Beyond One Million Genomes

D4.1

Secure cross-border data access roadmap — 1v1r1

Project Title (grant agreement No)	Beyond One Million Genomes (B1MG) Grant Agreement 951724
Project Acronym	B1MG
WP No & Title	WP4 - Federated Secure Cross-border Technical Infrastructure
WP Leaders	Tommi Nyrönen (CSC), Ilkka Lappalainen (CSC), Bengt Persson (UU), Sergi Beltran (CNAG-CRG)
Deliverable Lead Beneficiary	4 - CSC
Deliverable	D4.1 - Secure cross-border data access roadmap - 1v0r1
Contractual delivery date	31/05/2021	**Actual delivery date**	23/11/2021
Delayed	No
Authors	Dylan Spalding (CSC), Tommi Nyrönen (CSC), Heikki Lehväslaiho (CSC), Riku Riski (CSC), Ilkka Lappalainen (CSC), Bengt Persson (UU), Sergi Beltran (CNAG-CRG)
Contributors	Regina Becker (UNILU), Juan Arenas (ELIXIR Hub)
Acknowledgements (not grant participants)	
Deliverable type	Report
Dissemination level	Public

B1MG

Document History

Date	Mvm	Who	Description
01/04/2021	0v1	Dylan Spalding (CSC)	Initial draft circulated to WP participants for feedback
20/05/2021	0v2	Nikki Coutts (ELIXIR Hub)	Version circulated to B1MG-OG, B1MG-GB & Stakeholders for feedback
28/05/2021	0v3	Dylan Spalding (CSC)	B1MG-OG/GB/stakeholder comments addressed
28/05/2021	1v0	Juan Arenas (ELIXIR Hub)	Final version uploaded to the EC Portal
23/06/2021	1v1	Xenia Perez, Juan Arenas & Nikki Coutts (ELIXIR Hub)	Template issue fixed. Reformatted final version resubmitted to EC and Zenodo
7/09/2021	1v1r0	Dylan Spalding (CSC)	Revision started (due to rejection of Deliverable by EC)
29/10/2021	1v1r0.1	Dylan Spalding (CSC)	Revised version circulated for feedback
18/11/2021	1v1r0.2	Dylan Spalding (CSC)	Revised version - comments addressed
23/11/2021	1v1r1	Juan Arenas (ELIXIR Hub)	Revised version sent to EC for review

Table of Contents

B1MG

1. Executive Summary

A roadmap is required to communicate the plans for development of the infrastructure to provide secure cross border data access to genetic and phenotypic data to interested stakeholders. The roadmap is a live document, with the initial deliverable being version1, which will evolve as the requirements of more use cases and updates to technology and standards are taken into account.

The main aim of the roadmap presented here is to define the timelines to provide a proof of concept demonstrator of secure cross border data access, describe the technologies required to

B1MG

develop such a demonstrator, and to detail the work that has gone into determining the roadmap and requirement for the proof of concept demonstrator. The roadmap also outlines how it is expected that the work carried out towards the initial proof of concept will support the developing roadmap for 2022 and beyond, as the proof of concept evolves from a demonstrator with synthetic data to a production system accessing real genetic and phenotypic data.

B1MG

2. Contribution towards project objectives

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

	Key Result No and description	Contributed
Objective 1 Engage local, regional, national and European stakeholders to define the requirements for cross-border access to genomics and personalised medicine data	**1.** B1MG assembles key local, national, European and global actors in the field of Personalised Medicine within a B1MG Stakeholder Coordination Group (WP1) by M6.	No
	2. B1MG drives broad engagement around European access to personalised medicine data via the B1MG Stakeholder Coordination Portal (WP1) following the B1MG Communication Strategy (WP6) by M12.	No
	3. B1MG establishes awareness and dialogue with a broad set of societal actors via a continuously monitored and refined communications strategy (WP1, WP6) by M12, M18, M24 & M30.	No
	4. The open B1MG Summit (M18) engages and ensures that the views of all relevant stakeholders are captured in B1MG requirements and guidelines (WP1, WP6).	No
Objective 2 Translate requirements for data quality, standards, technical infrastructure, and ELSI into technical specifications and implementation guidelines that captures European best practice	**Legal & Ethical Key Results**	
	1. Establish relevant best practice in ethics of cross-border access to genome and phenotypic data (WP2) by M36	Yes
	2. Analysis of legal framework and development of common minimum standard (WP2) by M36.	Yes
	3. Cross-border Data Access and Use Governance Toolkit Framework (WP2) by M36.	Yes
	Technical Key Results	
	4. Quality metrics for sequencing (WP3) by M12.	No
	5. Best practices for Next Generation Sequencing (WP3) by M24.	No
	6. Phenotypic and clinical metadata framework (WP3) by M12, M24 & M36.	No
	7. Best practices in sharing and linking phenotypic and genetic data (WP3) by M12 & M24.	No
	8. Data analysis challenge (WP3) by M36.	No
	Infrastructure Key Results	
	9. Secure cross-border data access roadmap (WP4) by M12 & M36.	Yes
	10. Secure cross-border data access demonstrator (WP4) by M24.	Yes

B1MG

Objective 3	**1.** The B1MG maturity level model (WP5) by M24.	No
Drive adoption and support long-term operation by organisations at local, regional, national and European level by providing guidance on phased development (via the B1MG maturity level model), and a methodology for economic evaluation	**2.** Roadmap and guidance tools for countries for effective implementation of Personalised Medicine (WP5) by M36.	No
	3. Economic evaluation models for Personalised Medicine and case studies (WP5) by M30.	No
	4. Guidance principles for national mirror groups and cross-border Personalised Medicine governance (WP6) by M30.	No
	5. Long-term sustainability design and funding routes for cross-border Personalised Medicine delivery (WP6) by M34.	No

3. Methods

This document describes the evolving roadmap to provide the infrastructure required to support secure cross border access to the genetic, clinical, phenotypic and pedigree data generated within the European Union by the end of 2022. The document defines the current standards, functionalities, and components that are required for the Proof of Concept (PoC) demonstrator required in 2022. It outlines the status of each of these, and whether further development is required to ensure they can form a functional part of the proposed infrastructure for the PoC demonstrator.

To ensure the infrastructure solution proposed by WP4 supports all the requirement of other work packages in both Beyond 1 Million genomes (B1MG) and the working groups of the 1+ Million Genomes (1+MG), a scoping paper[1] was written and three workshops were arranged (one on the requirements for synthetic data[2], one on the rare disease use-case[3], and one on Ethical, Legal, and Societal Issues (ELSI) and Data Protection by Design with WP2[4]).

The synthetic data workshop was arranged in December 2020 to investigate the availability of synthetic datasets and the requirements of the different 1+MG use-case working groups for synthetic datasets. By matching the synthetic datasets to the requirements of the use cases, and the capability of each use-case to utilise the proposed infrastructure in time for the PoC demonstrator, a synthetic dataset was chosen to be used along with the rare disease use case.

The scoping paper was written by 1+MG working group 5 and B1MG WP4. The paper defines the five core functionalities required to support secure cross border data access. The paper listed the early adopter nodes that can deploy the proposed infrastructure for the PoC demonstrator, and proposed the first use-case to be addressed by the demonstrator to be the rare disease use case.

A rare disease workshop was held, where the PoC was described, and gap analysis performed to identify gaps where the PoC would not support the proposed rare disease use case. To ensure

---

[1] https://docs.google.com/document/d/1L5imuKcL0wZNQQ1vQmPAXpR8SG42EL4iq1HZJH87Yv0/edit
[2] https://docs.google.com/document/d/1PVUDsVccjuLHynSX5fG1aDB8dCHiqa9woGa7kvF5bE8/edit
[3] https://docs.google.com/document/d/1T4x575AgU3QPzBHS5wYb20ANasll_8_mSJ-Hd4WrPAA/edit
[4] https://docs.google.com/document/d/1kCiCftFTlODjRZOZ-MTktCCCqdgohZDebTEGIlAg588/edit

B1MG

maximum alignment with existing rare disease projects, such as the European Joint Program on Rare Disease[5] (EJP-RD which is a driver project for the Global Alliance for Genomics and Health[6] (GA4GH), ensuring the PoC utilised relevant standards applicable to rare disease was agreed.

