Deliverable D7.2 Implementation and documentation to create an operational EGA node

Project Title (Grant agreement no.):	ELIXIR-CONVERGE: Connect and align ELIXIR Nodes to deliver sustainable FAIR life-science data management services (871075)		
Project Acronym (EC Call):	ELIXIR-CONVERGE (H2020-INFRADEV-2018-2020)		
WP No & Title:	WP7 Federated European Genome-phenome Archives for trasnational access of COVID-19 host data		
WP leader(s):	Jordi Rambla (CRG) & Thomas Keane (EMBL-EBI)		
Deliverable Lead Beneficiary:	30 - CRG		
Contractual delivery date:	31/05/2021	**Actual delivery date:**	02/06/2021
Delayed:	No		
Partner(s) contributing to this deliverable:	EMBL-EBI(ELIXIR-EMBL), CSC(ELIXIR-FI), UU(ELIXIR-SE), BSC(ELIXIR-ES), UIO(ELIXIR-NO), FCG-IGC and INESC-ID(ELIXIR-PT), DKFZ/GHGA(ELIXIR-DE), Uni-Tübingen/GHGA(ELIXIR-DE), EMBL/GHGA(ELIXIR-DE), UP(ELIXIR-HU), SIB (ELIXIR-CH)		
Authors: Jordi Rambla (CRG), Teresa D'Altri (CRG), Frédéric Haziza (CRG) **Contributors:** - **Acknowledgments (not grant participants):** -			
Reviewers:	ELIXIR-CONVERGE Management Board (MB) members.		

Log of changes

DATE	Mvm	Who	Description
05/05/2021	0v1	Jordi Rambla, Teresa D'Altri (CRG)	Initial version
14/05/2021	0v2	Jordi Rambla, Teresa D'Altri (CRG)	Sent to PMU after incorporating internal WP feedback
17/05/2021	0v3	Nikki Coutts (ELIXIR Hub)	Circulated to the MB for final review before submission

27/05/2021	0v4	Jordi Rambla, Teresa D'Altri (CRG)	MB comments addressed
02/06/2021	1v0	Nikki Coutts (ELIXIR Hub)	Final version to be uploaded into EC Portal

Table of contents

1. Executive Summary

The scope of this deliverable is to develop a technical implementation which the federated EGA nodes need to become operative within the Federated EGA network. We have created a solution that enables sharing sensitive genetic data with associated or related metadata. The underlying problem is that sensitive genetic data has to be stored safely according to GDPR law and in some instances cannot move across borders (depending on different interpretations of GDPR), while public metadata can. The chosen solution allows files to be stored encrypted in the Local EGAs located in different countries, while public metadata is hosted at Central EGA. We have produced and made public a reference implementation and extensive documentation, to help the nodes set up their Local EGA.

2. Contribution toward project objectives

With this deliverable, the project has reached or the deliverable has contributed to the following objectives/key results:

Objective no. / Key Result no. Description	**Contributed to:**
Objective 1: Develop a sustainable and scalable operating model for transnational life-science data management support by leveraging national capabilities **(WP1, WP5**)	
Key Result 1.1: Established European expert network of data stewards that connect national data centres and similar infrastructures and drive the development of interoperable solutions following international best practice, including national interpretations of the General Data Protection Regulation (GDPR)	**No**
Key Result 1. 2: Development of joint guidelines and common toolkit that are adopted into funder recommendations, with support available nationally and in local languages	**No**
Key Result 1.3: The catalogue of successful national business models incorporated into national strategies	**No**
Key Result 1.4: The developed "sustainable and scalable operating model for transnational life-science data management support" is adopted into national ELIXIR Node	**No**
Objective 2: Strengthen Europe's data management capacity through a comprehensive training programme delivered throughout the European Research Area **(WP2, WP6)**	
Key Result 2.1: A comprehensive ELIXIR Training and Capacity building programme in Data Management, directed at both data managers and ELIXIR users, and connected to the national training programmes in Data Management in the ELIXIR Nodes and prospective ELIXIR Member countries.	**No**
Key Result 2.2: Development of a collective group of trainers that support scalable deployment of Data Management training across ELIXIR Nodes.	**No**
Key Result 2.3: A substantial cohort of data managers, Node coordinators and researchers with specific data management skills, business planning and knowledge of transnational operations across the ELIXIR Nodes	**No**
Objective 3: Align national data management standards and services through a sustainable, scalable and cost-effective data management toolkit **(WP2, WP3, WP5**)	
Key Result 3.1: Assemble a full-stack harmonised common toolkit comprising all aspects of data management: from data capture, annotation, and sharing; to	**No**

