Beyond One Million Genomes

D4.2

Secure data access demonstrator

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	D4.2 - Secure data access demonstrator		
Contractual delivery date	31/05/2022	**Actual delivery date**	31/01/2023
Delayed	Yes		
Authors	Dylan Spalding (CSC)		
Contributors	Tommi Nyrönen (CSC), Regina Becker (UniLU), Wei Gu (UNILU), Mallory Freeberg (EMBL-EBI), Juan Arenas (EMBL-EBI), Attila Patocs (NIO), Andreas Scherer (UH), Sergi Beltran (CNAG-CRG), Jeroen Belien (VUMC)		
Acknowledgements (not grant participants)			
Deliverable type	Report		
Dissemination level	Public		

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

Document History

Date	Mvm	Who	Description
16/12/y2022	0v1	Dylan Spalding (CSC)	Initial draft circulated to WP participants for feedback
19/01/2023	0v2	Dylan Spalding (CSC)	WP comments addressed. Version circulated to B1MG-OG, B1MG-GB and Stakeholders for feedback
31/01/2023	1v0	Dylan Spalding (CSC) & Nikki Coutts (ELIXIR Hub)	Comments addressed. Version uploaded to the EC Portal

Table of contents

1. Executive Summary **3**

2. Contribution towards project objectives **5**

Objective 1 5

Objective 2 5

Objective 3 6

3. Methods **6**

4. Description of work accomplished **7**

4.1 Overview 7

4.2 Rare Disease Use Case 8

4.2.1 Synthetic Data 8

4.3 User Story 10

4.4 Applications 11

4.4.1 Beacon 12

4.4.2 LifeScience Authentication and Authorisation Infrastructure 12

4.4.3 Resource Entitlement Management System 13

4.4.4 Genome-Phenome Analysis Platform 13

4.4.5 MatchMaker Exchange 13

4.5 Extension into Cancer Use Case 13

4.5.1 Synthetic data 14

4.5.2 User story 14

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

4.6 Issues 15

5. Results **16**

6. Discussion **16**

7. Conclusions **17**

8. Next steps **17**

9. Impact **18**

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

1. Executive Summary

The aim of the 1+MG initiative with coordination and support from the Beyond 1 Million Genomes (B1MG) project is to recommend technologies, methodologies, and governance models for the 1+MG signatories (member states) that support cross-border data access to both genetic and phenotypic data. Prospective users are researchers and clinicians who facilitate the development of personalised medicine across the European Union.

The 1+MG data infrastructure needs to be defined to ensure data managed within the federated network are compliant with the European and national legislation on data protection, security, and ELSI principles agreed in 1+MG. At the same time the aim is to maximise the Findability, Accessibility, Interoperability and Reusability of the 1+MG data, according to FAIR principles.

In this deliverable a Proof of Concept (PoC) was built using existing standards, applications, and services to demonstrate cross border data access for two 1+MG use cases; rare disease and cancer. The intention is to create a technical infrastructure baseline and advancement for the next iteration of implementation data protection principles according to the GDPR, and agreed in discussion with B1MG WP2. The work is not complete, and will continue during 2022-2027 in the European Genomics Data Infrastructure project. As these data are envisioned to be located in distributed nodes hosted in different 1+MG signatory countries, these nodes must be interoperable with each other. Ideally, compatibility with other infrastructures or data spaces in Europe can be maximised as well, while considering the timelines to achieve the ambition of the 1+MG initiative, and hence the user stories described here do not necessarily correspond to the user stories required to conform with the 1+MG data governance. Global standards were chosen, where possible, within the PoC to maximise interoperability between the PoC, the 1+MG data infrastructure (as provided by organisations such as CSC in Finland), and with other European-level data infrastructures and data spaces. In 1+MG the applications or services utilised within the PoC to construct the infrastructure service functionalities and the data analysis workflows are strongly recommended to be open-source to enable security review, as well as have permissive software licence allowing redistribution and modifications. The Genome Phenome Analysis Platform, which was used to demonstrate the 'processing' functionality via visualisation for the rare disease use-case, does not yet comply with this overall recommendation, but the underlying infrastructure does not restrict the applications that can be used by the use of common open source standards to facilitate communication between components. Applications and services, where possible, are utilised by users in service production environments across organisations providing existing (research) data infrastructures or resources, again to maximise interoperability and leverage existing or previous developments. Additionally, all applications needed to provide the five functionalities of data reception, data discovery, access management, storage and interfaces, and processing (data analysis) are in active development within their respective communities. Input into the PoC was taken from both WP/WG2 and WP3 / WGs 3 & 4, as well as other related projects such as TEHDAS, CINECA, EJP-RD, as well as the European Health Data Space. The PoC demonstrates cross border data access for secondary use by a researcher for both use cases as demonstrated by a video[1] uploaded to Youtube and a presentation[2] to the 1+MG Special Group meeting in November 2022 (Link).