The joint workshop with WP2 on ELSI and Data Protection by Design detailed the proposed Proof of Concept infrastructure, the nine main principles of ELSI that must be satisfied, and outlined the process to ensure that the proposed infrastructure conforms to those principles and General Data Protection Regulation (GDPR).

4. Description of work accomplished

In this section, the technological solutions proposed are outlined, including the standards, functionalities, and components chosen to provide the core infrastructure. The core components are the separate components that currently exist and are being actively developed and maintained through a variety of other sources, and these components can be comprised of one or more services or applications. These components together provide the five functionalities that were defined in the scoping paper as being required to provide the required infrastructure. Linking the services and components together is done via global standards where possible, ensuring that the infrastructure is interoperable with external infrastructures where possible.

---

[5] https://www.ejprarediseases.org/
[6] https://www.ga4gh.org/

B1MG

4.1 Infrastructure Functionalities

To provide the two main user-facing functionalities, the Data Discovery and Processing functions, the three other functions are required – Storage and Interfaces, Data Access management Tools, and Data Reception.

As described in the scoping paper, the data processing functionality can include running analytical pipelines (see later under GA4GH), but can also include visualisation or other processing activities. For example, the RD-Connect Genome-phenome Analysis Platform (GPAP) can be regarded as a processing activity as it manipulates the data it accesses to allow the user to visualise or analyse these data. As detailed in the scoping paper, the infrastructure responsibilities of WP4 are the functionalities within the dashed line in Figure 1, which can in general be seen as a Federated EGA instance with extended functionalities (see Table 1 for functionalities of a FEGA instance), such as the Resource Entitlement Management System[7] (REMS).

For this proof of concept, it is put forward that the infrastructure proposed by B1MG demonstrates both user-facing functionalities, supported by the core infrastructure in a federated environment.
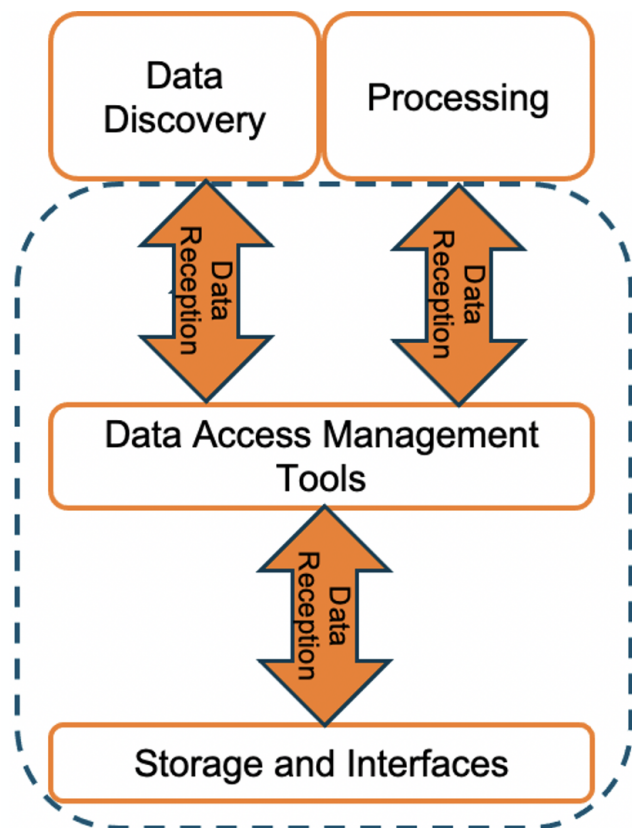

Figure 1: Functionalities described in the scoping paper, and their relationship to the user facing functionalities of Data Discovery and Processing

---

[7] https://github.com/CSCfi/rems

B1MG

4.2 Main Services

4.2.1 Beacon

The Beacon service provides the Data Discovery functionality. It is a GA4GH standard and a service provided by the GPAP. It allows a user to query a controlled access resource without having gained prior approval from a Data Access Committee (DAC) to access that resource while respecting the privacy of the individuals whose data exists within that resource. As such, the Beacon currently allows queries based on the presence of a particular allele at a certain genomic locus. The Beacon may return a binary yes / no answer, an allele frequency of that particular allele within the queries resource, and additionally the dataset within the resource that the allele (or alleles) exist. The standard supports three levels of access, anonymous, registered, and controlled. Anonymous access is effectively public access, where a user needs to have an identity associated with their query. Registered access requires that the user has an identity, such as an ELIXIR identity, and controlled access requires that the user has an identity and has also been granted permission to query the particular Beacon by the data controller or data access committee. Depending on the individual Beacon, registered access may require that a 'bona fide' researcher attribute is set in the user's identity, which ensures the user is a member of a recognised institution. Each Beacon may determine the type of response for each access level, depending on the requirements of the DAC or data controller of the information within the Beacon.

4.2.2 Federated EGA

The Federated EGA (FEGA) in this context is a technical infrastructure that provides a solution for archiving and distributing genetic and phenotypic data to authorised users only. As such it has components, the archive or file storage, the user permissions service, and the data distribution service. The archive utilises the file storage infrastructure provided by the node, and ensures the files archived at the FEGA are encrypted using the Crypt4GH encryption standard so that the files can only be read by authorised users. The user permissions service stores the access permissions for each user who can access data held at the FEGA node. Each user has an access account, and the files or datasets that the user can access are associated with the user account. This account can be linked to an ELIXIR identity using the GA4GH passport standard. The data distribution service is the service that is responsible for distributing the data from the archive to the user or other authorised location or service. The data distribution service utilises the GA4GH htsget standard, which ensures the data is transmitted over https, a secure encrypted data transfer protocol. The FEGA provides the Data reception and Storage and Interfaces functionalities, and as part of these functionalities maintains logs of user access permissions, file location and access, and data distribution processes.

4.2.3 Federated GPAP

The federated Genome-phenome Analysis Platform allows users to interact with, and visualise genetic and phenotypic data and run analyses. For example, the user can search and visualise variants that occur in individuals with a certain phenotype or disease. The GPAP provides a Beacon service for data held within the GPAP. The GPAP also has a user permissions database that stores the access permissions. The access permissions service of the GPAP is currently being

B1MG

modified to ensure it is compatible with the GA4GH Passport standard. The GPAP provides the processing functionality.

4.2.4 Resource Entitlement Management System

The Resource Entitlement Management Systems[8] (REMS) is a service that allows users to apply for data access, and data controllers or DACs to administer data access for data held in one or more FEGA nodes or data hubs. It provides both a web interface (see here[9] for a demonstration web interface) and an API, and utilises the ELIXIR AAI ensuring it is GA4GH Passport and AAI compatible. For example, after discovering a dataset of interest via a Beacon, a user is directed to make an access request to a file or datasets which is listed in REMS. The user must be identified using an ELIXIR identity, where they can apply for access to the dataset of interest. Once the application has been made, REMS then automatically notifies the relevant DAC of the request that has been made. The DAC can then log into REMS web interface, and grant or deny the access request. The process of applying for dataset access can be configured by the DAC to support different procedures for applying for data access, for example ensuring the Data Access Agreement is signed, agreeing to any specific terms and conditions, or making sure the proposed access request has been passed by an Ethics Review Board. If the access request is granted by the DAC, REMS ensures the user now has access permissions to the required dataset using the ControlledAccessGrant visa in the GA4GH Passport associated with that user. The ControlledAccessGrant has a configurable validity period (for example minutes), so that after a set period of time the ControlledAccessGrant visa is checked again with REMS to ensure the user still has access to these data. The DAC can revoke data access via REMS at any time. REMS provides the Data Access Management Tools functionality.

