Federated EGA Node Operations Guidelines

Authors: Giselle Kerry, Thomas Keane, Jordi Rambla, Dylan Spalding, Paul Flicek, Arcadi Navarro, Helen Parkinson, Mallory Freeberg, FEGA Operations Committee

Version 2.0, approved May 2022

Introduction

This document provides an overview of the areas which require resources in order to establish and operate a Federated EGA node. It is based on more than 10 years' experience of EMBL-EBI and CRG operating the EGA and initial experiences of the inaugural Federated EGA nodes. The operational areas of responsibility are inspired by the Federated EGA Maturity Model (MM[1]) which guides establishment and operation of nodes as part of the Federated EGA network of human data resources.

Governance and Legal

Before a Federated EGA node can start to accept and process sensitive human data, governance and legal frameworks must be established, data protection policies must be in place, and legal agreements must be drafted and signed by the relevant parties. This work requires resources for developing and maintaining these components, often in collaboration with institutional legal departments and legal experts, such as Data Protection Officers.

Sufficient resources are required to support the following responsibilities:

- Document node structure, vision, strategy, and sustainability model
- Establish and implement node KPIs
- Draft and review data protection policies and legal agreements (e.g. data processing agreement) between node and end-users
- Review and sign the Federated EGA Node Collaboration Agreement
- Perform and monitor risk assessments

Helpdesk

The Helpdesk is the first point of contact for submitters, Data Access Committees (DACs), and data requestors and should operate using an email-based ticket tracking system to provide an auditable record of all external interactions, to track response times, and to facilitate staff handover of ongoing issues. All helpdesk tasks should be carried out according to a set of standard operating procedures (SOPs), escalating issues to another team, when needed. Helpdesk personnel often contribute to outreach and user engagement activities, as well.

[1] https://inab.github.io/fega-mm/

Sufficient resources are required to support the following responsibilities:

- Establish a Federated EGA node Helpdesk team
- Develop and maintain SOPs and documentation for all Helpdesk activities
- Ensure that Helpdesk tasks are performed in accordance with SOPs and relevant security guidelines, and that user inquiries are responded to in a reasonable timeframe
- Assist submitters with data preparation, data upload, metadata submission, DAC creation, and data access policy creation
- Assist DACs with account and permissions management
- Assist requesters with data discovery, applying for access, and downloading data
- Assist all users with general queries
- Maintain and develop relevant user-facing documentation (e.g., FEGA node website, submission guidelines, training materials)

Operations

The operations team runs, monitors, and provides support for the Exchange, Transform, and Load (ETL) processes that are performed on all submissions prior to long term archiving, indexing, and presentation of datasets. The operations team also provides technical support for issues that Helpdesk personnel cannot directly resolve. Some of the responsibilities could also be shared or carried out by Helpdesk personnel.

Sufficient resources are required to support the following responsibilities:

- Establish a Federated EGA node Operations team
- Develop and maintain SOPs and documentation for all operations activities
- Support data submission, archive, release, access, and distribution services. Tasks can include:
    - Validation of submission integrity, syntax, and semantics
    - Data file encryption and archiving
    - Metadata validation, loading, and indexing
    - Ensuring data release process completes successfully
    - Monitoring infrastructure and resource usage and capacity
    - Monitoring operational integrity of services
    - Deploying data submission, access, and distribution services
    - Coordination of maintenance periods and service downtime
- Support the Helpdesk team with operational issues, as needed
- Deploy and maintain services that interface with Central EGA
- Deploy and maintain FEGA node-specific services in production with documented specifications

Additionally, the following internal service support features are recommended:

- Logging and tracking
    - *Purpose*: To provide a fully queryable audit trail of all data submitted to the archive, ETL operations, user account creation, data access permissions, and data distribution.
    - *Features:* Log and store all actions/events related to FEGA node services; sufficient retention policy that aligns with relevant governance and legal

frameworks; appropriate security measures in place to prevent unauthorised access to sensitive or personal information (e.g., user names, IP addresses).
    - *Central EGA solution*: Kibana (ELK stack) and databases (MySQL and Postgres) to manage monitoring and logging information accessible only by relevant staff members.
- Helpdesk issue tracker
    - *Purpose*: Issue tracker system for all external user requests and interactions to be tracked and traced transparently.
    - *Features*: Access restricted to the appropriate Helpdesk and supporting staff; support for semi-/automating responses; single point of contact for users; transparency for relevant staff to track ongoing issues and reference previous issues/resolutions.
    - *Central EGA solution*: Best Practical RT Tracker.
- Operations issue tracker
    - *Purpose*: Internal tracking system for operational issues and resolution.
    - *Features*: Access for the appropriate operational and supporting staff and traced; transparency for relevant staff to track ongoing work and reference previous work.
    - *Central EGA solution*: JIRA.
- Software development work tracker
    - *Purpose*: Internal tracking system for software development work.
    - *Features*: Access for the appropriate software development and supporting staff and traced; transparency for relevant staff to track ongoing work and reference previous work.
    - *Central EGA solution*: JIRA.
- Documentation and SOP system
    - *Purpose*: A system for storing, tracking, and versioning internal operational documentation such as SOPs.
    - *Features: Ability to version and track changes, share for review and implementation, and link to supporting materials.*
    - *Central EGA solution*: Confluence.

