EGA Federated Node Operations

Authors: Giselle Kerry, Thomas Keane, Jordi Rambla, Dylan Spalding, Paul Flicek, Arcadi Navarro, Helen Parkinson
EMBL-EBI
Version 1.0, January 2020

1 Introduction

This document gives an overview of the operational areas which require resources in order to create a federated EGA node. The document is based on more than 10 years experience of establishing and operating the EBI and CRG Central EGA nodes. We provide a breakdown of the operational areas of responsibility into Helpdesk Services, Technical Operations, Software Development, and IT Infrastructure.

2 Helpdesk Services

The EGA helpdesk is the first point of contact for submitters, Data Access Committees (DACs), and data requestors). The helpdesk operates via 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. Helpdesk tasks are carried out according to a set of standardised operating procedures (SOPs). Helpdesk personnel also coordinate and deliver EGA outreach and user engagement activities.

Areas of responsibility include:

- Ensuring that all tasks are performed in accordance with EGA standard operating procedures (SOPs) and Security guidelines;
- Assisting all submitters in their requests and queries including, data preparation, data upload, metadata submission, DAC creation, and data access policy creation;
- Assisting DACs with account creations & DAC maintenance (e.g. changes to DAC structure, updating of contact details etc.);
- Advising requesters on how to search, apply for access, and download data
- Assisting with general EGA queries;
- Maintaining and developing documentation, e.g. EGA node website, submission documentation, user documentation;
- Sufficient staffing to provide a response in less than 48 business hours;
- Review and tracking of SOPs for all helpdesk activities: Security, submission, distribution and DAC management.

3 Node 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 group also provides technical support for issues that the helpdesk personnel cannot directly resolve. Some of the responsibilities could also be shared or carried out by helpdesk personnel.

Areas of responsibility include:

- Support the ETL (Extract, transform, and load) pipeline which receives data and metadata from submitters. Tasks include:
    - Validation: Checking submission integrity, syntax, and semantics;
    - Operate data encryption and archiving service;
    - Metadata validation, loading, and indexing;
    - Monitor the usage of the upload staging disk space;
    - Monitor that files are moving consistently through pipelines and into archive;
    - Monitors the usage of the IT infrastructure to ensure efficient resource usage (e.g. disk storage in submitter upload areas) and all ETL and data distribution services are operating as expected;
    - Plan and coordinate maintenance and downtime of services.
- Second level operations support to help-desk
    - Assist help-desk in archiving issues/queries;
    - Assist help-desk in download issues.
- Data distribution service
    - Monitoring and operating the service;
    - Ensuring data release process completes successfully;
    - Deploying EGA data distribution services.
- Provision of publicly accessible API endpoints for EGA federated network uptime monitoring.

4 Software Development

The federated EGA network will be based on a set of public APIs (to be published separately) for services that require inter-node communication, e.g. MetaData API for sharing submission metadata with the Central EGA, Permissions API for managing data access via Central EGA, Data Distribution API for providing EGA users with a unified interface for data access. The Local EGA solution[1] will provide an implementation of these services, and we anticipate that many of the 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 IT infrastructure, we will publish all APIs so that nodes could participate in the network by implementing those APIs in their existing systems.

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

Responsibility for:

- Implementation of the EGA APIs within existing node IT infrastructure (optional) and/or contributing to Local EGA software solution[2];
- Implementation of metadata sharing service with Central EGA nodes;
- Deployment of new versions of EGA APIs
    - e.g. Permissions API, Metadata API, Data distribution API;
- IT infrastructure: A testing/development instance of the EGA node (compute+storage required).

5 IT 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 node staff members. All data transfers in/out of the EGA node must be encrypted while in transit. Further details of EGA security guidelines are documented here. Below we provide the components and protocols must be in place for the operation of a production EGA node with examples of the current adopted technology at Central EGA:

Helpdesk issue tracker

- Purpose: Issue tracker system for all external requests and interactions to be tracked and traced transparently;
- Central EGA adopted technology: Best Practical RT Tracker.

Operations issue tracker

- Purpose: Internal issue tracking system for operational issue tracking and resolution to the appropriate engineer and traced;
- Central EGA adopted technology: JIRA (ATLASSIAN).

Documentation and SOP system

- Purpose: A system for storage, tracking, versioning internal operational documentation such as SOPs;
- Central EGA adopted technology: Confluence (ATLASSIAN).

Compute

- Purpose: Computing power to ensure that day to day EGA operations can be supported in a timely fashion;
- Features
    - Support ETL pipeline, encryption/decryption, quality assessments, calculations on what is being archived, and a development/testing environment;

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

- Central EGA adopted solution: HPC compute farm and network storage in an isolated part of institutional network only accessible to authorised personnel. Compute resources required depends on submission throughput.

Upload staging storage

- Purpose: Sufficient space for upload of data files at all times. Requirements depend on submission sizes and throughput;
- Require regular monitoring service to track usage, notify submitters, and enforce usage limits.

Archival storage

- Purpose: Store data in a secure, indexed and sustainable fashion, yet allows for onward 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
- Example: Object-based storage system such as CleverSafe. Could also be implemented via an appropriately encrypted Posix file management system.

Data distribution service

- Purpose: To enable secure delivery of data archived at EGA to approved users
- Features
    - Re-encryption of data on a per user basis (compute requirement)
    - Servers to run the distribution services;
    - Sufficient scalable compute to scale the service according to user demand;
- Example: Aspera, secure FTP, GA4GH htsget, Globus.

Tracking databases

- Purpose: To provide a fully queryable audit trail of all data submitted to the archive, ETL operations, data access permissions, and data distribution log;
- Example: MySQL, Postgres, Oracle.