4.3 GA4GH Standards Used in the Proof of Concept

In this roadmap, we propose to base the infrastructure technologies on global standards where appropriate as this ensures maximum interoperability both within Europe and worldwide. The Global Alliance for Genomics and Health (GA4GH) is an organisation that sets the technical standards and frames the policy for the responsible sharing of human genomic data within a human rights framework. Every GA4GH standard must go through a defined approval process (Appendix 1), which helps ensure that these standards address a real use-case and meet minimum security and ELSI standards as defined by the Data Security and Regulatory and Ethics foundational workstreams. GA4GH standards are also semantically versioned, which ensures that interoperability within the infrastructure can be maintained and tracked as the standards and technology evolves, ensuring sustainability.

As defined in the scoping paper, there are 5 main functionalities proposed for the B1MG infrastructure which can be mapped to one or more existing GA4GH standards (table 1). An overview of each respective GA4GH standard is given below.

---

[8] https://github.com/CSCfi/rems
[9] https://rems-demo.rahtiapp.fi/

Table 1. Example mapping of GA4GH standards used in the Proof of Concept services and components.

Functionality	GA4GH Standard	PoC Component	PoC Service
Data Discoverability	Beacon, DUO	REMS	Beacon
Data Reception	htsget, phenopackets	Data API	
Storage and Interfaces	VCF, SAM / BAM / CRAM, Crypt4GH, phenopackets		FEGA
Data Access Management Tools	AAI / Passports, DUO	REMS	ELIXIR AAI
Processing	TES, WES	IGV	GPAP

A majority of the proposed components or services required for the PoC are open source (Table 2), with the exception of the GPAP.

Table 2: Licenses applying to the service or component implementation for the PoC.

Service or component	Open source	License	Link
Beacon	Yes	Apache 2.0	https://github.com/ga4gh-beacon/beacon-elixir/blob/master/LICENSE
Federated EGA	Yes	Apache 2.0	https://github.com/EGA-archive/LocalEGA/blob/master/LICENSE
Data API	Yes	Apache 2.0	https://github.com/EGA-archive/ega-data-api/blob/master/LICENSE
REMS	Yes	MIT	https://github.com/CSCfi/rems/blob/master/LICENSE
IGV	Yes	MIT	https://github.com/igvteam/igv/blob/master/license.txt
GPAP	No		
ELIXIR AAI	Yes	Apache 2.0	https://github.com/elixir-cloud-aai/elixir-cloud-aai/blob/dev/LICENSE

4.3.1 Beacon

The GA4GH Beacon standard is a data discoverability standard that underpins the Beacon service described above. The current Beacon standard is relatively limited in its utility, allowing users just to query for a particular allele at a particular locus. Within the rare disease use case this allows users to determine if a node contains an individual who has the same variant as to one of interest, for example a candidate causative monogenic variant for a certain disease, which would allow the user to request full access to the underlying data. But as by definition rare disease variants make the individual more susceptible to re-identification, the use of the current Beacon standard within the rare disease use case is limited.

To try and address this, the new Beacon V2 standard is being developed. This major version release, which will need to be reviewed and approved by the GA4GH, will extend the Beacon to allow gene or region level queries, and phenotype queries. This significantly extends the utility of the Beacon for the rare disease use case to allow researchers to query for nodes containing

B1MG

individuals who share variants in the same region or gene, share similar phenotypes or disease, or datasets that correspond to certain data use conditions.

4.3.2 Data Use Ontology

The Data Use Ontology (DUO) is an ontology that describes the data use conditions that are applied to one or more datasets held within a node. As the DUO is an ontology, it is by design machine-readable. The conditions can affect the type of secondary use that can be performed on the original data. For example, the data can be restricted to a certain location, or to research within a certain disease. Each release of the DUO is versioned, ensuring that historic DUO attributes are maintained when the definition of the term is modified to reflect changes in the legal and technological landscape around genomic data.

The DUO can be seen as part of the data Discoverability functionality, and the Data Access management Tools functionality. This is because the machine readable attribute of the DUO allows prospective users to query data sets based on the data use conditions - for example user A may wish to research disease X, so can filter out from available datasets all but that disease. The DUO can also be used to determine data access, for example by determining if the researcher who is applying for data access intends to use the data for a purpose allowed by the Data Access Committee or data controller.

4.3.3 htsget

htsget is a data streaming protocol that allows genetic data to be securely streamed from one location to another. It supports range-based queries, for example a user can request a single region, exon or gene to be streamed to a location. htsget is a secure protocol, with all data transmission running over https. The use of htsget can be used to restrict access to genomic regions, preventing users streaming data from certain genes for example.

4.3.4 Phenopackets

The GA4GH Phenopacket standard is designed to allow phenotypic information to be exchanged in a consistent standardised and structured format, linking detailed phenotypic information with disease, patient, and genetic information for both clinicians and researchers. Phenopackets support the use of ontologies and terminologies, but recommend the use of ontologies to maximise the ability to do algorithmic analyses of phenotypic data. From inception it was envisioned that Phenopackets will map to HL7 FHIR[10], and this work is ongoing[11] to support the standardized exchange of phenotypic information from one system to another, including electronic health records (EHR). Phenopackets are already used by the rare disease community, but are not specific to the rare disease community, for example in the Solve-RD[12] project and RD-Connect[13], which demonstrates their real world use as a standard for the exchange of phenotypic information between nodes or services.

4.3.5 VCF, BAM / CRAM / SAM

The VCF and SAM / BAM / CRAM file formats[14] are used to store genetic data in a standard format. SAM / BAM / CRAM files store the complete genetic data for a genomic region, including

---

[10] https://www.ga4gh.org/news/phenopackets-standardizing-and-exchanging-patient-phenotypic-data/
[11] https://phenopackets-analysis.readthedocs.io/en/latest/fhir-mapping.html
[12] http://solve-rd.eu/
[13] https://rd-connect.eu/
[14] https://samtools.github.io/hts-specs/

B1MG

the depth of coverage for aligned data across these regions, and the quality of the sequenced reads. BAM is a binary version of a SAM file, while CRAM is a newer file format that compresses the data to a greater extent than SAM or BAM, reducing the amount of storage required for a set amount of genomic data.

VCF files store the variation data detected in one or more individuals, ensuring that the file sizes are reduced, but detailing the specific alleles that are different from the reference sequence.

All file types can be indexed creating a separate index file, which is a requirement for distributing the files by htsget.

4.3.6 Crypt4GH

Crypt4GH[15] is an encryption standard that uses one or more encryption keys to encrypt one or more regions of a file, such as a VCF or SAM / BAM / CRAM files, but can be used for any file type of format. The encryption key or keys that are used to encrypt the file itself are added to the header of the file encrypted with secondary keys. To allow decryption of one or more regions of the file, the user just needs the specific key or keys to decrypt the portion of the header which encrypted the region of interest in the file, and then use that key to decrypt the required region. This has the advantage that encrypting the file for different users just needs the specific keys to be encrypted per user, reducing re-encryption overhead. It also allows only specific regions of a file to be decrypted for certain users, restricting access to certain regions or genes.

4.3.7 Authentication and Authorisation Infrastructure and Passports.

The required Authentication and Authorisation Infrastructure (AAI) and Passport standards are defined by the GA4GH. These standards specify the way a user's identity and access permissions can be distributed and understood across a federated network. The standards also allow a user to link multiple identities together. For example, the ELIXIR AAI[16] conforms to the GA4GH AAI[17] standard, and supports the GA4GH passport[18] standard, which are in turn based on the OAuth2.0[19] / Open ID Connect[20] authentication and authorization standards. This allows users who have an ELIXIR identity to link other identities to their ELIXIR identity. This allows individual nodes to maintain their own identity provider, if required, but still allow a user to use their ELIXIR identity to access authorised resources across a federated network. For example, a user can link their ELIXIR identity to an EGA identity, which would allow the user to access the Resource Entitlement Management System[21] (REMS) hosted by CSC[22] to administer user permissions using their ELIXIR identity, and then access the EGA using the same identity while internally the EGA identifies the user using the EGA account. The granularity of data access can be defined by the DAC depending on the ELSI requirements, for example data access can be granted to specific files, files associated with an individual, or to a cohort.