[1] https://www.youtube.com/watch?v=6MtIJA4xXdU
[2] https://drive.google.com/file/d/1FfrJ9TdKmAvXOXYoG50fQ9-cP6BSSC76/view

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

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.	Yes
	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	No
	2. Analysis of legal framework and development of common minimum standard (WP2) by M36.	No
	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.	Yes
	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 has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

Objective 3 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	**1.** The B1MG maturity level model (WP5) by M24.	Yes
	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 deliverable describes and links to a proof of concept (PoC) or demonstrator that explains how secure cross border data access can be achieved utilising a range of international standards, open source or community based applications, for two 1+MG use cases. To do this it has taken input from Work Packages 2 and 3, to ensure that the standards used are relevant to the use cases and that the Ethical, Legal, and Societal issues (ELSI) are addressed and the demonstrator allows a research user to access data across borders without infringing the rights, data security or privacy of the participant.

The demonstrator will allow an analysis into the gaps within the demonstrator which will need to be addressed. The aim is that a production version of the PoC can be later deployed, for example via the Genomic Data Infrastructure (GDI) project.

The PoC does not address user access to genomic or phenotypic data in a primary healthcare or clinical professional role for primary use.

As B1MG is a Coordination and Support Action (CSA) project, development and adaptation of solutions to cross border data access have to be done by other projects, such as CINECA[3] and European Joint Programme on Rare Disease[4] (EJP-RD). Therefore the intention of the PoC was to take existing applications and standards and link these together to demonstrate the five functionalities identified by the B1MG scoping paper[5] to enable cross border data access. Where possible applications and standards were chosen which are not specific to the chosen use cases, to try and ensure the infrastructure is applicable for as many use cases as possible.

[3] https://www.cineca-project.eu/
[4] https://www.ejprarediseases.org/
[5] https://doi.org/10.5281/zenodo.6089583

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

The PoC utilises synthetic data to ensure no risk to identifiable individuals, and these data were modified versions of whole genome, variation, and pedigree files derived from open access resources and the generation of these data for the PoC is described in section 4.2.1.

4. Description of work accomplished

4.1 Overview

WP4 has created a proof of concept (PoC) to define a set of standards, services, applications, and components that would support the five functionalities of the prospective service architecture. The PoC demonstrates these for the 1+ Million Genomes (1+MG) use cases, in the first instance for the rare disease use case (WG8) in 2021, and subsequently in 2022 the cancer (WG9) use case.

The PoC needed to demonstrate primarily Technical Interoperability, as shown in Figure 1, which maps features needed in the 1+MG infrastructure into the European Open Science Cloud[6] (EOSC) interoperability framework[7]. There the PoC needed to define the architecture, shared interfaces, and possible virtualisation platforms required to achieve technical interoperability In addition, the proposed infrastructure must support semantic interoperability, so a variety of international standards were proposed to support this, taking account of the work in WP3. Additionally the PoC must not prevent Organisational or Legal interoperability by applying the Data Protection by Design and Default[8] (DPbDD) principle, and complies with the General data Protection Regulation[9] (GDPR).

