EXCELERATE Deliverable D 9.5

Project Title:	ELIXIR-EXCELERATE: Fast-track ELIXIR implementation and drive early user exploitation across the life sciences	
Project Acronym:	ELIXIR-EXCELERATE	
Grant agreement no.:	676559	
	H2020-INFRADEV-2014-2015/H2020-INFRADEV-1-2015-1	
Deliverable title:	Report on implementation of ELSI and policy consideration for controlled access data	
WP No.	9	
Lead Beneficiary:	20 - CSC	
WP Title	Use Case D: ELIXIR framework for secure archiving, dissemination and analysis of human access-controlled data	
Contractual delivery date:	31 August 2018	
Actual delivery date:	31 August 2018	
WP leader:	Thomas Keane Jordi Rambla	1 - EMBL-EBI 8 - CRG
Partner(s) contributing to this deliverable:	EMBL-EBI, CRG, CSC, UIO, UU, NBIC, EMBL (ELIXIR Hub)	

Authors and Contributors:

Dylan Spalding (EMBL-EBI), Jordi Rambla (CRG, Spain), Juha Törnroos (CSC, Finland), Jaakko Leinonen (CSC, Finland), Mikael Linden (CSC, Finland), Teemu Kataja (CSC, Finland), Minna Ahokas (CSC, Finland), Alexander Senf (EMBL-EBI), Heikki Lehväslaiho (CSC, Finland), Tommi Nyrönen (CSC, Finland), Niclas Jareborg (UU, Sweden), Thomas Keane (EMBL-EBI) and Ilkka Lappalainen (CSC, Finland)

Table of contents

1. Executive Summary

The EXCELERATE WP9 focuses on the human data use case. It provides technical extension to the permanent archive for all types of controlled access data, the European Genome-phenome Archive (EGA), that will support genomic information deposition, data management and sharing within local jurisdiction.

This report focuses on describing the improvement on the EGA processes that support the use of controlled data access protocol and compliance to the ELIXIR ELSI Policy and the European General Data Protection Regulation (GDPR). It also provides terminology translation between EXCELERATE WP9 technical reports, the ELSI terminology and the GDPR.

Currently the EGA service is provided by two European organisations based on a bilateral agreement. The technical solutions delivered during the EXCELERATE WP9 will require a legal framework to enable federated EGA services.

2. Impact

2.1 Webinars:

- Tommi Nyrönen, Alexander Senf, Ilkka Lappalainen, Jaakko Leinonen, Juha Törnroos, Oscar Martinez Llobet, Angel Carreno Torres (2017): The transfer of large volumes of electronic, confidential, human data. *ELIXIR Webinar.* Available at https://www.elixir-europe.org/events/elixir-webinar-transfer-large-volume-data
- Mikael Linden, Juha Törnroos (2018): Access to Sensitive Human Data with the ELIXIR AAI. *ELIXIR Webinar.* Available at https://www.elixir-europe.org/events/webinar-access-to-sensitive-data-aai
- Juha Törnroos, Frederic Haziza, Niclas Jareborg (2018): Local EGA - Core functionalities demonstration. *Tryggve development team review meeting.* Available at https://www.youtube.com/watch?v=d8tIDZvDGKQ

2.2 Posters

- Mikael Linden, Tommi Jalkanen, Juha Törnroos: Sensitive dataset access management. Poster, ELIXIR All Hands 2018, Berlin. Available at https://f1000research.com/posters/7-754

2.3 Presentations

- Presentation at ELIXIR Beacon F2F meeting, Barcelona March 2018: Technical team update, roadmap and Beacon Network presentation. Available at https://drive.google.com/drive/u/0/folders/1CJS1aPf8cVEfSdygH-OdiOmAE-jG_g55
- Presentation at ELIXIR All Hands Beacon workshop, Berlin 2018: Adding AAI functionality to your Beacon implementation.
- Presentation at ELIXIR All Hands 2018, Berlin 2018: Local EGA

2.4 Documentation and publications

- Juha Törnroos, Mikael Linden: Architecture for delivering dataset permissions on ELIXIR AAI. Available at https://docs.google.com/document/d/1rqCD75HRA99HKwq0s-OkWWBKiaEo2JO-_MYjlsyiSQ4/edit?usp=sharing
- Stephanie O. M. Dyke, Mikael Linden, Ilkka Lappalainen, Jordi Rambla De Argila, Knox Carey, David Lloyd, J. Dylan Spalding, Moran N. Cabili, Giselle Kerry, Julia Foreman, Tim Cutts, Mahsa Shabani, Laura L. Rodriguez, Maximilian Haeussler,