---

[15] https://doi.org/10.1093/bioinformatics/btab087
[16] https://elixir-europe.org/services/compute/aai
[17] https://github.com/ga4gh/data-security/tree/master/AAI
[18] https://github.com/ga4gh-duri/ga4gh-duri.github.io/blob/master/researcher_ids/ga4gh_passport_v1.md
[19] https://oauth.net/2/
[20] https://openid.net/connect/
[21] https://www.csc.fi/en/rems-kayttovaltuuksien-hallintajarjestelma
[22] https://www.csc.fi/en/home

B1MG

4.3.8 Workflow and Task Execution Service

The GA4GH Workflow Execution Service[23] (WES) standard specifies how complex analytic pipelines can be specified in a standardised way to allow these pipelines to run in different locations, for example at different federated node: an important requirement in allowing users to send the analysis to the data rather than bring the data to the analysis which helps to ensure the data remains secure and conforms to the ELSI issues that apply to it.

The Task Execution Service[24] (TES) allows analysis pipelines to run on a wide range of heterogeneous infrastructure. For example, a user may wish to run the same variant calling pipeline on data hosted at three different nodes, but the compute infrastructure used by these nodes is different. While WES allows the pipeline to be specified in a standardized way, TES ensures that the task required to run the pipeline can be executed on different types of compute infrastructure, such as different high performance computing (HPC) or Cloud environments. WES and TES together ensure that identical, repeatable, and consistent analysis can be run on different federated node irrespective of the underlying compute infrastructure.

4.4 Infrastructure Technologies

The infrastructure proposed is shown in Figure 2, which illustrates how the services and their constituent components support the five main functionalities and are supported by the range of GA4GH standards, which supports the requirements of the rare disease use case which are detailed within Section 4.6.

---

[23] https://github.com/ga4gh/workflow-execution-service-schemas
[24] https://github.com/ga4gh/task-execution-schemas
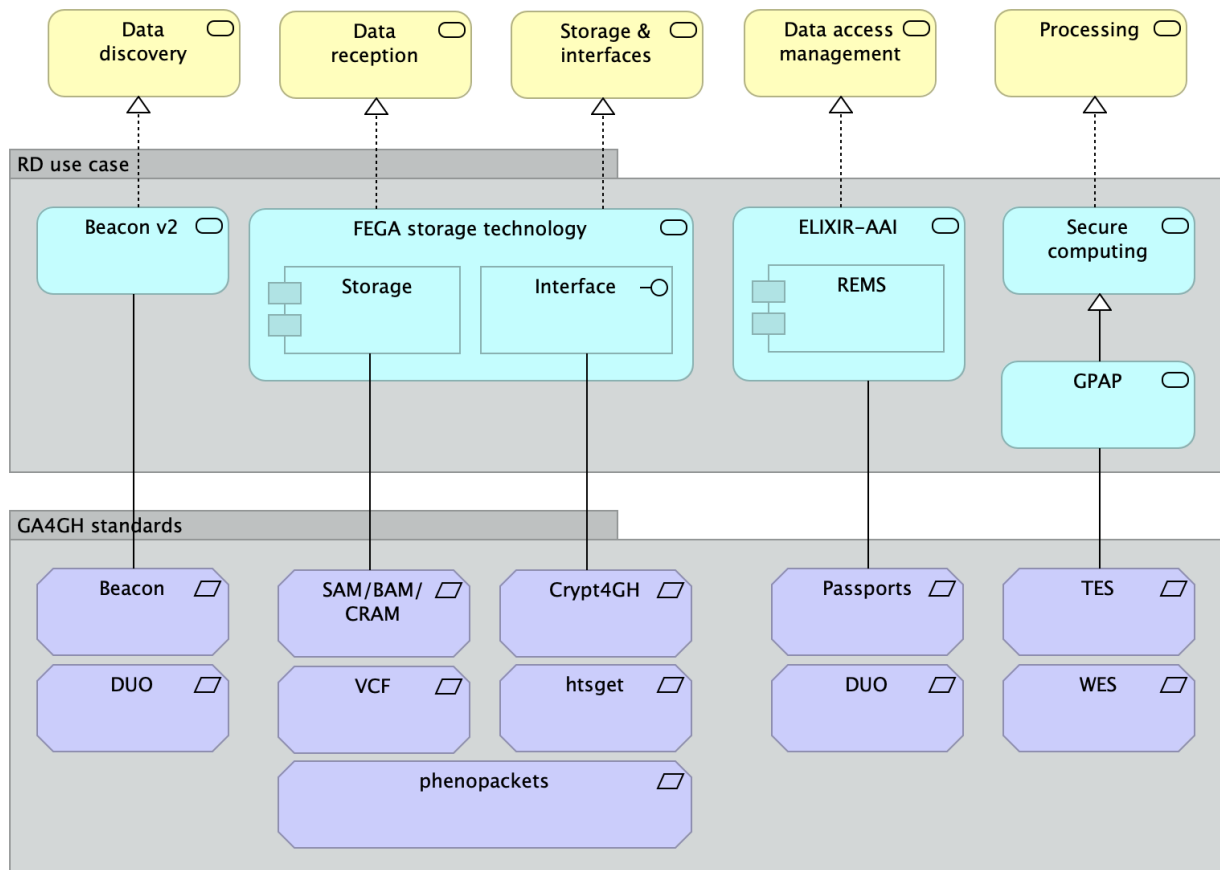
B1MG

Figure 2: The core functionalities as defined in the scoping paper are supported by the services and GA4GH standards. The Beacon, Federated EGA technology, ELIXIR-AAI, and GPAP are all used within the rare disease use case.

A use-case agnostics swimlane diagram is included in Figure 3, which details the use-case for the main actors (data consumer, data controller, and data hubs) in discovering and accessing data in a federated context. It details the steps required by the data consumer and data controller, and the functionalities in response that need to be supported by the infrastructure, and at what stage in the data discovery or data access process these actions occur.
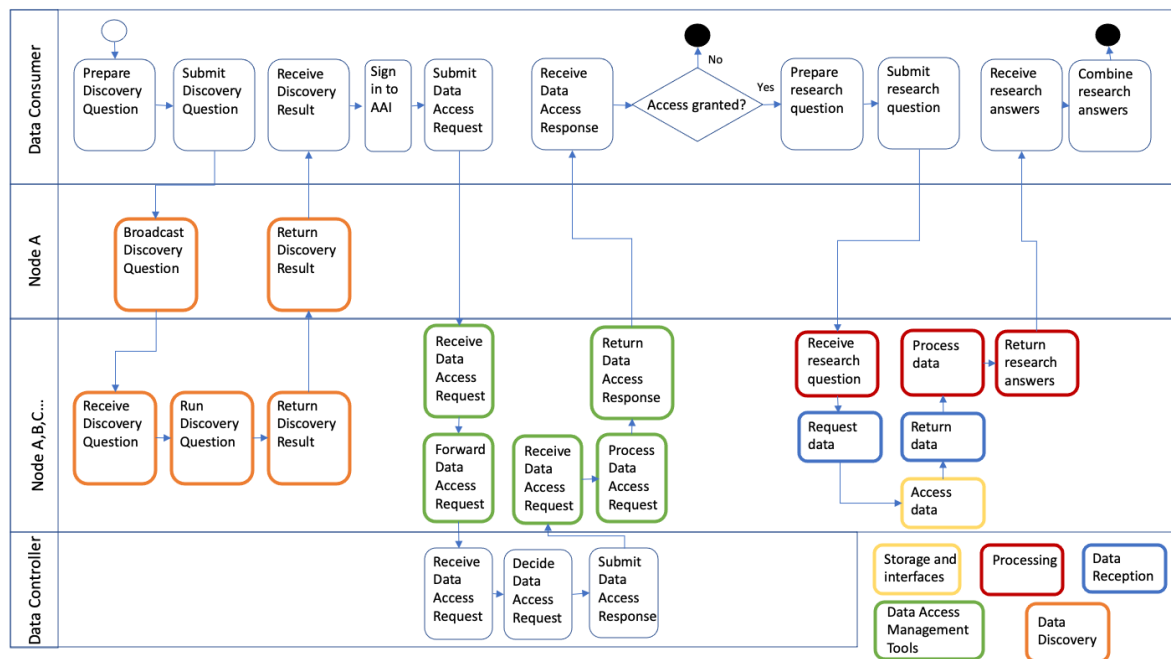
B1MG