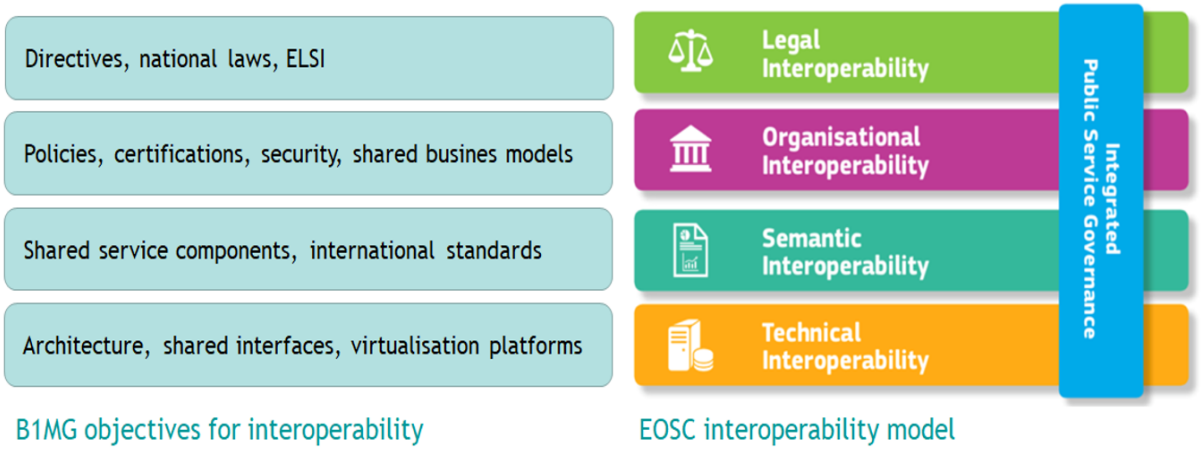


Figure 1: Example of the mapping of the B1MG objectives for interoperability to the EOSC interoperability model.

[6] https://eosc-portal.eu/
[7] https://data.europa.eu/doi/10.2777/620649
[8] https://gdpr-info.eu/art-25-gdpr/
[9] https://gdpr.eu/

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

4.2 Rare Disease Use Case

The scenario that was demonstrated was of a rare disease researcher who was investigating congenital myasthenic syndromes. A child patient was suspected of having a monogenic mutation which caused the clinical features observed in the child, including neonatal hypotonia and distal arthrogryposis, as well as an inability to walk and lower respiratory tract infections. As the mutation was suspected to be monogenic, the child and parents all had their genomes sequenced and the associated bioinformatic analysis was performed which discovered a de-novo mutation in the RYR1 gene within the child. Given this information, the researcher wanted to know:

1. Are there any other individuals with the same mutation or allelic variant?
2. If there are, what is their diagnosis i.e. phenotype? What can be derived from them for the prognosis?
3. What is the variant frequency across populations?
4. Is there an association between the gene mutation and the disease?

4.2.1 Synthetic Data

To enable the PoC to demonstrate the functionalities for such a use case, a synthetic dataset[10] was generated which could be split between different locations. The synthetic dataset was derived from 1000 Genomes[11] data with known deleterious variants 'spiked' into the genomes of the proband with relevant associated phenotypes. The dataset consists of 18 individuals, formed from 6 trios, with two child probands with similar phenotypes and variants in the RYR1 gene. The other probands had other phenotypic and clinical features, as detailed in Figure 2. The dataset consists of genomic information in BAM[12], gVCF[13], and FASTQ[14] format for each individual, and phenotypic information in phenopacket[15] format, as well as pedigree information in PED[16] format. The clinical and phenotypic information was in Orphanet Rare Disease Ontology[17], Human Phenotype Ontology[18] (HPO), and Online Mendelian Inheritance in Man[19] (OMIM) catalogue.

[10] https://ega-archive.org/studies/EGAS00001005702
[11] https://www.internationalgenome.org/
[12] https://samtools.github.io/hts-specs/SAMv1.pdf
[13] https://github.com/broadinstitute/gatk-docs/blob/master/gatk3-faqs/What_is_a_GVCF_and_how_is_it_different_from_a_%27regular%27_VCF%3F.md
[14] https://maq.sourceforge.net/fastq.shtml
[15] http://phenopackets.org/
[16] https://zzz.bwh.harvard.edu/plink/data.shtml
[17] https://www.orpha.net/consor/cgi-bin/index.php
[18] https://hpo.jax.org/app/
[19] https://www.omim.org/