Brian Walsh, Xiaoqian Jiang, Shuang Wang, Daniel Perrett, Tiffany Boughtwood, Andreas Matern, Anthony J. Brookes, Miro Cupak, Marc Fiume, Ravi Pandya, Ilia Tulchinsky, Serena Scollen, Juha Törnroos, Samir Das, Alan C. Evans, Bradley A. Malin, Stephan Beck, Steven E. Brenner, Tommi Nyrönen, Niklas Blomberg, Helen V. Firth, Matthew Hurles, Anthony A. Philippakis, Gunnar Rätsch, Michael Brudno, Kym M. Boycott, Heidi L. Rehm, Michael Baudis, Stephen T. Sherry, Kazuto Kato, Bartha M. Knoppers, Dixie Baker & Paul Flicek (2018): Registered access: authorizing data access. *European Journal of Human Genetics*. Available at https://www.nature.com/articles/s41431-018-0219-y

3. Project objectives

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

No.	Objective	Yes	No
1	To upgrade and make more portable -omics data collection and submission tools utilizing the European Genome-phenome Archive (EGA) as the core of an ELIXIR community secure data sharing network for - omics data.		X
2	To enable value-added services at project specific, regional, or national resources by establishing ELIXIR-wide community facing tools that allow local resource owners and developers to add value to their systems through data and metadata services from the EGA.	X	
3	To extend and generalise the system of access authorization management and high volume secure data transfer developed in the EGA project to address the secure data access needs across ELIXIR resources and open new modes of secure data access such as through public and private clouds.	X	

4. Delivery and schedule

The delivery is delayed: Yes ☑ No

5. Adjustments made

None

6. Background information

Background information on this WP as originally indicated in the description of action (DoA) is included here for reference.

Work package number	9	Start date or starting event:	month 1
Work package title	**Use Case D: ELIXIR framework for secure archiving, dissemination and analysis of human access-controlled data**		
Lead	1 - EMBL-EBI		
Participant number and person months per participant **1 - EMBL (34 PM)**, 5 - UTARTU (16 PM), 6 - NBIC (0 PM), UMCG (LTP to NBIC) (5 PM), 8 - CRG (22.3 PM), 14 - UPF (23.5 PM), 20 - CSC (24 PM), THL (LTP to CSC) (3 PM), 24 - UiO (6 PM), 45 - UU (2 PM), SU (LTP to UU) (4 PM).			
Objectives This Work Package has three main objectives: To upgrade and make more portable -omics data collection and submission tools utilizing the European Genome-phenome Archive (EGA) as the core of an ELIXIR community secure data sharing network for - omics data. Tools developed here will support submission of all types of -omics data from human samples consented for biomedical research from disease consortia such as International Cancer Genome Consortium (ICGC), Rare Diseases (Rd-Connect), national cohorts, and biobanks. Emphasis is given to supporting investigator and locally driven research projects with human data consented for biomedical research. To enable these projects, the data submission tool chain will be made more portable and user-friendly with the goal of distributing a common toolset "in-a-box" to enable local and national groups to collect -omics data and meta data in a distributed manner which is consistent across European groups through ELIXIR coordination. To enable value-added services at project specific, regional, or national resources by establishing ELIXIR-wide community facing tools that allow local resource owners			

and developers to add value to their systems through data and metadata services from the EGA. For example, local research projects would be enabled to make their data discoverable and searchable, and linked with available -omics data from various sources, by leveraging stable unique EGA identifiers. Further, locally developed project specific data portals will be enabled through defined standard APIs using real time secure data links which allow -omics big data archived in the EGA to be presented in combination with biobanks or cohort data.

To extend and generalise the system of access authorization management and high volume secure data transfer developed in the EGA project to address the secure data access needs across ELIXIR resources and open new modes of secure data access such as through public and private clouds. For example, a trusted ELIXIR Cloud service can receive local copies of selected datasets through a secure data mirroring system and provide access to data and compute to those users that already have data access permissions available from appropriate Data Access Committees stored in the EGA system. The WP will partner first with 2-4 large resource owners to gain the required expertise, document the process in multiple ELIXIR member states and finally to propose a way to scale up these services to match wider European requirements. This WP will also be used to drive creation of the ELSI framework that supports the workflow (WP12).