Figure 3: User story for the primary actors, the data consumer, the data controller, and the data hubs or nodes. Each process performed by the infrastructure is colour coded, detailing how the requirements of the data consumer and data controller are met.

4.4.1 Justification for infrastructure choices

As B1MG is a CSA project, one of the main requirements for selection of the services and components is the use of existing services or components that are currently being used in a production instance in other projects for accessing controlled access genetic data, and are available currently or will be soon available under open source licence. Additionally the requirement to maximise the FAIRness[25] of the data within the infrastructure by the use of globally defined standards, and consideration of the long term sustainability,for example continued funding and support, and maturity of the services and components were considered in the selection of the technologies used. Federated EGA is based on the technology of the European Genome-phenome Archive, which has been securely archiving and distributing genomic and phenotypic data for over 10 years. Funded development of FEGA is continuing via other projects, such as ELIXIR-CONVERGE[26]. REMS is a service that is currently used to manage access permissions to controlled access resources, such as Findata[27] and the biobank of the Finnish Institute for Health and Welfare (THL) Biobank[28], the Finnish data permission authority Findata used as a model process in TEHDAS[29] joint action, and Euro Bioimaging[30]. Beacon is both a service which can be used to query genomic data over existing resources and a GA4GH standard. The Beacon aids interoperability and federation by natively supporting federated queries, for example through the Beacon Network[31]. Phenopackets are a phenotypic data exchange protocol, which is designed to map to FHIR and hence support the secondary use of EHR data in a research context. htsget is a secure genomic data transfer protocol used in production by EGA among other resources. By using existing production resources, or standards

---

[25] https://www.nature.com/articles/sdata201618
[26] https://elixir-europe.org/about-us/how-funded/eu-projects/converge
[27] https://findata.fi/
[28] https://thl.fi/en/web/thl-biobank
[29] https://tehdas.eu/
[30] https://www.eurobioimaging.eu/
[31] https://beacon-network.org/#/

that have gone through extensive data security and ethics review, such as all GA4GH approved standards, the security and interoperability of the infrastructure is maximised.

4.5 Synthetic Data

To ensure that the proof of concept does not risk exposing real participants' data, it will use synthetic data initially to demonstrate and test the functionality of the system.

There are different requirements for synthetic datasets to meet, for example initial functionality testing, stress and capacity testing, and security and data protection testing. For the secure access demonstrator, the requirements are to demonstrate the functionalities and how they support the use cases, elicit feedback from interested stakeholders, and allow WP4 to extend and develop the roadmap over the final year of the project. The proof of concept will use a modified version of a synthetic dataset[32] developed for the H2020 Common Infrastructure for National Cohorts in Europe, Canada, and Africa[33] (CINECA) project. This dataset includes genetic data from the 1000 Genomes project[34], with phenotypic data generated from the UKBiobank[35] data dictionary[36]. The phenotypic data includes demographic data, diabetes and cardiovascular data, cancer data, and Covid data, making this dataset also applicable for the common complex disease, cancer, and Covid-19 use cases. The dataset currently does not have rare disease Human Phenotype Ontology[37] (HPO) or Orphanet Rare Disease Ontology[38] (ORDO) terms associated with it, so these terms will need to be generated. Additionally, the genetic data of the individuals will be edited to include plausible causative mutations for the associated phenotypes and disease. This will allow any analysis performed to return scientifically relevant results, ensuring that any proposed analysis is fit for purpose.

We will continue to collaborate with other projects and grants who are developing datasets for other use cases, such as INTERVENE[39] WP3 where a range of synthetic datasets are being developed for a variety of use cases. Initial versions of these datasets are due by the end of 2021, with final versions due by mid-2022. Additionally we will continue to collaborate with CINECA, especially WP4 and WP5 who are developing datasets for Joint Cohort Analysis (WP4) including variant calling, eQTL analysis, and genetic risk score determination, and WP5 (Healthcare Interoperability and Clinical Applications) which include cancer biomarker discovery.

As the infrastructure develops, stress and scalability testing will be required, and to support this a larger synthetic dataset will be required which will be for loading at each data hub depending on the expected size of the data hosted by each hub. A synthetic dataset with 1 million genomes has been funded by the Finnish Ministry of Social Affairs and Health which can be utilised for stress and capacity testing of the network in 2022.

---

[32] https://ega-archive.org/datasets/EGAD00001006673
[33] https://www.cineca-project.eu/
[34] https://www.internationalgenome.org/
[35] https://www.ukbiobank.ac.uk/
[36] https://biobank.ctsu.ox.ac.uk/crystal/index.cgi
[37] https://hpo.jax.org/app/
[38] https://www.orpha.net/consor/cgi-bin/index.php
[39] https://www.interveneproject.eu

B1MG

4.6 Rare Disease Use Case

During the B1MG Synthetic Genomes workshop, three primary rare disease scenarios were identified from the WG8 - Rare Diseases working group as detailed in Table 3.

Table 3. 3 example use cases as defined by WG8 and given to WP4 as example use cases.

Scenario	Context	Questions
1	An undiagnosed child with mental retardation and facial dysmorphisms with a negative array-CGH analysis underwent WES, which disclosed a de novo, likely pathogenic, rare variant in FBX011 gene.	1. Is there a relationship between the gene mutation and disease? 2. Are there any other individuals with the same mutation or allelic variant? 3. If there are, what is their phenotype? What can be derived from them for the prognosis? 4. What is the variant frequency across different populations?
2	A girl was diagnosed affected by a Noonan-like syndrome, but she was negative after analysis of an extended panel for Rasopathies. Following WES, a de novo likely pathogenic mutation was found in MAPK1 gene.	1. Are there any other individuals with the same mutation or allelic variant? 2. What is their phenotype and natural history of the disease? 3. What is the variant frequency across different populations?
3	A patient affected by spondylometaphyseal dysplasia with corner fractures, a rare form of AD osteocondrodysplasia of unknown genetic origin underwent WES analysis, which disclosed a de novo likely pathogenic mutation in FN1, a gene previously associated with glomerulopathy with fibronectin deposits, a rare kidney disease (KD) not present in the patient.	1. Are there any other individuals with the same mutation or allelic variant? 2. What is their phenotype? 3. Is it possible to establish any genotype-phenotype correlation? 4. What is the variant frequency across different populations?

By removing duplicated questions from the 3 use cases above, we can determine a minimal set of questions that would support all 3 use cases, and map these to the required user-facing services (Table 4).