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

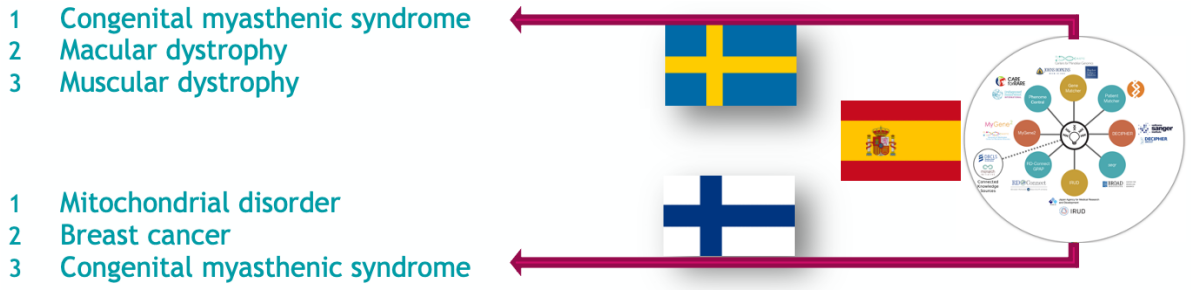
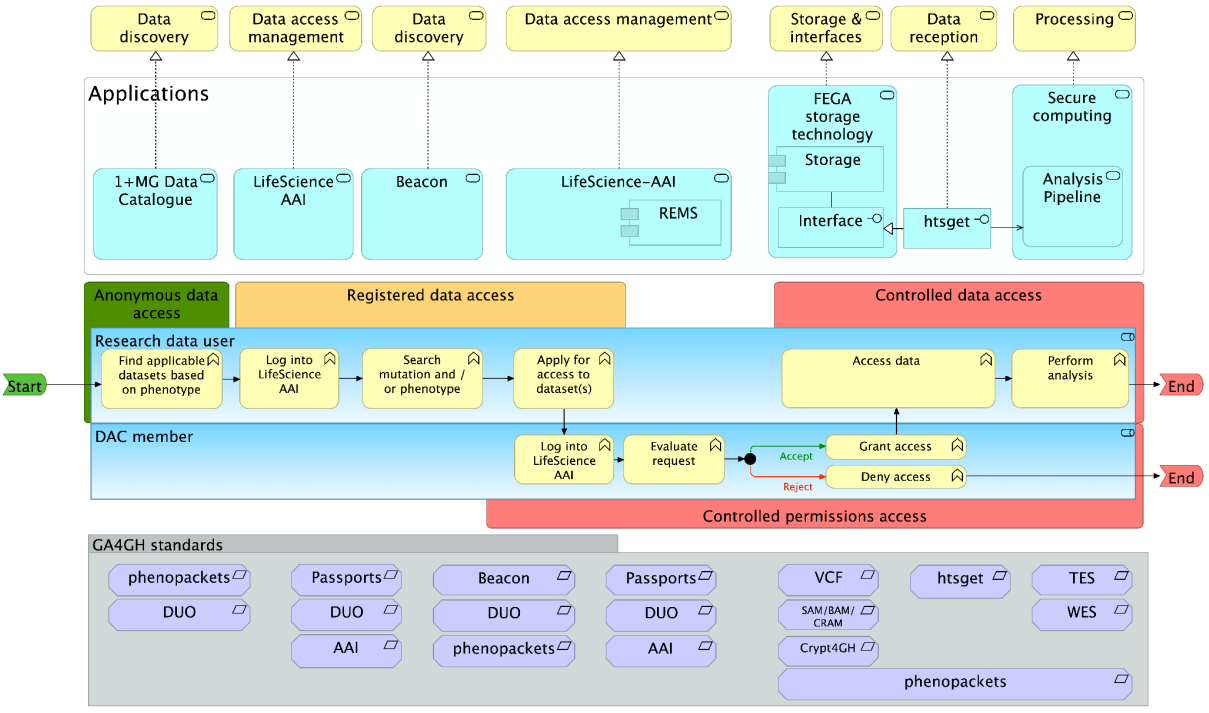


Figure 2: This diagram shows the six trios in the synthetic dataset, as well as the associated diagnosis. Of the six trios, three were located in Sweden and three in Finland. To ensure a positive discovery query could be made between the Finnish and Swedish nodes, a proband with the same diagnosis, but slightly different genetic variants was in both nodes.

The PoC utilises the technologies defined in the architecture diagram Figure 3. The architecture image shows the research user journey (as business functions) from the viewpoint of 2 actors, the research data user and data access committee (DAC) or Data Permit Authority member, with the associated applications (blue) or services above supporting the five functionalities (as business services) which were defined in the Beyond 1 Million Genomes scoping paper - Data Reception, Storage and Interfaces, Data Discoverability, Data Access Management Tools, and Processing . Additionally the Global Alliance for Genomics and Health[20] (GA4GH) standards used to link the different applications are shown at the bottom of the diagram as motivation requirements.



[20] https://www.ga4gh.org/

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

Figure 3: Architecture diagram of the PoC. The functionalities are listed at the top, supported by the different services and applications. These in turn fulfil the requirements of the user story in the middle, and all of this is based on the GA4GH standards listed at the bottom.

To ensure semantic interoperability between the five defined functionalities and the research user journey as defined by TEHDAS[21], the functionalities were mapped to the data journey described in Table 1.

Table 1: Mapping of the TEHDAS data journey definitions to the B1MG functionalities.