Software Development

Interoperability between nodes in the Federated EGA network are based on a set of shared APIs (to be published separately) for services that require inter-node communication. Examples include for sharing submission metadata with the Central EGA, for managing data access via Central EGA, and for providing users with a unified interface for data access. The Local EGA solution[2] provides a basic implementation of these services, although many nodes will contribute to the development of this software solution with the intention to use it for the local node. For nodes that have significant existing infrastructure, all APIs will be published so that nodes could participate in the network by implementing those APIs in their existing systems. Nodes are likely to develop additional software solutions for node-specific services according to their remits.

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

Sufficient resources are required to support the following responsibilities:

- Establish a Federated EGA node Software Development team
- Develop and maintain SOPs and documentation for all software development activities
- Adopt software development best practices for managing a production service
- Implement services to communicate with Central EGA microservices
- Implement the Local EGA software solution and/or contribute to developing the Local EGA software solution (optional)
- Implement node-specific software solutions for providing FEGA node services
- Implement technical benchmarking and performance tests, including compliance testing and stress testing
- Ensure sufficient technical resources are provisioned to run the developed/deployed services

Technical Infrastructure

Since EGA is a resource for sharing controlled access human genetic and phenotype data, all node IT infrastructure access needs to be accessible to only authorised staff members. All data transfers in/out of FEGA nodes must be encrypted while in transit. Further details of EGA security guidelines are documented[3]. Below are listed the components and protocols that must be in place for the operation of a production FEGA node with examples of the currently adopted technology at Central EGA.

Storage

- Submission staging storage
    - *Purpose*: To provide a temporary location for submitters to upload their data files and other submission materials prior to the ETL processes starting.
    - *Features*: Sufficient space for upload of data files from all submitters; monitoring services to track usage; designation and enforcement of limits; mechanisms to notify users of activity. Requirements depend on submission sizes and throughput; ability (or at least a plan) to scale capacity in response to user demand.
    - *Central EGA solution:* Large disk provisioned for submission staging areas, with a unique account provisioned for each submitter. Hard and soft space limits placed on each account, with monitoring emails sent to an internal mailing list when limits are reached.
- Permanent archival storage
    - *Purpose*: Store data in a secure, indexed and sustainable fashion, allowing for on-demand access and distribution.
    - *Features:* At least two geographically distinct backup copies; replication of all submitted data between archive locations prior to communicating to submitter that archiving is complete; sufficient i/o bandwidth for ingest and data distribution; ability or plan to scale capacity in response to user demand.

[3] https://ega-archive.org/files/European_Genome_phenome_Archive_Security_Overview.pdf

    - *Central EGA solution*: Object-based storage system such as CleverSafe. Could also be implemented via an appropriately encrypted Posix file management system.

Network / Transfer

- *Purpose*: To enable secure data upload to (data submission) and data access/download from (data distribution) FEGA node in a timely manner, as well as internal transfer of data within FEGA node for data archive and processing services.
- *Features:* Sufficient network bandwidth to support both data in and out of the FEGA node; availability of secure network transfer protocols; ability (or at least a plan) to scale capacity in response to user demand.
- *Central EGA solution*: Use of EMBL-EBI's high-speed, and UK academic high-bandwidth network (Janet) for moving data in/out of EGA over secure protocols such as Aspera, secure FTP, and Globus.

Compute

- *Purpose*: To ensure day-to-day node operations can be supported in a timely fashion.
- *Features*: Support ETL processes, encryption/decryption, quality assessments, calculations on what is being archived, and a development/testing environment.
- *Central EGA solution*: HPC compute farm and network storage in an isolated part of institutional network only accessible by authorised personnel. Compute resource requirements depend on submission throughput and scope of data processing activities performed.

Community Engagement

Sufficient resources are required to support the following responsibilities:

- Develop communication materials and channels, emphasising the commitment of the Federated EGA Node in research data sharing
- Develop and maintain training materials for the node's targeted users
- Customise communication package to Federated EGA Node stakeholders to be audience-specific (e.g., for end-users, funders)

Conclusion

This document provides an overview and descriptions of the areas which require resources in order to establish and operate a Federated EGA node. Its alignment with the FEGA Maturity Model enables interested nodes to understand what areas of responsibility are required for establishing and operating a node as part of the FEGA network of human data resources. This document will be periodically reviewed and updated as part of the ongoing collaboration between nodes in the FEGA Network. For example, it can be expanded to include additional examples of how FEGA nodes have met the resource requirements.