Table 4. List of questions and associated functionality, service and type of goal in the PoC

Question	Primary Functionality	User facing Service
Search for individuals with a particular variant.	Data Discoverability	Beacon
Determine the frequency of specific genes with variants in a particular node or set of nodes.	Data Discoverability	Beacon
Explore variants from selected individuals.	Processing	Genome-phenome Analysis Platform

B1MG

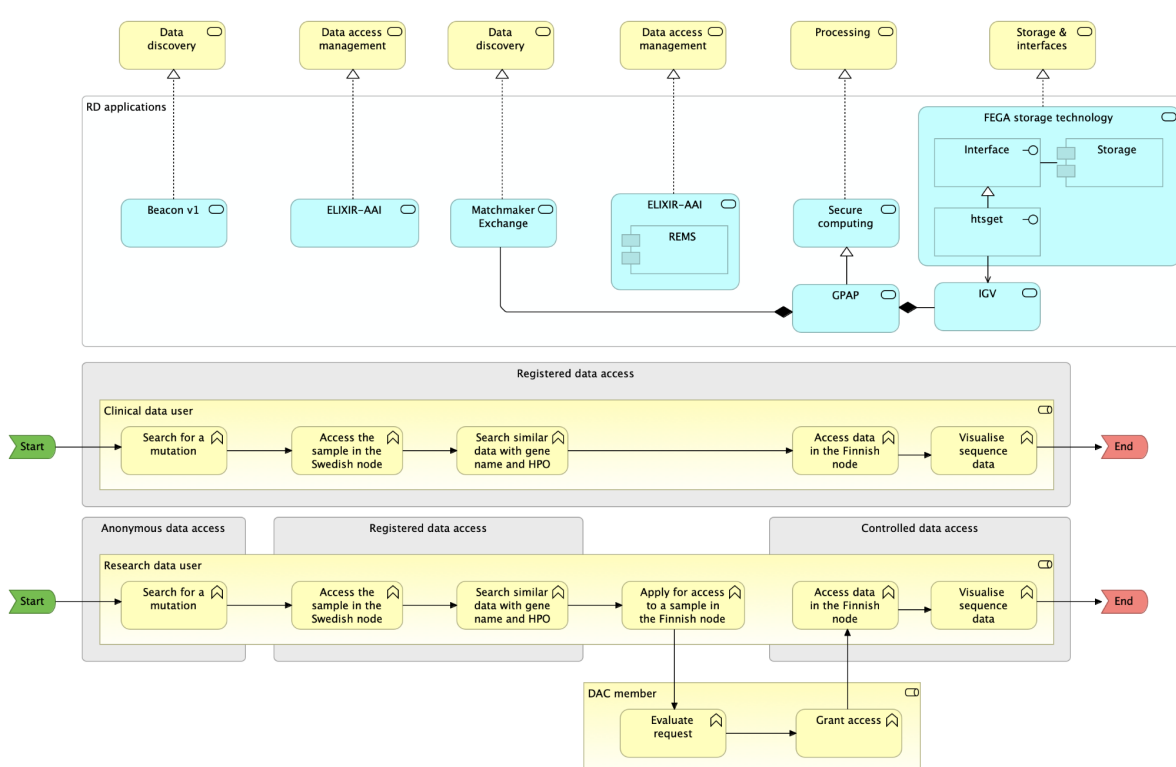|		
Explore phenotypes / disease of selected individuals.	Processing	Genome-pheno me Analysis Platform
Search for individuals with variant(s) in a particular gene or region.	Data Discoverability	Beacon V2
Search for individuals with a particular phenotype or disease.	Data Discoverability	Beacon V2
Determine the frequency of variants, or specified genes with variants, in a particular population.	Data Discoverability	Beacon V2
Determine the frequency of variants in a particular node or set of nodes.	Data Discoverability	Beacon V2
Return the identifiers of the individuals with a certain phenotype and variant, or certain phenotype and variants in a specified gene,or a certain disease and variant, or a certain disease and variants in a specified gene.	Data Discoverability	Beacon V2
Return a set of variants called on different individuals in different nodes which were called using the same analysis pipeline to ensure data harmonisation.	Processing	WES, TES

Figure 4: Diagram detailing how both clinical and research users utilise the proposed standards, components, and services to answer some of the research questions proposed by the rare disease use case.

B1MG

The PoC demonstrates how a rare disease researcher, clinician, and DAC member will be able to utilise the infrastructure to discover an individual with a certain variant of interest across a range or data hubs, apply for access to these data, grant or deny access to these data, access these data, and perform ma set on analyses or processes on these data. In the initial PoC use-case, the case of a rare disease researcher is considered, who utilises the Beacon Network to identify a variant of interest that exists within one or more data hubs, requests access to these dataset using REMS and the ELIXIR AAI, a DAC member who reviews the access request and decides to grant access, and the processes used to allow the researcher to access and visualise the raw genetic data of the participant with the variant of interest plus their immediate family.

As described above in the GA4GH section, the Beacon V2 has not been submitted to GA4GH for approval, and hence the approval of this new standard is required to ensure the full functionality of the proof of concept. Hence a distinction has been made between the existing Beacon standard, and the rare disease use case questions that it can answer, versus the questions which require the extended functionality of the Beacon V2, for which a stretch goal has been defined.

4.7 Other use cases

Over the next 12 months, it is proposed that WP4 elicits feedback from the other working groups to extend the requirements for the core infrastructure to the cancer, infectious disease, and common complex disease use cases. It is understood that there is a risk the initial infrastructure proposed here is too focussed on the rare-disease use-case, but all the standards underlying the infrastructure are agnostic to the use case, and have been developed by the global community to facilitate the use of genetic data within health care and for the secondary use of these data in research. The initial PoC is the way WP4 proposes to elicit more extensive feedback from the other use cases. Additionally, by working with the CINECA project, specifically WP4 and WP5 (federated analysis and cancer use case respectively), B1MG hopes to leverage their developments for B1MG and believes that the fact the standards and services proposed here are also in use with CINECA mitigates that risk for the initial PoC. For example, the inclusion of the federated analysis stretch goal for the rare disease use case is in response to the CINECA WP4 federated variant calling pipeline which utilises the GA4GH TES standard. B1MG is working closely with CINECA for their GWAS use case, investigating the federated genotype / phenotype association analysis using GWAS.

4.8 Data security and GDPR

To maintain trust in a system that allows federated analysis of genomic and phenotypic data, it is crucial that the data is maintained in a way that respects the nine core ELSI principles as defined by WP2. To this end, an initial gap analysis was done with Ethical, Legal, and Societal Issues (ELSI) and associated requirements, and the main General Data Protection Regulation[40] (GDPR) principles as supplied by WP2 to WP4, and the functionalities, components, and services proposed for the core infrastructure and to be implemented as part of the PoC. The initial results are summarised below.

A bottom up approach was initially taken to try and identify any services or associated procedures or processes (where processes are a constituent process of a large component or service) that may not conform to data protection principles (DPPs) or GDPR as part of the Data

---

[40] https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

B1MG

Protection by Design and Default approach described by WP2[41]. This enabled the creation of a heat map (Figure 5), which detailed the data protection principles which were fully met, needed clarification or supporting documentation or protocols, and not met. For example with the logging process within the PoC, this conforms to the accuracy DPP but procedures regarding the length of time these logs should be kept for (Storage Limitation) need to be clarified when these logs contain data subject to GDPR.

Figure 5. Heatmap of how core services, processes and procedures comply with ELSI principles.

Initial DPPs selected for prioritisation during the workshop include data minimization and purpose limitation. Data minimization means here that only data needed for a certain project will be selected and accessed. This applies to both datasets of data subjects as well as data types or parts of the genome relevant for the envisaged use. On the practical level, this is related to purpose limitation as this DPP also requires a limitation of accessed data, here determined by the data use conditions as defined. Both, again, require as a safeguard also access restriction to as few people as possible. In addition, safeguards to ensure purpose limitation need to make sure that neither malicious nor accidental processing for other purposes can take place, thus having a focus on the security setup of the processing.

These two DPPs will be focussed on initially to ensure that all services, procedures, and documentation are identified to ensure these are fully met by the PoC.

