EGA Federation: Structure and organisation

Authors: Thomas Keane, Jordi Rambla, Dylan Spalding, Paul Flicek, Arcadi Navarro, Helen Parkinson

January 2020, Version 1.1

The EGA project is currently a collaboration between EMBL-EBI and the CRG, regulated by agreements between the two institutions. This document is a proposal for the structure of an EGA federated network and service expectations. We propose to organise the EGA into three types of nodes: Central EGA, Federated EGA nodes and EGA Community nodes; we outline the goals of such an organization, and summarize the commitments and services provided by the nodes.

1 Introduction

The EGA is a resource for secure archiving and sharing of all types of potentially identifiable genetic and phenotypic data resulting from biomedical research projects. It provides access to human research data, fosters data re-use, enables reproducibility and speeds up biomedical and translational research in line with the 'FAIR' (Findable, Accessible, Interoperable, and Reusable) principles. The EGA is an ELIXIR Core Data Resource and one of the inaugural Global Alliance for Genomics and Health (GA4GH) driver projects. The EGA was set up by the EMBL-EBI in 2008 and since 2013 the resource has been jointly operated with the CRG.

In the last 10 years, most potentially identifiable human -omic data was generated in the context of research consortia. We have recently seen the emergence of large cohorts of human samples not only from research, but also from national or regional healthcare initiatives. Many countries now have nascent personalised medicine programmes meaning that human genomics is undergoing a step change from being a predominantly research-driven activity to one funded through healthcare. We envisage that a significant subset of this data will be made available for secondary research and models for this are already operating in some countries (e.g. Genomics England). However, genetic data generated in a healthcare context is subject to different information governance than research data. Healthcare is subject to national laws and it is unlikely that health data from one country will be exported outside regional or national jurisdictions. Strategic discussion with national stakeholders, including ELIXIR's National nodes, confirm the need for non centralised EGA-based systems that ensure the FAIRness of health data.

In coordination with ELIXIR and the GA4GH, the EGA is committed to the establishment of a federated and interoperable network of human data resources to enable data sharing in Europe, while guaranteeing a high level of visibility and prestige of the EGA brand.

2 EGA Federation Structure: EGA Central & Nodes

In what follows, we summarize the commitments and services offered by EGA federated nodes. The node descriptions below are subject to change and are also subject to a legal

agreement between the partner institutes. Appendix 1 provides the reader with definitions of service terminology (e.g. submission, data distribution, ETL, helpdesk etc.).

2.1 EGA Central

EGA Central (currently EMBL-EBI and CRG) offers submission, long-term data archiving, and data distribution to the international scientific community. EGA Central will:

1. Governs node entry and exit conditions to the federated network according to the legal agreement
2. Govern the use of the EGA brand and its use (e.g. grant applications, websites, public presentations)
3. Convenes and interacts directly with the EGA Scientific Advisory Board
4. Coordinates the admission of new nodes, and the assignment of nodes to the appropriate category. Reviews node performance and reports to the EGA Strategic Committee and SAB
5. Support international submissions, ETL pipeline, and data distribution of data held at EGA central nodes
6. Offers international helpdesk services to all EGA users
7. Provides helpdesk and training to Federated Nodes
8. Coordinates the development and maintenance of EGA APIs, tools, and resources (see EGA technical roadmap)
9. Offers coordination and support to EGA Local and EGA Community Nodes to operationalise their node
10. Provide rotating chairs of the EGA Federation Strategic Committee, held quarterly to report on the activity of nodes, review operations and performance, and to gather new requirements (see committees ToR documents).

2.2 Federated EGA Node

Federated EGA nodes are nodes that provide full EGA services (external helpdesk, submissions, archiving, and distribution) for a particular jurisdiction. The primary motivation for establishing a Federated EGA node is to enable the discovery and access of data that for consent or other reasons is required to be archived within the relevant jurisdiction (e.g. data for research generated in a healthcare context, but not limited to this). Publicly shareable metadata about studies/datasets archived at these nodes will be shared with Central EGA to enable discovery. Federated EGA nodes will offer the same APIs and interfaces as the Central EGA, and will provide independent data distribution to users. Dataset user access permissions should be synced with Central EGA. The minimum membership term will be four years.

Federated EGA nodes will:

1. Implement the EGA ETL pipeline (see Appendix 1)
2. Provide helpdesk services to submitters and Data Access Committees (DACs) within the jurisdiction (as defined in legal agreement)
3. Provide helpdesk services for all authorised users to access data held at the node
4. Provide a consistent user experience according to EGA SOPs

5. Provide data distribution for datasets held at their node
6. Submit validated public metadata with Central EGA within two business days of submission and validation at federated node
7. Contribute in the development of common EGA APIs, tools, and resources
8. Attend the EGA Federation Strategic committee
9. Provide summary level operational reporting to Central EGA
10. Optionally attend at the EGA SAB as an observer

2.2 EGA Community Node

EGA Community nodes may be individual institutions, national or regional healthcare genomics initiatives, consortia of institutions, and project-specific consortia.