Data Journey		Functionality
Data Preparation	Pre-processing to agreed standards, annotation with metadata etc	Data Reception
Data Inclusion	Physical transfer of data incl., legal transfer of data to 1+MG to enable visibility in data catalogue	
Data Storage & Management	Including GDPR Compliant processing environment, data versioning, backup etc	Storage and Interfaces
Data Discovery	Discovery of data using GDPR compliant APIs e.g., Beacon	Data Discovery
Data Access	A mechanism(s) by which the data controller can authorise access to select dataset(s)	Data Access Management Tools
Data Use	Data processing for the approved purposes	Data Processing

4.3 User Story

1. A registered user logs into the ELIXIR Beacon network using the LifeScience AAI to query for the same variant they observed in a participant of a research study into congenital myasthenic syndrome.
2. The user gets a positive response from the Swedish node, and the user is directed to the 1+MG REMS instance to apply for access.
3. The 1+MG data controller or DAC grants access to a virtual cohort of 1+MG data based on the research question data based on individuals with congenital myasthenic syndrome or an associated phenotype, which exist across the 1+MG network both the Swedish and Finnish nodes, but do not necessarily share the same genotype.
4. The user then logs into the Swedish node, using the LifeScience AAI,
5. and uses Matchmaker Exchange to query for other individuals within the virtual cohort with a similar phenotype and genotype.
6. The user receives a positive response from the Finnish node for another variant in the same gene that may have an effect on the phenotype of interest,

[21] https://tehdas.eu/

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

7. and logs into the Finnish node using the LifeScience AAI and visualises the raw data via htsget.

To conform to ELSI recommendations from WG2, the initial search using Beacon in step 1 was done at the registered level as opposed to anonymous level, as under GDPR record level data cannot be queried by anonymous users. An example of the requirements for registered access via the LS AAI are agreeing to the Acceptable Use Policy and Conditions of Use[22], privacy notice[23] and cookie policy, bona fide researcher status[24], or on advice from WP2 a statement of purpose and acknowledgement of joint controllership to ensure GDPR compliance. For step 2, the location of the positive response need not be identified to the end user, it is described here because the requirement of the PoC was to demonstrate cross border data access. For step 3, a 1+MG DAC could grant initial access, subject to veto from the national DAC, in this case Sweden or Finland, which is a proposed access scenario from WG2. Alternatively the PoC would support access solely via a 1+MG DAC, or via separate national DACS, by a single access request routed to separate national DACS, which would create a virtual cohort of individuals whose data are relevant for the research question and conform to ELSI requirements. For steps 4 and 7 depending on the cookie policy of the researcher's browser, plus the ELSI policies of the respective nodes, the researcher may not have to log into each node separately, but be automatically logged in with their LifeScience AAI identity, and additionally the researcher need not know the location of the source data. For step 5, MatchMaker Exchange was used as at this point Beacon Version 2 standard had not been approved by GA4GH, so to query both genotype and phenotype MME was used. Finally for step 7, htsget was used to demonstrate data access within a node or secure processing environment, or htsget could be used to stream data to an alternative compute location, which may be either within the national infrastructure or in another location depending on the ELSI issues, where the data could be processed using a variety of analytical pipelines or tools, such as the Genome-Phenome Analysis Platform (GPAP). Additionally the visualisation of data demonstrated may not conform to ELSI issues as this can be seen as another form of data export outside the node or SPE, but the process of visualisation could equally be another form of analysis within the node or SPE that only returns aggregated or pseudonymised results which conform to the ELSI requirements of the data.

It should be noted that the discovery functionality, via Beacon, is envisioned to happen at registered level to help ensure GDPR compliance. This ensures that any processing of an individual's data conforms to GDPR, and also that the data accessed by the researcher will be the minimal data required for that specific research. In essence, the researcher will utilise the Beacon network query to identify a dataset of interest, to which the researcher will apply for access. In such a scenario, the MatchMaker Exchange query would not occur at controlled access, but this could equally well occur at registered access level subject to the ELSI.

4.4 Applications

The PoC was made up of a set of applications or standards, including Beacon[25], Resource Entitlement Management System[26] (REMS), MatchMaker Exchange[27], LifeScience Authentication

[22] https://lifescience-ri.eu/ls-login/ls-aai-aup.html
[23] https://lifescience-ri.eu/ls-login/privacy-notice-for-life-science-login.html
[24] https://www.nature.com/articles/s41431-018-0219-y
[25] https://beacon-project.io/
[26] https://github.com/CSCfi/rems/
[27] https://www.matchmakerexchange.org/

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

and Authorisation Infrastructure[28], Beacon Network[29], and the Genome-Phenome Analysis Platform[30]. A feature of the PoC is that where possible all these applications utilise open-source standards to maximise interoperability, such as those from GA4GH.