WP4 will work to address the services or components that clearly fail to meet these DPPs, and work with WP2 to gain clarification where required on the requirements, and to detail the procedures or documentation that is needed to ensure compliance where required.

In addition to the bottom up approach, we will use the PoC to enable a top down approach, using the PoC to determine how the PoC meets the DPPs from a user perspective, and use this to test specific use cases, such as a user wanting to withdraw consent and needing information

---

[41] https://drive.google.com/file/d/1nhYP2aSX9v3dAqxwdOM0olegUtFz2DTv/view?usp=sharing

B1MG

on how to initiate the process, or a submitter wanting to ensure the data hub meets their responsibilities as a data controller.

Onced the PoC has been developed and released, a deep dive into how the respective services and components support and conform to the ELSI principles. For example, prospective data controllers will be able to manage data access permissions using the PoC REMS instance, and hence be able to give feedback to WP4 on gaps or requirements that are missing from the original PoC. The PoC will also be able to feed into the development of a Data Protection Impact Assessment (DPIA) with WP2. The heatmap (Figure 5) will be used to help prioritise the service or component for each deep dive.

WP4 will also reach out to other projects, such as CINECA, Solve-RD, EJP-RD, and Federated Human Data[42] to determine how these projects are meeting the DPPs, and if there are existing solutions that can be utilised within the 1+MG and B1MG projects. Extensive information is required for the nodes on how to set-up and run a data hub that conforms to all the DPPs, so we will strengthen our collaboration with Federated Human Data to identify, document, and share these requirements and procedures with the prospective data hubs. This documentation will also identify where the responsibility lies for each service or component, or procedure that is needed to ensure the data hub meets the DPPs. The use of the initial size limited synthetic dataset will allow the data hubs to test the processes while not breaking any of the DPPs.

Along with basing the infrastructure and components on GA4GH standards, which have been approved from a Data Security and Regulatory and Ethics respect before approval by GA4GH (Appendix 1), WP4 will use the PoC to engage with WGs 8 - 11 and WP2 to extend the gap analysis to all use cases. The PoC will also allow WP4 to define the Standard Operating Procedures (SOPs) required for each node hosting the infrastructure to ensure compliance with GDPR and to maintain the security of the hosted data. While WP4 defines the infrastructure required to support secure cross border data sharing, WP4 does not have the resources or responsibility to ensure that the hardware and physical infrastructure used to support the components and services recommended by WP4 are suitable and comply to the requirements set by GDPR, WP2, the WGs 8 - 11, and the data controllers. As the use-cases give feedback on the proposed infrastructure WP4 can define and clarify the responsibilities of the nodes with respect to security and data protection.

4.9 Planning

Figure 6 details the high level roadmap, presented as a Gantt chart, for the remainder of 2021, for the ELSI issues, the development of the synthetic datasets required to demonstrate the PoC, and the technical roadmap for deploying and running the infrastructure required for the PoC. There are a set of dependencies which need to be met to allow the PoC to be deployed, which determine the timelines.
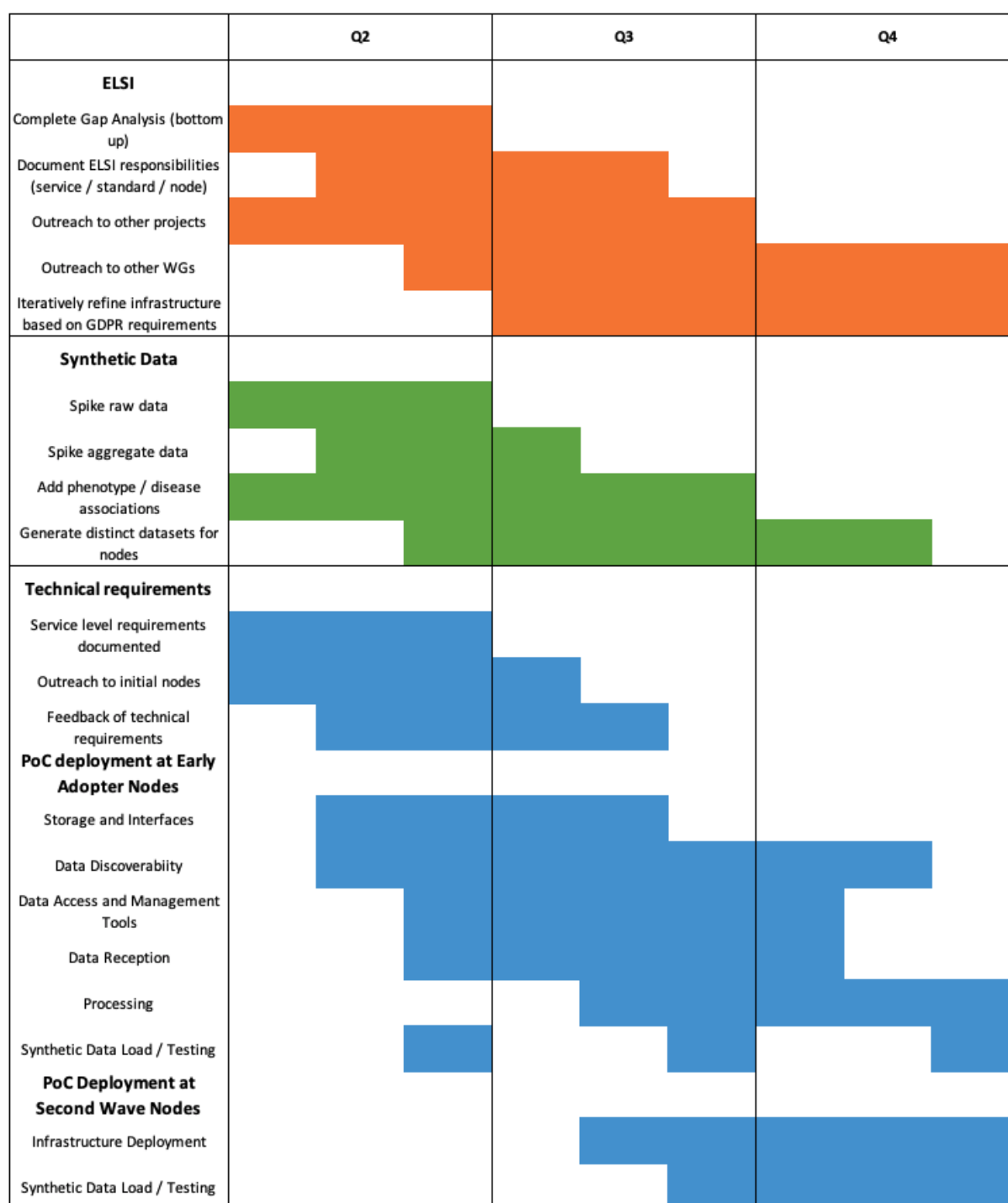
---

[42] https://elixir-europe.org/about-us/commissioned-services/federated-human-data

B1MG

Figure 6. The remaining high-level tasks needed to ensure the PoC is delivered on-time by Q4 2021.

Figure 7 defines the proposed high level roadmap for 2022. After demonstrating the rare disease PoC to the use cases and signatory member states, we will define any gaps in the requirements from the Covid-19, common disease, and cancer use cases not covered by the PoC. This will include extending and adapting the synthetic dataset where possible for the remaining use cases, for example by adding sets of variants with a known association to a common disease, such as diabetes, or cancer related genes. Where necessary additional synthetic datasets developed with collaborators will be used, as described in 4.5. A scaled up synthetic dataset will be required, to allow stress testing of the complete network, and split into more datasets

applicable to the number of individuals each data hub is expected to host. An expansion of the outreach will be needed to engage with the remaining signatories to demonstrate the infrastructure and elicit their expected data volumes and timelines.