Work Package Leads: Thomas Keane, EMBL-EBI (since 1/3/2017); Jordi Rambla, CRG EGA (since 1/3/2017); Justin Paschall, EBI (up to 1/3/2017); Arcadi Navarro, ES (up to 1/3/2017)

Description of work and role of partners

WP9 - Use Case D: ELIXIR framework for secure archiving, dissemination and analysis of human access-controlled data [Months: 1-48]

EMBL, UTARTU, NBIC, CRG, UPF, CSC, UiO, UU

This WP delivers the core ELIXIR workflow for long term archive and re-use of human data consented for biomedical research requiring access-control based on a data access agreement and approval process. The workflow supports data submitters and ELIXIR Node coordination on data deposition into the EGA archive in a manner that will maintain data ownership in the hands of the original research data owner, enable data release to authorised individual users from the archive and to partner with downstream secure ELIXIR data analysis platforms. This workflow ad supporting infrastructure will allow the data owners to focus on their unique areas of data generation and analysis expertise while being able to rely on EGA and the ELIXIR infrastructure for their common –omics big data storage, coordination and distribution needs under appropriate legal frameworks. The work described here will leverage the work of other ELIXIR-EXCELERATE Work Packages, for example WP10 to scale each service structure to cover all ELIXIR Nodes and with WP4 for technical service support, and relies on WP12 to establish the necessary legal framework that supports workflows.

The Workflow can be summarized as:

1. Data preparation, validation, and submission to the EGA making use of common supporting tools and data models (e.g. through Node data Network, WP10). Focus on providing software tools and remote APIs enabling local leadership and customisation within context of specific projects, supported by common ELIXIR coordinated tools and data models.

2. Bidirectional linking and secure data streaming between -omics data archived in EGA and local repositories or data portals that hold further information about the project and samples.

3. Management of user access-rights for release of archived data to authorized researchers under Data Access

Agreements using ELIXIR tools, such as the REMS (WP4), that allow resource owners to manage data access rights.

4. Expanded access through ELIXIR partner secure clouds that can host EGA datasets, requiring the provision of metadata and authorization APIs

5. Data synchronization between the main EGA archive and authorized project specific resources and access points, such as compute clouds.

Task 9.1: Enhanced secure data submission tools. (49PM)

This task will update the existing EGA submission tools and documentation to facilitate large-scale data submissions operations, emphasizing local leadership and customization within a common framework.

Partners: ES, EMBL-EBI, FI

Subtask 9.1.1: Support for large scale submission of -omics data and sample metadata to the EGA. (30PM)

Support for large-scale submission of -omics data and sample metadata to the EGA through improved online tools, automated verification, and tools for the application of standard vocabularies to phenotype collection. These tools will make use of table “spread-sheet” based views of data for submitters less comfortable with technologies such as XML. Further tools and reports supporting global EGA stable identifier mappings will allow easier integration with local identifiers, in support of federated global tracking of submitted samples and their derived -omics data.

Subtask 9.1.2: Portable submission toolkit. (21PM)

This task is composed of data format definitions and software components, a “mini-EGA in-a-box” will allow increased local control and coordination of data collection, and allow early validation of standardized data and metadata formats.

This implementation provides the practical means for distributed projects to collect access-controlled human biomedical data in a manner that maintains a coordinated data model and dataset registry, enabling federated and a centralized single-point of discovery and access.

Task 9.2: Integrating centralized and distributed projects through transparent access to secure data: enabling local projects within a European wide framework. (40PM)

This task will enable local projects, such as study-specific data portals, local cohort resources, and national bioinformatics hubs by providing developer level APIs and services such that local efforts can efficiently build customized project branded solutions which make use of underlying ELIXIR and EGA tools and data archives.

Partners: ES, EMBL-EBI, FI, EE, NL, NO, SE

Subtask 9.2.1: Support secure integration of EGA data to downstream project client websites. (10PM)

Support secure integration of EGA data and metadata to downstream project client websites by providing new EGA programmatic interfaces that support standardized REST calls and provides results in ELIXIR endorsed formats (WP3 and WP6).

Subtask 9.2.2: Access management workflow support. (10PM)

Support access management workflows by data access committees through ELIXIR for EGA and other projects through developing applications of the Resource Entitlement Management Systems (REMS) expanding on an existing pilot project. This effort is focused on providing tools to delegate management to local projects and ELIXIR Nodes through new administrative roles.

Subtask 9.2.3: ELIXIR and EGA access integration. (20PM)

Specific efforts supporting controlled access-omics data infrastructure for use of partner national cohort studies in terms of submission, permissions management, and local and customized presentation of data under the cohort branding. Services will be tailored to respect the unique policy and data protection requirements of national cohorts, allowing single point of request and download from cohort branded web-pages. Support will be provided for distributed local hosting of datasets, within a common ELIXIR framework, where restrictions exist on the movement or hosting of data based on national borders.