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

integration with analysis platforms and making the data publicly available according to international standards.	
Key Result 3.2: Provide exemplar toolkit configurations for prioritised demonstrators to serve as templates for future use.	**No**
Key Result 3.3: Establish national capacity in using as well as updating, extending and sustaining the toolkit across the ERA.	**No**
Key Result 3.4: Enable 'FAIR at source' practice for data generation, and analytical process pipeline implementation by flexible deployment of the toolkit in national operations	**No**
Objective 4: Align national investments to drive local impact and global influence of ELIXIR **(WP4,WP6)**	
Key Result 4.1: Development of a Node Impact Assessment Toolkit based on RI-PATHS methodology.	**No**
Key Result 4.2: Adoption of Impact assessment in ELIXIR Nodes, supported by Node coordinators network and feedback on applicability from dialogues with national funders.	**No**
Key Result 4.3: Creation of national public-private partnerships and industry outreach where open life-science data and services stimulate local bioeconomy	**No**
Key Result 4.4: Growth in reach, impact and engagement of stakeholder communication assessed by established ELIXIR Communications metrics	**No**
Key Result 4.5: Initiating and advancing discussions on Membership (EU and international) or strategic partnerships (international countries) following ELIXIR-CONVERGE workshops.	**No**
Objectives - WP7 - Federated European Genome-phenome Archives for transnational access of COVID-19 host data.	
O7.1: Federation architecture, interfaces, and compliance tests for the Federated EGA network (Task 7.1).	**Yes**
O7.2: Coordination of the development of a reference implementation sufficient to create a functional Federated EGA node (Task 7.1).	**Yes**
O7.3: Development of phenotype metadata model to enable mapping and linkage of COVID-19 host-clinical measures across European national nodes, and robust linkage between host and viral datasets (Task 7.3).	**No**
O7.4: The development of documentation and guidelines for the operational practices of federated EGA nodes (Task 7.4).	**Yes**

3. Introduction

The aim of CONVERGE WP7 is to create the Federated EGA Network (FEGA). This was raised by the need for efficient data management and secure sharing of human genomic and phenotypic data.

Genomic data is costly to obtain and extremely valuable for the scientific community and therefore should be made available to other researchers and reused as much as possible. Nonetheless, the same data is also subjected to Data protection (GDPR) regulation. In this context, EGA identified the need for the development of a federated network to enable secure sharing of data whilst enabling it to remain within the jurisdiction in which it was generated. The Federated EGA 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. 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 jurisdiction,
- **Community EGA nodes** are individual institutions or initiatives with human genetic data intended to be shared with the research community.

Deliverable scope. The scope of this deliverable is to develop the technical implements for federated EGA nodes to be interoperable in the Federated EGA network. In order to be operative, each federated EGA node will need to have the capability to accept local submissions applying international file format standards for the data and to carry out those processes that prepare data sets for being reused in subsequent research projects.

Methodology. For nodes with less mature infrastructure, we provided a packaged implementation to allow rapid deployment using the existing Local EGA technology developed by the partners. This solution enables the core functionalities for a Federated EGA node such as data submission, secure archiving, metadata storage, permissions management, and secure data distribution.

All EGA node services originate from the same codebase but due to infrastructural differences, the actual service deployments may differ. Hence, we have coordinated the development work to ensure their compatibility with the shared interfaces. We have also produced a reference implementation of these interfaces in order to enable a fast way for new nodes to start building the technical capability.

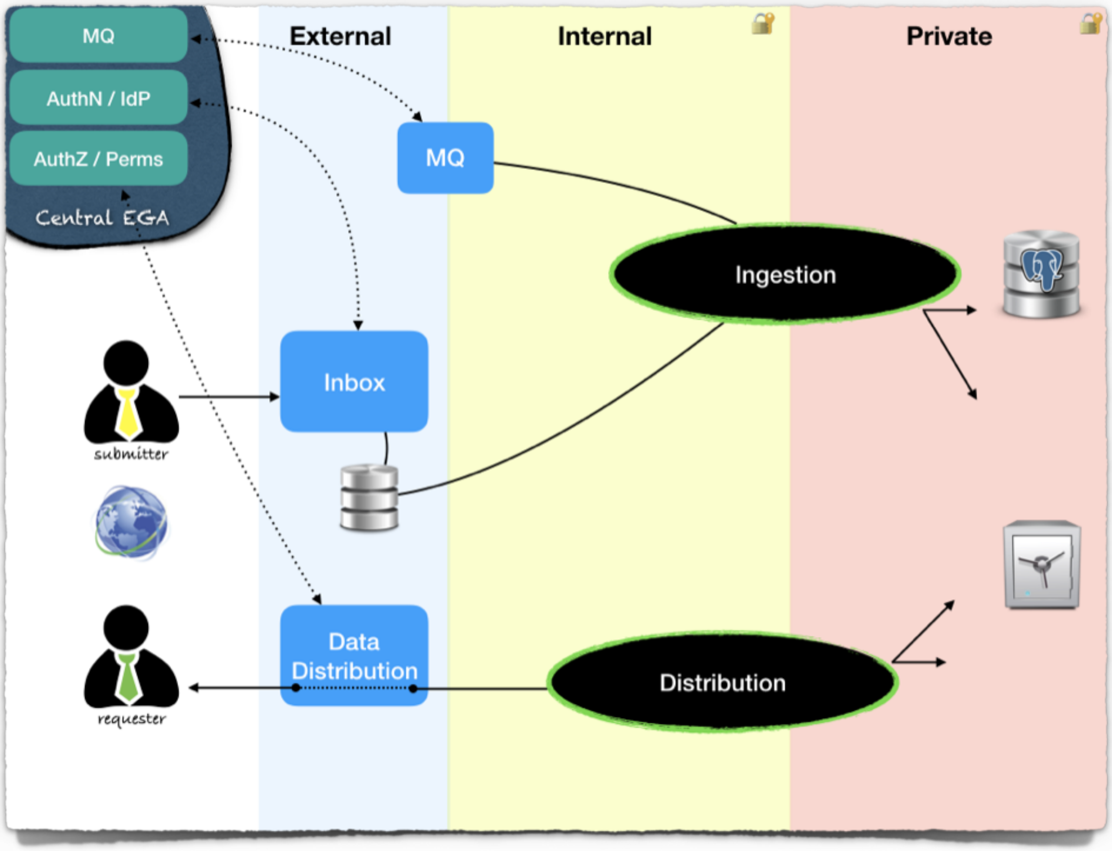
We have worked closely with the partners in several nodes (Germany, Finland, Sweden, Norway and Spain) to provide them real-time help for development and debugging. We did so through planned recurrent meetings complemented by spontaneous trouble-shooting sessions any time it was needed.

