and well documented. Policies would necessarily be aligned with European, National and institutional regulations.
- Have legal agreement documents needed to comply with the above-mentioned policies (i.e. GDPR). These documents, such as DPA (Data processor agreement) and DAA (Data Access Agreement), need to be provided by the node to the users to allow the management of the data.

The Federated EGA's main decision-taking organ, the Strategic committee, formed by members of central EGA and the federated nodes, is responsible to supervise that the federated node comply with their legal requirements. In addition, it provides information and support.

5 Conclusions

In conclusion, this document briefly describes some of the requirements needed to establish a Federated EGA node. More work is needed to:

- Better define the requirements. This will be enabled with the experience of the first nodes entering the federation during the year 2022.
- Generate a reliable and experience-tested tool for the evaluation of the level of maturity of the entities wishing to become a FEGA node, including all the requirements.

- Adapt and model the requirements to the different level of commitments that the entity might want to achieve, i.e. FEGA node or FEGA community.

This work will be carried out during the deployment of the IMPaCT-Data project, and consequentially delivered in the following deliverables, as per Grant Agreement.

Glossary

IMPaCT	Infraestructura de Medicina de Precisión asociada a la Ciencia y la Tecnología
IMPaCT-Data	Programa de ciencia de datos de IMPaCT
CD	Direction Committee (Spanish initials, Comité Dirección)
WP	Work Package
FEGA	Federated European Genome-phenome Archive
DPA	Data processor agreement
DAA	Data access agreement
EGA	European Genome-phenome Archive
DAC	Data access Committee
SOP	Standard Operating Procedure
EBI	European bioinformatics Institute (Cambridge, UK)
CRG	Center for Genomic Regulation (Barcelona)