Task 9.3: Federated authentication, large scale data management, and secure clouds in practice. (38PM)

This task is closely linked to the technically focused WP4 that provide the technical solutions required to deliver the outcomes of Task 9.1 and 9.2. In this task, technical components, including high volume secure data transfer and authentication and authorization management, are brought together to make -omics data from EGA and phenotypic data from cohort studies available for secure download, remote API access or from within public or private Cloud-based secure analysis environments. Cloud-based access to the EGA ecosystem provides a new access mode meeting a significant user need from research groups with limited local resources for compute and large-scale reference data storage.

Partners: EMBL-EBI, ES, FI, EE

Subtask 9.3.1: Large scale data mirroring support. (12PM)

Support for automated large scale data mirroring from the EGA archive to the authorized ELIXIR partner local services and cloud compute or HPC providers. This process instantiates concrete data flows based on data transfer technologies in WP4 to track domain specific files, versions of files, confirms transfer success, and tracks files available in different locations. Generic interfaces should provide transparent access to multiple underlying transfer and storage modules (e.g. gridFTP/irods/object store etc.)

Subtask 9.3.2: EGA data access authorization integration. (12PM)

Integrate EGA data access authorizations to local project data portals and Cloud access providers. This is a new service that allows authorized third-party services to programmatically check compliance with the current user data access authorizations from the ELIXIR coordinated repositories such as the EGA database each time user accesses a file in the cloud or other remote service. A first planned project using EGA data within the private, secure, cloud at CSC in Finland will provide our reference implementation.

Subtask 9.3.3: Data access APIs. (14PM)

Develop and implement standard data access APIs to be to used for inter and intra cloud communication and for secure remote REST API access in coordination with the Global Alliance for Genomics and Health (GA4GH).

For tasks 1-3 we expect to list a number of updates to the submission tools while we work with the first 2-4 chosen resources. These updates will be prioritised in the scope of this WP. WP4 will provide AAI support for WP 9, and vice versa WP9 will work with WP4 to set the requirements for ELIXIR AAI services. WP9 needs to information on service component availability and this information is expected to be available from technical services registry such as cloud resource allocation, valid EGA data access authorizations, and file mirroring status if data are not yet ready to be used in the cloud. WP12 will Create a set of Legal Frameworks for ELIXIR-related operations that will be integrated within WP9 with the technical solutions devised for particular EGA needs.

Appendix 1: Implementation of ELSI and policy consideration for controlled access data

A1.1. Introduction

The ELIXIR EXCELERATE WP9 promotes secure and safe data management processes for human Genetic Data compliant with the Open Science policy and the FAIR (Findable, Accessible, Interoperable and Reusable) principles. The data types collected from the study participants are typically consented for research but not for fully open and public distribution. As a consequence, Data User will be able to access these data only using a controlled access protocol. This process validates the identity of each Data User and restricts data access to only authorized Data Users based on the original Informed Consent constraints.

The ELIXIR core data resource for controlled access data is called the European Genome-phenome Archive[1] (EGA). The EGA follows strict protocols for information management, data storage, security and dissemination. The EXCELERATE WP9 expands existing core services with four distinct actions to support a larger distributed network in Europe:

1. facilitate national data deposition to the local federated EGA Nodes;
2. enable remote access or caching of data sets from the core EGA to the Nodes to support local data analysis in a secure manner;
3. integrate the core EGA Data User identification and access permissions as part of the ELIXIR Authentication and Authorization Infrastructure (AAI) services; and
4. develop shared EGA metadata and security standards in collaboration with European and global initiatives such as the Global Alliance for Genomics and Health to support national use cases.

The core technical components and the process requirements for sensitive data management were described in Deliverable 9.1 and their implementation in Deliverable 9.4. This report links both of these reports to the ELIXIR Ethical, Legal and Societal Issues (ELSI) Policy[2,3]. The ELIXIR Board has approved the ELSI Policy and it is applicable to all ELIXIR Services as part of the Node Collaboration Agreements, Service Delivery Plans or those provided as part of the Commissioned Services contracts. The ELIXIR ELSI Policy is compliant with the national laws and relevant international regulations such as Universal Declaration of Human Rights, Charter of Fundamental Rights of the European Union and the recent EU General Data Protection Regulation (GDPR[4]). The Policy implementation within the Nodes is mandated as part of the Article 11 of the ELIXIR Consortium Agreement[5].

[1] https://ega-archive.org/
[2] https://www.elixir-europe.org/news/re-using-life-science-data-ethically
[3] https://drive.google.com/file/d/0BxqILhwJcm1qME00QWRKUmtEVXM/view
[4] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&from=EN
[5] https://drive.google.com/file/d/0B7btK9HAXhx1dEVIMmRsaEpDSzA/view

This document also aims to clarify the responsibilities of sensitive data management between Data Provider, Service Provider and Data User as defined in ELIXIR ELSI Policy document section “Definition of Terms”, and the GDPR. The ELSI document terminology is used throughout in this report in preference over the technical terminology (see Table 1 that also provides mapping to GDPR terms).