4.4.1 Beacon

Beacon is a GA4GH standard that allows genetic queries to be made to repositories of genetic data without revealing the full genetic sequence of the individual or individuals within the repository. It does this by responding with either a boolean yes or no response, or an allele frequency. Additionally Beacon supports tiered access, which means a Beacon can be open access, where any user can perform a query, registered access[31], where a user must be a bona-fide researcher to query the Beacon, or controlled access, where the researcher must have already been granted access to the data. The particular instance of a Beacon can be set up to respond to queries from users at different tiers, depending on the Ethical, Legal, and Societal Issues (ELSI) requirements of that particular Beacon.

In April 2022 GA4GH approved version 2[32] of the Beacon standard, which extends the supported queries to individuals, biosamples, cohorts, analyses and experiments. With the queries extended to individuals, Beacon version 2 now supports genotype and phenotype queries, similar to those supported by MatchMaker Exchange.

4.4.2 LifeScience Authentication and Authorisation Infrastructure

The LifeScience Authentication and Authorisation Infrastructure (AAI) provides a way for a user to use their institutional identity to access a range of federated resources, without having to remember different identities. It also provides support for access management by listing the datasets or resources the user has access to, as well as the user's roles and attributes. The ELIXIR AAI, which was used in the PoC, was migrated to the LifeScience AAI in April 2022[33]. The LifeScience AAI is compatible with the GA4GH AAI[34] standard, and via the ELIXIR AAI supports the GA4GH Passport[35] standard. The AAI standard leverages the use of the OpenID Connect specification[36] (which provides an identity layer on the industry standard OAuth 2.0 protocol[37]) to authenticate the identity of individuals wishing to access controlled access resources, including both clinical and research genomic data sources. The GA4GH Passport is a standardised way to represent attributes related to a user's identity, as a collection of Visas. There are different visa types, including AffilliationAndRole, AcceptedTermsAndPolicies, ResearcherStatus, LinkedIdentities, and ControlledAccessGrants. The specification also supports custom visa types. For the PoC, the two main visas used were the LinkedIdentities visa, which is used to link institutional identities to a LifeScience AAI identity, and the ControlledAccessGrants which indicates which datasets or resources the user has access to.

[28] https://lifescience-ri.eu/home.html
[29] https://beacon-network.elixir-europe.org/
[30] https://pubmed.ncbi.nlm.nih.gov/35178824/
[31] https://www.nature.com/articles/s41431-018-0219-y
[32] https://github.com/ga4gh-beacon/beacon-v2/
[33] https://elixir-europe.org/AAI-migration
[34] https://github.com/ga4gh/data-security/tree/master/AAI
[35] https://github.com/ga4gh-duri/ga4gh-duri.github.io/blob/master/researcher_ids/ga4gh_passport_v1.md
[36] https://openid.net/connect/
[37] https://oauth.net/2/

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

4.4.3 Resource Entitlement Management System

The Resource Entitlement Management System, or REMS, allows a user to apply for access to controlled access resources, and for Data Access Committees or Data Permit Holders to grant, deny these access requests. Additionally the Data Access Committee (DAC) or Data Permit Authority (DPA) can request additional information for the access request. Both actors, the research user or DAC / DPA member, utilise the LifeScience AAI to access a REMS instance. The REMS instance records all actions related to an access request, and also supports the GA4GH Passport standard. This means that each individual node can administer the access requests for data they administer, but REMS also supports central administration of multiple nodes with data. When a user attempts to access a controlled access resource, the Passport is checked for the appropriate visa, and this visa is issued by the REMS instance that controls access to the resource. This access can be revoked within a REMS instance , and subject to the previous visa validity time, access is revoked immediately.

4.4.4 Genome-Phenome Analysis Platform

The Genome-Phenome Analysis Platform[38], or GPAP, was utilised to demonstrate processing functionality relevant to the rare disease use case. In step 2 of the PoC, the GPAP served the variants through the Beacon standard. In sep 5 of the PoC, the GPAP served the patient phenotypes and candidate genes through the MatchMaker Exchange standard. In the final step of the PoC, the user was able to visualise the raw genetic data archived at the node via htsget, and hence determine the coverage or read alignment over the variant of interest. This processing functionality can be replaced by other applications depending on the use case, for example by Singularity[39], Docker[40] or Podman[41] containers to run bioinformatic analysis or annotation pipelines that are specific to individual use cases. In the PoC, the user would also be able through the GPAP interface to query the data further, setting up variant filters, for example.

4.4.5 MatchMaker Exchange

Matchmaker Exchange[42] is GA4GH driver project and IRDiRC[43] task force with the aim to facilitate the discovery of individuals with similar genotype and / or phenotype across a diverse range of federated nodes via standardised APIs. In the PoC it is utilised via the GPAP to discover an individual with a similar phenotype and genotype in another node. However it is envisaged that to conform with the Data Minimisation Principle of GDPR the discovery process will be done at registered level to only grant access to those data that are both applicable and available with respect to ELSI for the proposed research.