EGA Community nodes regulate access to their own data and provide data distribution to all locally approved users. It is recommended that they offer the same APIs and interfaces as the Central EGA but can develop their own APIs and interfaces. Publicly shareable metadata about studies/datasets archived at these nodes will be shared with Central EGA to enable discovery. In order to provide a seamless user experience it is recommended that EGA Community instances share EGA's production practices and APIs for helpdesk, data discovery, permissions, authentication and authorization, and data distribution. The minimum membership term will be four years. EGA Community Nodes will:

1. Offer data distribution services for data archived at their node
2. Offer helpdesk to external users for the provision of data access and distribution
3. Primarily accepts submissions from within their organisation
4. Share public metadata with Central EGA to enable data discovery

3 Node Services and Infrastructure

	EGA Central	**Federated EGA**	**EGA Community**
External services			
Data submission (see Appendix 1)	Offers international submissions service.	Offers submissions service for submissions in a particular jurisdiction.	Does not offer an external submissions service.
Helpdesk support (see Appendix 1)	Provides external international helpdesk	Provides helpdesk support for submitters in its jurisdiction and for approved users of data managed at its facilities	Does not provide external helpdesk support for submitters, but only for approved users of data managed at its facilities

Data Access Committees	Provides support and tooling for DAC activities	Conforms to Central EGA DACs SOPs	Manages the DACs linked to their data/studies. Integration with EGA DAC tooling is recommended to avoid fragmentation of the user experience
Data Distribution (see Appendix 1)	Manages worldwide distribution for data hosted at Central EGA	Manages worldwide distribution for data hosted at Federated EGA node	Distribution for data hosted at EGA Community
Discovery Services	Provides worldwide data discovery services	Integrates with the EGA data Discovery Services	Integrates with the EGA data Discovery Services
EGA Branding			
EGA logo and branding	Full use of EGA branding in funding requests and external communications	Full use of Federated EGA branding in funding requests and external communications	Full use of EGA Community branding in funding requests and external communications
Infrastructure			
Compute and storage	Sufficient compute and storage to complete submission, QC, ETL, and distribution	Sufficient compute and storage to complete local submission, QC, ETL, and distribution	*N/A* *Resource requirements are out of scope for this node category*
Authentication and authorisation (see Appendix 1)	Hosts or proxies the AAI services to other nodes	Conforms to EGA APIs, is integrated and compatible with Central EGA	Recommend integration with Central EGA
APIs	Develops and maintains EGA APIs, e.g. permissions, distribution, submission	Conforms to EGA APIs, e.g. permissions, distribution, submission	Recommended to use the EGA APIs to allow a smooth integration of services
Data replication (see Appendix 1)	At least two geographically	At least two geographically	Determines local data replication

	distinct copies of all data	distinct copies of all data. One of the copies may be located in EGA Central.	policies
Helpdesk service (see Appendix 1)	Hosts and coordinates EGA helpdesk system. Integrated helpdesk system with other central EGA nodes. Provides operational contacts for all EGA nodes.	Helpdesk systems integrated with Central EGA. Must have a designated operational contact person available for Central EGA requests	Use their own user request/issue management mechanism. Must have at least a designated contact person available for Central EGA requests
Software development licensing	Permissive licensing of all EGA code, e.g. Apache 2.0[1].	Permissive licensing of all EGA-related code, e.g. Apache 2.0	Recommend permissive license for EGA-related code
EGA Organisation			
(Central) EGA Strategic Committee (see Appendix 1)	Membership	Not member	Not member
EGA Federation Strategic Committee (see Appendix 1)	Co-chairs	Membership	Observer
EGA Federation Operations Committee (see Appendix 1)	Co-chairs	Membership	Observer
(Central) EGA Scientific Advisory Board	Required attendance, host of annual SAB meeting	Optional attendance as observer	Optional attendance as observer
Basis for membership	Legal agreement for entry/exit	Legal agreement for entry/exit depending on approval by EGA Central. Minimum term four years.	MoU (or equivalent) for entry/exit signed with EGA Central. Minimum term four years.

[1] Apache 2.0 license: https://www.apache.org/licenses/LICENSE-2.0

Appendix 1 - Glossary of Terms

Activity or Organisation	**Definition of Activity**
Data submission	A node has the capability to offer: • A secure user account login system • Secure upload of raw data and metadata • Validate the raw data and metadata according to current EGA protocols Full details are provided in additional Federated EGA technical and SOP documents.
ETL pipeline (Extract, transform, and load)	The software pipeline that takes an EGA submission (data+metadata) and • Validates the input (data integrity, syntax, and semantics) • Archives the data in a replicated long term storage • Loads the metadata into a queryable database system
Data distribution	A service that enables authorised users to securely download either entire datasets (e.g. secure FTP, Aspera) or subsets (e.g. htsget API)
Helpdesk support	A service where external users can send operations support queries such as: • Completing data submissions • Finding datasets • Data download or access issues The service should have sufficient personnel resources to provide timely responses (e.g. <48 business hours).
Data QC	A service that can generate and publicise high level quality metrics from submitted data, primarily data from genome sequencing data.
EGA Strategic Committee	A committee formed of senior management from Central EGA nodes that provide the strategic governance for the project.
EGA Federation Operations Committee	A committee formed of operational supervisors and leads from the EGA nodes

	to review, communicate, harmonise operational aspects of the EGA nodes (e.g. service technical compatibility, updates in policies and SOPs, helpdesk service capacity). See ToR for committee for more details.
EGA Federation Strategic Committee	A committee formed of senior management representatives from all the EGA federated nodes to provide strategic governance and policy setting for the federated nodes. See ToR for committee for more details.
Data replication	A geographically separated physical copy of the raw data and metadata.
Authentication and authorisation	Authentication - verification of a person's identity. Authorisation - verification from a data controller that a user has permission to access data.