Table 1. Correspondence of ELSI, GDPR and technical terminology for EGA services.

ELIXR ELSI	**EU GDPR**	**Technical description used for EGA**
Service Provider	Processor	EGA
Data Provider	Controller	Primary institution(s) or project that generated the data and provided leadership for the study
Genetic Data	Genetic Data	Genomic data generated from the study participants
Informed Consent	Consent	Legal document signed by each study participant authorizing data for research but not fully public dissemination.
Data Subject	Data Subject	Study participants whose genomic data are accessed and analysed
Data Access Committee (DAC)	Controller (through Data Provider)	A Committee established by the Data Provider to make data access decisions on the basis of scientific and ethical criteria and and in accordance with the original informed consent agreements
Data User	Controller	Researcher as part of an institution applying data access for secondary research uses. In the context of data use research organisation is responsible for legal obligations.
Data Transfer Agreement (DTA)	Data Processing Agreement (DPA)	Agreement between EGA and the projects or institution(s) that generated the data (Data Provider)
Data Transfer Agreement (DTA)		Data Access Agreement (DAA) between the Data Access Committee and authorized researcher with an institutional affiliation
Anonymous Data	*GDPR does not apply to anonymous data*	Not traceable to an identifiable natural person

Pseudonymised Data	Pseudonymous data	Data traceable to an individual only through separately kept information
Personal Data	Personal Data	Any information relating to an identified or identifiable natural person
Processing	Processing	Any operation or set of operations performed on personal data

General requirements for processing sensitive human data

The ELIXIR develops and supports interoperating data management services including data transfer, storage, compute and user authentication and authorization. These resources are made available to Data Users through Service Providers such as the EMBL-EBI Embassy or the CSC ePouta Cloud as described in the Terms of Use. Use of these resources requires biomedical research projects to comply with the Informed Consent, include an ethics approval and follow relevant laws and regulations. For example, the EU GDPR defines the roles and responsibilities for Service Provider, Data Provider and Data User (Figure 1).

> *Anonymisation should be achieved using state of the art techniques that are used in practice. It is sufficient that the data is* de facto *anonymised, i.e. individuals cannot be re-identified by the use of reasonable means. When assessing re-identification risks all factors have to be considered, including for example context knowledge that is available to the data users, access control systems that ensure data is only used for biomedical Research purposes in order to to reduce available context knowledge, sensitivity of the data etc.* (ELIXIR ELSI Policy)

The main issue for data management arises from the requirement to support secure data processing environment. Biomedical research projects collect Genomic Data from the Data Subjects that cannot be rendered Anonymous without compromising the scientific value for research. For example, a whole genome sequence explicitly identifies a natural person given it can be linked to a piece of DNA derived from the same person or their relative. Furthermore, aggregated data products describing for example genomic allele balance across the genome for a Biobank cohort consisting of thousands of participants may be used to re-identify the Data Subject using statistical methods with prior knowledge of genetic information from the person[6,7,8].

[6] 10.1016/j.ajhg.2015.09.010

[7] 10.1371/journal.pgen.1000167

For the purpose of this report, sensitive Genetic Data collected from the Data Subject is defined as Pseudonymised Data. The data types are linked to the Data Subject by an abstract identifier. Any Personal Data attached to Genetic Data (such as name, address, phone number or national health care provider identifier) are removed before deposition into ELIXIR services. Hence, the Service Provider cannot map deposited data back to the Data Subject (this information may be kept by the Data Provider). The EXCELERATE WP9 infrastructure and processes developed for managing Pseudonymised Data are based on identifying, assessing and managing the risk for re-identification, data leak or security breach appropriately. The standards and best practices for protecting ELIXIR services, including the identity management, data security, privacy protection and service assurance, have been worked out jointly with the Global Alliance for Genomics and Health (GA4GH) Data Security work stream.

A1.2. The federated EGA data access policy

> *At the point of data submission ("data in") and where applicable, any such agreement must reflect that the Data Provider is submitting the data to the Service only after having secured the necessary Informed Consent by the Data Subject and/or taking into account any other limitations, such as for example deriving from ethics reviews or relevant law, or the requirement to feedback Incidental Findings. It is advisable that the DTA also include a provision for the Data Provider to inform the Service Provider should the need to remove Personal Data arise for example when Consent is withdrawn.* (ELIXIR ELSI Policy)

The EGA accepts only data sets that support the general data management requirements described above. The Data Provider must also authorize a named person(s) to upload the data files and metadata to the EGA on behalf of the project or organisation that approved and monitored the original study. These requirements constitute a formal agreement between the Data Provider and Service Provider (EGA) that are recognised as Data Processing Agreement (GDPR terminology) or Data Transfer Agreement (ELSI terminology). As a consequence this agreement also defines Data Provider as a GDPR Controller and EGA as a Processor providing the necessary secure data management processes (Figure 1).