4.5 Extension into Cancer Use Case

As a CSA project B1MG does not have funds to develop nor deploy solutions, so once the Beacon version 2 standard was approved in April 2022 there was a delay before a suitable open source implementation became available to test with the PoC. Once one became available, it was

[38] https://pubmed.ncbi.nlm.nih.gov/35178824/
[39] https://docs.sylabs.io/guides/3.5/user-guide/introduction.html
[40] https://docs.docker.com/get-started/
[41] https://podman.io/
[42] https://www.matchmakerexchange.org/
[43] https://irdirc.org/

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

incorporated into the PoC, in this case the Beacon reference implementation[44] was utilised. This implementation uses phenopackets and VCF files, both GA4GH standards utilised by the rare disease community and within the PoC to ensure the data reception functionality is realised in a standardised way. VCF is also utilised by the cancer use case, and we are working with WP3 and WG9 to map the minimal metadata model within the cancer use case to phenopackets to facilitate data reception within Beacon instances.

4.5.1 Synthetic data

The rare disease dataset utilised for the rare disease PoC was not suitable for the cancer use case, so WG9 described a possible dataset that could be used. The data was based on a melanoma use case from the COLO-829 cell line. The variant data were taken from the previous deep WGS of 3 cancer cell lines[45] as vcf files, and these were loaded to a version 2 Beacon deployed at CSC[46]. WG9 supplied a subset of example data on the melanoma patient, mapped to the evolving minimal metadata model[47] being defined in conjunction with WP3, which was then mapped the National Cancer Institute thesaurus[48] (NCIt) via the Ontology Lookup Service[49] and this subset of data was converted to phenopackets, and loaded to the Beacon.

4.5.2 User story

The discovery part of the user story was also changed for the cancer use case, a patient with metastatic melanoma has a mutation in the BRAF gene, but stops responding to therapy due to an unknown resistance mutation. Subsequent sequencing finds a mutation of interest in the PTEN gene. The researcher wants to know if any other individual has been seen with PTEN mutations conferring resistance to BRAF metastatic melanomas. Hence the discovery query now is 'Do you have any individuals with a mutation in the PTEN gene AND BRAF biomarker'. Additionally the Beacon V2 can also respond with treatments, such as vemurafenib.

[44] https://github.com/EGA-archive/beacon2-ri-api
[45] https://www.nature.com/articles/s41598-019-55636-3
[46] http://128.214.255.34:5053/api/info/
[47] https://doi.org/10.5281/zenodo.6573853
[48] https://ncithesaurus.nci.nih.gov/ncitbrowser/
[49] https://www.ebi.ac.uk/ols/index

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

```
"meta": {
    "beaconId": "org.B1MG.beacon-test-count",
    "apiVersion": "v2.0.0",
    "returnedGranularity": "count",
    "receivedRequestSummary": {
        "filters": [
            {
                "id": "NCIT:C64768",
                "scope": "individuals"
            }
        ],
        "requestParameters": {
            "gene": "PTEN"
        },
        "requestedGranularity": "count",
    },
},
"responseSummary": {
    "exists": true,
    "numTotalResults": 2
}
```


Figure 4: Example of a Beacon V2 POST request and response to a query asking if the Beacon has any individuals with a variant in the PTEN gene who are also being treated with vemurafenib (NCIT:C64748).

As well as the discovery step, the processing functionality for the cancer use case was different to the rare disease use case, with WG9 proposing to replace the GPAP with cBioPortal[50]. There are issues still to be resolved around the operation of cBioPortal as a user controlled application running on HPC resources, as it is not easy to orchestrate the containers within a secure processing environment, but this is a known issue and we have initiated contact with the EOSC4Cancer[51] project to investigate a solution or best practice.

4.6 Issues

The updated cancer PoC utilises an implementation of the Beacon version 2 standard, but the ELIXIR Beacon network does not currently support the full Beacon V2 standard, therefore federated queries based on genotype and phenotype cannot be performed over the network, but only directly with a specific Beacon V2 deployment. Work is ongoing with the ELIXIR Beacon Network Implementation study and the GDI project to support Beacon V2 queries over the network.

Additionally at this point the Beacon Reference Implementation used does not currently support authentication or authorisation via the LifeScience AAI. This functionality was provided by the

[50] https://www.cbioportal.org/
[51] https://eosc4cancer.eu/

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

ELIXIR Beacon Network in the original rare disease PoC. Once the ELIXIR Beacon Network supports the Beacon V2 standard, this functionality will automatically be supported. Additionally, it is expected that one or more Beacon V2 implementations will support the LifeScience AAI in future.