Included in the outreach will be the results of the ELSI work, defining the SOPs and responsibilities needed to ensure the data hub corresponds to the DPPs, utilising the information gained by installing, deploying, and running the PoC demonstrator. Collaboration will continue with relevant projects, such as CINECA, Federated Human Data, ELIXIR-CONVERGE, and Federated EGA to maximise leverage work in these areas. For example, Federated EGA are writing the documentation and procedures required to operate a Federated EGA node; it is expected that some of the procedures will be applicable to the B1MG data hubs.

Close collaboration with GA4GH will continue to be required through the ELIXIR and GA4GH strategic partnership, to ensure the standards already deployed are continue to be developed, maintained within the B1MG infrastructure, and kept secure but also to develop the cloud compute processes required to enable more secure federated analyses to be performed on the data as required by WGs 8 – 11.
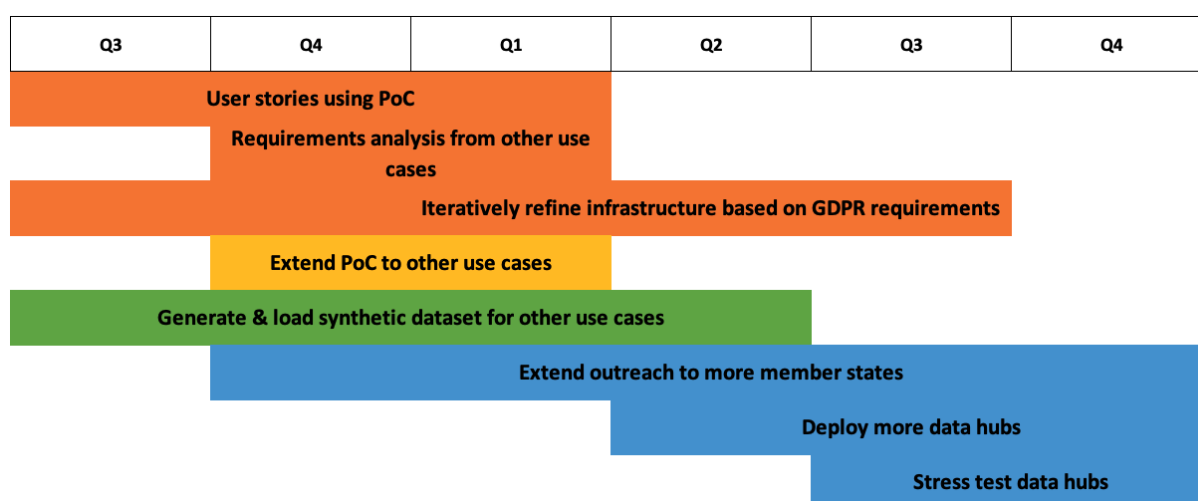

Figure 7: Outline of the plan to Q4 2022 - extending the PoC to other use cases and iterating the infrastructure definition based on GDPR requirements, gaining additional requirements, generating and loading the final synthetic datasets, and deploying the data hubs to the member states.

Risk Mitigation

Due to B1MG being a Coordination and Support Actions (CSA) project; development, implementation, and deployment of the service and components described here cannot be done via B1MG, creating a clear risk that delays and changes to timelines can alter the proposed B1MG roadmap. WP4 is mitigating this as much as possible by using existing services and components that are where possible in production use on real genetic and phenotypic data, or are developments of existing services providing access to genetic and phenotypic data in a secure GDPR compliant manner. The coordination role of B1MG also influenced the use of GA4GH standards as a basis for the infrastructure, and this allows B1MG to help coordinate the requirements of 1+MG, European Health Data Space[43] (EHDS), and other programs to standard

---

[43] https://ec.europa.eu/health/ehealth/dataspace_en

B1MG

setting organisations such as GA4GH which maximises interoperability, but also the re-use of components and services developed based on these standards and using open source licenses.

Data security and GDPR are addressed in collaboration with WP2, but risks of missing gaps are mitigated for by the use of GA4GH standards, each of which must go through an approval[44] process (Appendix 1) including both data security and ethical approval by a group of experts from a range of counties.

The above reasons were the main reasons why the infrastructure has chosen these standards, services, and components.

As B1MG is a CSA project, one of the main requirements for selection of the services and components is the use of existing services or components that are currently being used in a production instance in other projects for accessing controlled access genetic data, and are available currently or will soon be available under open source licence. Additionally a requirement is to maximise the FAIRness of the data within the infrastructure by the use of globally defined standards. Federated EGA (FEGA) is based on the technology of the European Genome-phenome Archive, which has been securely archiving and distributing genomic and phenotypic data for over 10 years. Funded development of FEGA is continuing via other projects, such as ELIXIR-CONVERGE. REMS is a service that is currently used to manage access permissions to controlled access resources, such as Findata and the biobank of the Finnish Institute for Health and Welfare (THL) Biobank, the Finnish data permission authority Findata used as a model process in TEHDAS joint action, and Euro Bioimaging. Beacon is both a service which can be used to query genomic data over existing resources and a GA4GH standard. The Beacon aids interoperability and federation by natively supporting federated queries, for example through the Beacon Network. Phenopackets are a phenotypic data exchange protocol, which is designed to map to FHIR and hence support the secondary use of EHR data in a research context. htsget is a secure genomic data transfer protocol used in production by EGA among other resources. By using existing production resources, or standards that have gone through extensive data security and ethics review, such as all GA4GH approved standards, the security and interoperability of the infrastructure is maximised.

5. Results

The document describes the proposed evolving roadmap towards a demonstrator for secure cross border data access by M24.

6. Discussion

This document allows all interested stakeholders to understand the proposed timelines for the PoC demonstrator, and therefore at what point their requirements or responses will be required to continue with the development of the roadmap. As the roadmap is a living document input is encouraged from all stakeholders to ensure the document covers as many use cases as possible.

---

[44] https://www.ga4gh.org/how-we-work/ga4gh-product-approval/

B1MG

7. Conclusions

The document delivers the first version of the roadmap to provide the infrastructure to support cross border data access of restricted access human genetic and phenotypic data.

8. Next steps

Continue to engage with stakeholders, especially WG 8 - 11, to continue to evolve the roadmap to ensure it tracks the required development to ensure the demonstrator supports all use cases.

9. Impact

The roadmap, along with the initial scoping paper, establishes the prerequisites and timelines required to provide the demonstrator, and the timings for use case interaction.

10. Appendices

Appendix 1

Once a proposed product has been incorporated into the GA4GH roadmap, the product is submitted to the following bodies for review:

1.   The Data Security Foundational Workstream (DSWS),
2.   the Regulatory and Ethics Foundational Workstream (REWS),
3.   and a specially convened Product Review Committee (PRS).

The PRC consists of three members who are nominated by the submitter but must comprise of one representative from each of the groups below:

1.   A workstream leader form another technical workstream,
2.   a member of a third, different, technical workstream,
3.   and a representative of a driver project which has been involved in the product development.

The PRC must return a unanimous 'Accept' verdict for the product to pass this stage in the approval process. The PRC may also refer the product back to the DSWS or the REWS with any concerns they may have. Additionally, the REWS and DSWS must both 'Accept' the product.

Once all three bodies have made positive assessments of the product, the product is presented to the GA4GH Steering Committee, where the steering committee will vote on whether to approve the product or not. Rejection will require re+submission, though reasons for rejection will be given. Approval by the steering committee will formally make the product an approved GA4GH standard.

B1MG

Once a standard has been approved, changes to the standard must be versioned using semantic versioning. This ensures that interoperability is maintained within the standard while allowing the standard top to evolve as technology and requirements change.

By utilising GA4GH standards where possible, the B1MG infrastructure ensures that the standards have gone through an approval process which addresses security, ethical, and legal issues. Additionally, the approval process ensures that the standard does not duplicate, or replicate the functionality of an existing standard, reducing the proliferation of standards. Major changes require that the standard is re-submitted to GA4GH for approval.

B1MG