[8] 10.1093/jamia/ocw167

The Data Provider delegates the responsibility for overseeing the control access authorization process to a Data Access Committee (DAC) and not to the Service Provider that simply acts as a secure data broker (Figure 1). The ownership of the data is not transferred to the DAC as part of this process. The EGA retains itself the right to refuse to serve any DAC where data access is denied selectively based on scientific competition or for reasons that are not based on the original Informed Consent.

The secondary use of data is defined as part of the legal contract (Data Access Agreement, DAA) signed between the DAC and the authorized Data User[9]. Currently data are downloaded from the EGA to a preferred institutional or personal compute environment. In the future, Federated EGA will provide a secure processing environment for the data analysis. In GDPR terms, access to the data will require a DPA between the Service Provider (Processor) and Data User (Controller) as described in Figure 1.

In compliance with the new EU GDPR and the ELIXIR ELSI policy, the EGA service is able to remove all deposited data from a Data Subject. This requires a formal notification from the Data Provider that authorizes the Service Provider to contact all authorized Data Users and instruct them to remove all data collected from the Data Subject from any further analyses. Data Providers are required to confirm that it has happened. In practice, it might not be possible to trace the data through personal computers and institutional infrastructure. This problem is avoided with the authorized data analysis environments as described in ELIXIR EXCELERATE D9.4 report.

A1.3. Preparing for Federated EGA services

> *Controlled access to Personal Data should be implemented unless the available consent or other considerations allow fully unrestricted open access. The Service Provider is responsible for ensuring levels of data security appropriate for the type of data held.* (ELIXIR ELSI Policy)

The EGA service is currently provided by two European institutes, the EMBL-EBI and the Centre for Genomic Regulation, Barcelona, Spain. The EXCELERATE WP9 provides the technical solutions to support use of data deposited in the EGA service within the ELIXIR Nodes compliant with the federated EGA data access policy. Further, it includes technology transfer from the EGA to the participating ELIXIR Nodes in order to facilitate

[9] 10.1038/ng.3312

national data depositions compliant on shared data model and FAIR principles. These actions prepare for a larger distributed EGA network of data archiving, analysis environment and dissemination services. Hence, the Service Provider in this context should include the current institutions providing the EGA service and the new national Nodes that jointly will form the Federated EGA service. The plans will include Data User authentication and authorization information integrated as part of the ELIXIR AAI services for easier user experience (single username and password for all services), for increased security and integrity of the processes across the involved institutes (such as data access permissions) and to provide a full audit trail when needed.

The Resource Entitlement Management System (REMS) is part of the ELIXIR AAI services[10]. The EXCELERATE WP9 coordinated integration of REMS as part of the EGA services to support the DAC processes for accepting data access applications, reviewing these applications and approving authorized use of the governed data sets. The new electronic workflows based on REMS improve DAC authorization processes by minimizing human error through standardization and automation and providing provenance model across all DACs. The EGA provides access to REMS only for those DAC persons that have been explicitly named by the Data Provider. DAC members log into the REMS service using two-factor authentication process. Identification of the acting DAC member and time stamped logging of all actions provides a full audit trail. Changes on Data User access rights are encrypted and submitted from the REMS to the EGA service using a secure connection.

The EGA service maintains the master copy of all data access rights for DAC authorized Data Users. Using the ELIXIR Authentication and Authorization Infrastructure (ELIXIR AAI) this information is federated as digitally signed snippets of data (called access tokens) allowing Data User to access the current EGA service or the future sensitive data services including the federated EGA Nodes. For example, the EXCELERATE WP9 made possible remote access from the EGA service to the Finnish ELIXIR Node secure compute environment using the new GA4GH htsget transfer protocol[11] or accessing locally stored genomic data through a Filesystem in User SpacE (FUSE). The Data User may only use the remote or local files with valid access token retrieved from the EGA master database using secure communication between the EGA and ELIXIR Node. Importantly, this process enforces locally the DAC authorized data access permission for each Data User accessing potentially a shared computing environment (for example a

[10] https://f1000research.com/articles/7-1199/v1
[11] https://doi.org/10.1093/bioinformatics/bty492

virtual machine allocated for project members). The original data itself is permanently stored in an encrypted format on the Service Provider's managed secure data storage system. The access tokens have a limited lifetime supporting the ELIXIR ELSI Policy requirement on authorized and controlled access to the data.