One of the recommendations from WP2 was that to conform with GDPR the Beacon queries must be performed at registered level when non-aggregated or record level data is queried. Without support for either the LifeScience AAI within a Beacon implementation, or support for Beacon V2 within the ELIXIR Beacon network, the Beacon cannot be used with non-synthetic data.

5. Results

The PoC demonstrates cross border data access for a researcher who discovers data within a remote node which relates to their research question, and then accesses this data.

The PoC has been demonstrated via a video[52] on Youtube. This video has been extensively shared and played, with 843 views as of 1 December 2022. It has been played and advertised at a variety of meetings, such as the ELIXIR All Hands Meeting[53] 2022, the GA4GH 9th Plenary[54] in 2021, and the NORDUNet in September 2022 to demonstrate network technology needs for genomics data management in Europe.

Requirements from the WG9 (Cancer) use case have been utilised at the ELIXIR Biohackathon to support developments in federated data analyses, both via containerisation technologies and federated workflows, such as GA4GH compatible Nextflow pipelines.

6. Discussion

The PoC has been demonstrated using two 1+MG use cases, the rare disease use case from WG8 and the cancer use case from WG9. The cancer use case successfully demonstrated an implementation of the Beacon version 2 standard to enable federated discovery queries based on genotype and phenotype, one of the first demonstrations of the Beacon version 2 standard in a real world use case. The delay in the cancer PoC has illustrated the risk to B1MG as a CSA in that it cannot develop existing standards or applications, just propose developments and improvements. However this risk has been mitigated to some degree by the GDI project, which specifically supports the development required for the deployment of the 1+MG infrastructure until 2026.

[52] https://www.youtube.com/watch?v=6MtIJA4xXdU
[53] https://elixir-europe.org/events/elixir-all-hands-2022
[54] https://www.ga4gh.org/wp-content/uploads/Meeting-Report-9th-Plenary.pdf

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

7. Conclusions

The PoC utilises existing applications, services, and standards and demonstrates cross-border data access with these. The use of GA4GH and other open standards where possible helps to ensure long term sustainability of the infrastructure and allows individual applications or services to be updated as requirements evolve without affecting other parts of the PoC. It also facilitates the way the federated network of nodes supports updates, as versioning of each application or standard can be tracked, and coordinated across the network minimising down time. Additionally it maximises the FAIRness of the data within the network, as these standards are used by many partner projects, such as EJP-RD and CINECA, as well as other genomic data repositories and distribution projects, such as EuCanCan[55], CONVERGE[56], and the Federated European Genome-phenome Archive[57]. Within the example projects listed the areas of cancer, rare disease, infectious disease (Covid-19), and population scale genomics are all represented.

8. Next steps

The PoC will form the basis of the GDI Starter kit, a way of deploying a set of 5 vanguard nodes across Europe to demonstrate cross border data access. As part of GDI the development will take place to develop the applications and standards within the PoC, as well as identify other applicable applications or standards via the gap analyses workshop that have already been scheduled, and the combined work package 2, 3 and 4 workshop to ensure the whole recommendations and best practices conform the the Data Protection by Design and Default principles, as well as maximising interoperability between the different data types from each 1+MG use cases.

Once the starter kit has been built, it can be used to help onboard other nodes and use cases to build capacity and knowledge transfer across the 1+MG infrastructure. Additionally it can be used to demonstrate a possible end-to-end solution for cross border data access that can be used to help build collaborations and and with other projects, research infrastructures, and data spaces. While a majority of the components, such as REMS, GPAP, and the Life Science AAI are already in use in production environments, while the PoC deployment is in development phase, and hence the GDI Starter kit will be an evaluation and onboarding tool, while the production level developments will occur within GDI over the next 24 months.

[55] https://eucancan.com/
[56] https://elixir-europe.org/about-us/how-funded/eu-projects/converge
[57] https://ega-archive.org/federated

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724

9. Impact

The PoC has enabled the proposed infrastructure to be better communicated to other 1+MG use cases, as well as both internal and external stakeholders and projects. The initial PoC for the rare disease use case was presented to WG9, enabling WG9 to start to envision how the PoC could address some of their specific questions. Additionally it has been presented to WG10, the infectious disease use case. It has also helped demonstrate the GA4GH standards, and initiated discussion on gaps that exist within these standards, particularly with respect to data use restrictions, identity management, and other ELSI.

B1MG has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 951724