4. Description of work accomplished

4.1 The Local EGA

We developed a software which enables sharing sensitive genetic data and its associated metadata. As depicted in the image below, the distributed solution we created works so that submitters upload encrypted files into a Local EGA inbox, located in the relevant country. 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. The transport depends on the inbox and files are stored using the Crypt4GH file format. The metadata of the encrypted files and the permissions to access them are located at Central EGA.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.



4.1.1 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.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

Data distribution. If a file access has been granted by a DAC to a data requester, the file can be served in Crypt4GH format .

4.1.2 Dissemination and training material

During the development of the work described above, and as an intrinsic part of itself, we carried out several dissemination and training activities. The aim of these was to keep partners updated regarding the evolution of the software, train them about its deployment and, at the same time, gather their valuable feedback. We can highlight:

- **Seminars and presentations.** A webinar was recorded during the Mini-symposium for Federated Human Data[1] on June the 9th 2020 and the speaker, head of IT and Local EGA architect, Frédéric Haziza explains the different components involved in securely depositing and retrieving data in a remote location, while metadata is centralized at CentralEGA.
- **Periodical meetings.** We hold periodic meetings to update partners and synchronize progress.
- **Testing session with partners.** We have carried out testing sessions with our partners to provide mutual feedback regarding the development of the solution.

5. Results

The main result of this deliverable is **that we have developed the above mentioned solution with all its components and made it available to the partners**. Several partners are deploying their side of the infrastructure to be able to connect their LocalEGA to Central EGA.

Responding to the need highlighted by the work done in task 7.1, the solution has been designed to ensure that private data remain encrypted at the Local EGAs in the respective countries, while public metadata is shared with Central EGA, in order to make the data searchable across the federation.

In addition we have developed:

- Extensive documentation for the implementation of the Local EGA in the Nodes: Local EGA — Local EGA[2]. The documentation describe all needed information regarding:

[1] https://www.youtube.com/watch?v=k9R8W3V3ugU

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

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

  - 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
- A reference implementation[3]: This repository contains the necessary code and instructions to set up a LocalEGA. It also contains CentralEGA stubs in order to help interested developers to set up a locally-deployed LocalEGA and have it run against CentralEGA (be it the fake instance or not).

6. Conclusions

We conclude that a highly important work has been done, which will contribute to the set up of the Federated EGA network.

7. Impact

The described work has greatly impacted the feasibility of Federated EGA network creation. Thanks to the results highlighted above, the nodes have been provided with the technical tools for becoming operational and establishing the expected services. The impact of this work can therefore be seen as a crucial piece in the broader picture of the invaluable power that the creation of the Federated EGA network will have over the European Research Community.

8. Next Steps

In the near future we will continue to work in collaboration with the nodes to establish their Local EGAs connection with Central EGA. This work is directly supported and supervised by the Federated EGA Operations committee, where members of all interested nodes and Central EGA meet fortnightly to discuss aspects related to the technical implementation of the Federated nodes.

In addition, we are working in the frame of the ELIXIR Federated Human data community to bring the efforts ahead.

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

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.

This work will greatly contribute and complement the Maturity Model developed in the context of Task 7.4, to aid nodes progressively moving through the technical implementation requirements to become operational.

9. Deviation from Description of Action

Not applicable.

This project has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No **871075**.