The enhancements presented in the EXCELERATE WP9 to EGA further expand its compliance to FAIR principles[12]. The ELIXIR Recommended Interoperability Resource (RIR)[13] together with Fair Metrics[14] state these criteria in more practical terms: The EGA is a widely used (Community), established ELIXIR Core Data Resource (Governance, Funding & sustainability plan) for all human genotype and phenotype data (Non-redundancy of the resource, Comprehensiveness of the resource) that WP9 extends to a federated sharing and computational service (Inter-organisational collaboration of the resource, Infrastructure for integratable data collections). While genotype-phenotype data by its nature can be shared through control-access mechanism, the central EGA continues to share open and fully public metadata adhering to community standards to maximise data discoverability (Metadata exposure). The federated approach will allow EGA to better respond to national legislation and practices (Legal framework, funding, and governance) that offers better management of data use in dedicated computing platforms combined with tight user identification of users through development of global metadata and security standards (Quality of resource).

[12] http://FAIRsharing.org
[13] http://www.elixir-europe.org/platforms/interoperability/rir-selection
[14] http://fairmetrics.org/

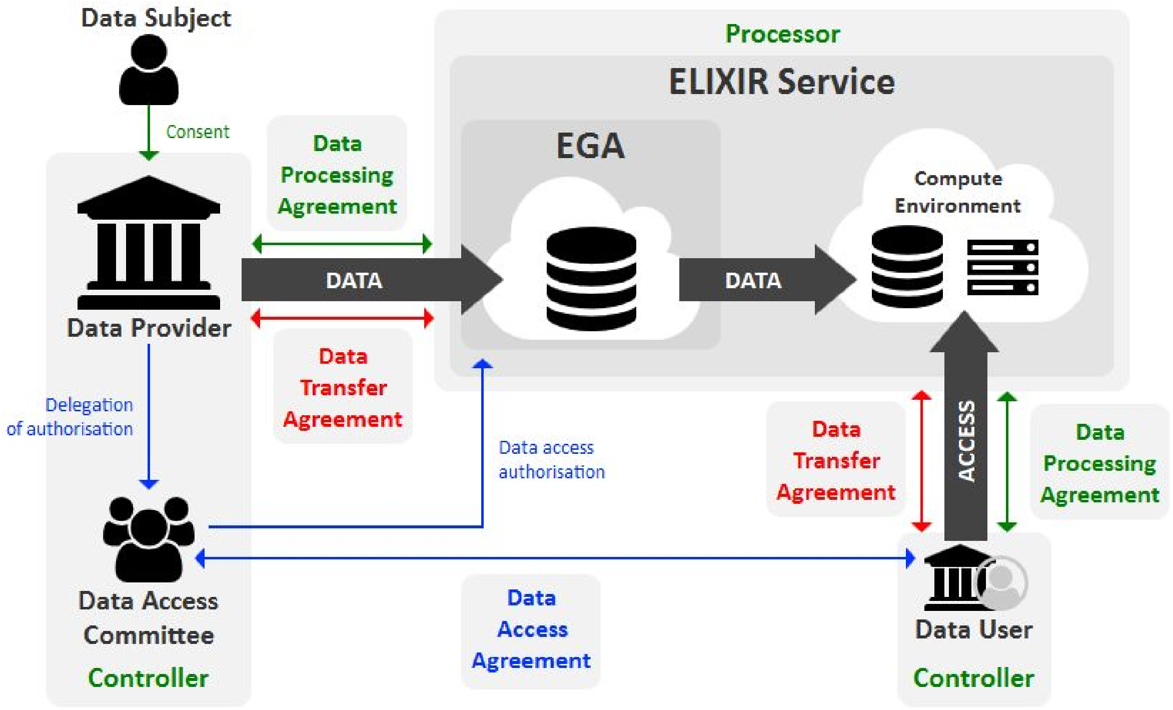

Figure 1. The EGA federated access policy (marked in blue) demonstrating ELIXIR ELSI (red) and GDPR (green) compliance. Data Provider may only deposit Pseudonymised data to Service Provider (EGA). The data access granting responsibility is delegated from Data Provider to Data Access Committee (DAC). Data User may access DAC approved data sets from the Service Provider secure computing environment using ELIXIR AAI services as described in Delivery report 9.4. The DTA and DPA between Data User and ELIXIR Service Provider typically takes the form of general terms of service or terms of use.

A1.4. Third Party-Managed IT resources

> *Increasingly, IT services are moving from local servers and hardware managed by the organisation's own staff to systems owned and managed by third party providers. Transfer and storage of controlled access data on third party systems require additional considerations.* (ELIXIR ELSI Policy)
>
> *... Thus, it is advisable that institutions validate that they are partnering with a reputable third party provider and develop appropriate security plans and service agreements before data is migrated to third party providers. Migration of person-related data to servers located in other countries also require consideration of relevant data protection regulation (e.g. in the case of non-EU cloud services, it should be ensured that it is legally allowed, for example because the foreign country has been declared a "safe" country by the EU*

> *Commission or the Consent of the Data Subject explicitly allows the transfer).* (ELIXIR ELSI Policy)

Data Provider always retains complete ownership of the data through the federated data access policy. The data deposited to the EGA service can be stored already in the two sites - EMBL-EBI and CRG, Barcelona based on a bilateral agreement between these institutes. The establishment of the new Federated EGA service will require formalization of the 3rd party status within the consortium to clarify what type of legal agreements will be required for the multinational federation in order it to be classified as GDPR Data Processor for the Data Providers and for the Data User (as shown in Figure 1). Further, the GDPR and national laws will dictate how data will flow within the European Union and especially between the Federated EGA Nodes. The Informed Consent, or implementation through National legislation, will be needed to explicitly authorize data transfer to a non-EU country or integration as part of services based outside of the European Union or data analysis using third party managed IT service platforms (such as use of international private sector clouds) for data analysis.

A1.5. Conclusions and Future

A number of European countries have an active genome strategy that aims to use advanced genomics for national health care. As a result, it is expected that a large number of human genomes and other sensitive biomolecular data will be collected from citizens and made available for secondary research[15]. The rapid increase in data volume and complexity creates risks for interoperability on controlled access data for international research that ELIXIR EXCELERATE WP9 has been trying to address.

The EXCELERATE project has increased the ability of the participating ELIXIR Nodes to manage sensitive data collected from biomedical research projects. The technical human data management solutions demonstrate how large scale data depositions, transfers or remote access to genomic data can be used to support the local researchers while complying to the federated EGA data access policy, European legislation, and the secure shared data model. To realise the full potential of these technical solutions, they need to be supported by comprehensive process documentation and user training. The EGA federation also requires clear definition of roles and responsibilities for contractual service delivery. The ELIXIR ELSI Policy, national laws and the EU legislation form the necessary legal basis for these operations across the Europe. What is clearly required is

[15] http://ec.europa.eu/newsroom/dae/document.cfm?doc_id=50964

understanding how the EGA service may collaborate with services outside of the EU or potentially include Nodes that are not governed by the EU GDPR.

This report has analysed how the EGA controlled access process complies with the ELIXIR ELSI Policy. The GA4GH has recently described a novel method for data dissemination - the registered data access model[16]. This model is intended for data types where security risk analysis on the type does not require DAC access approval process for each data set, for example the risk of providing aggregated data products (derived from individual level Genomic Data) for *bona fide* researchers that accept terms and conditions preventing re-identification of Data Subjects within the data set. The ELIXIR AAI supports already registered access model by providing a mechanism for the Data User to demonstrate their *bona fide* status and facilitates the necessary attestations to confirm that they are committed to the terms of registered access data. As a result, the users' ELIXIR identity is amended with an extra attribute declaring that they have a status as a bona fide researcher. The model is currently tested as part of the ELIXIR Beacons - project[17].

Another potential source of complexity and confusion lies in defining, describing and understanding the exact permissions granted by the Data Access Committee to the Data User. Work has already begun to standardise these data use conditions that will be codified as Consent Codes[18] and DUO ontology[19] within the GA4GH Data Use and Research Identifier (DURI) work stream. The future work should ensure that the Consent Codes are fully integrated as part of the ELIXIR AAI facilitating faster access to EGA deposited data sets from secure Cloud environments.

Interpreting the genomic data for research and for the benefit of health services requires collaboration at the global level that is not possible without understanding how the national laws and the EU GDPR impact on data sharing with partners outside of the European Union. The work within GA4GH work streams is important to ensure interoperability and development of the necessary standards for the data and processes. The EXCELERATE WP9 work will continue in future H2020 funded projects such as the EOSC-Life and CINECA. The ELIXIR coordination efforts are included into the proposed Federated Human Data implementation study for 2019-21.

[16] https://doi.org/10.1038/s41431-018-0219-y
[17] https://f1000research.com/posters/6-425
[18] https://doi.org/10.1371/journal.pgen.1005772
[19] https://docs.google.com/document/d/1970NY0lvKw2Geu0lYz75FfeigdT-pwmSyLFs8zNyXpg/edit#