An Overview of the Proof of Concept of Federated Infrastructure for the EU 1+ Million Genomes Initiative

Executive Summary

In February 2021 the EU 1+ Million Genomes initiative approved federated infrastructure scoping paper[1]. A building of a Proof of Concept was subsequently initiated with the support of resources from volunteer signatory member states Finland, Spain, Sweden and Germany. The proof of concept focuses on the 1+MG rare disease use case data access and analysis workflow, but the architecture is agnostic and expandable to other use cases.

A proof of concept delivered in December 2021 demonstrates federated genomic data access to address a clinically relevant use case proposed by the rare disease expert network in Europe. The proof of concept leverages existing global standards, as well as software services and components built in European framework projects like Horizon 2020. We demonstrate how the infrastructure enables access to genomic data for research.

This paper outlines components and standards, and how these standards and components may be used to address similar use cases outside the 1+MG Initiative, and our rationale for choosing these standards and components for the Proof of Concept.

Additionally we address possible risks, and future requirements that the infrastructure will mitigate when it is moving toward service deployment. Procedural, deployment, or resourcing issues are not yet implemented by the EU 1+ Million Genomes Initiative. Agreement on the concept and basis of the infrastructure needs to be obtained next. The agreement needed consists of the infrastructure and the core components and standards that would allow the proof of concept to be expanded with “vanguarding” or deployment-ready data hubs from signatory member states. Procedures and protocols for deploying and maintaining such an infrastructure are expected to advance rapidly during 2022 in collaboration with the 1+MG ELSI group if the proposed infrastructure is seen by the signatory member states as a step forward in the right direction for the EU 1+ Million Genomes Initiative.

[1] https://docs.google.com/document/d/1L5imuKcL0wZNQQ1vQmPAXpR8SG42EL4iq1HZJH87Yv0/edit

Table of Contents

Introduction

The 1+MG declaration describes the intention to have at least 1 million accessible genomes within Europe by 2022. The signatory countries are trying to ensure that the appropriate technical infrastructure is available across the European Union to allow for secure, federated access to genomic data. Interpreting health data, including genomic data, can help health professionals to predict, prevent, and diagnose disease and hence improve the subsequent treatment. To enable this, research and clinical data needs to be linked and analysed while protecting the individual's privacy and conforming to data protection principles, and to do this without moving the data into a central location. With the increasing importance of genomics globally, ensuring the infrastructure is interoperable with and supports global standards helps to ensure that the use of these data are citizen-focussed and patient friendly.
The 1+MG roadmap includes four use cases based on disease category; rare disease, cancer, common and complex disease, and infectious disease. The infrastructure proposed by the Beyond 1 Million Genomes (B1MG) work package 4 must support the requirements of these

four use cases, plus industry and the Genome of Europe. The B1MG provides support and coordination for the implementation of the 1+MG roadmap[2], which in 2021 has been tasked with the translation of the 1+MG mission to pilots of the infrastructure.

Scoping Paper Recommendations and Functionalities

In February 2021 the B1MG WP4 scoping paper[3] was approved by WG5 with the request to endorse the building of a Proof of Concept based on the WG8 rare disease use case. The paper also defined 3 recommendations regarding the proposed infrastructure in 2021-2022:

1. Define a data infrastructure based on shared architecture and interfaces supporting transnational access to existing data sets stored within the signatory states.
2. Move to an infrastructure that provides access to a single, harmonised 1+MG dataset by joining nationally stored data collections together with globally accepted interoperability standards.
3. Provide a 1+MG use case required data discovery, analysis and access to the 1+MG dataset as a service either on national or transnational data processing platforms.

The scoping paper identified five functionalities which must be provided - data discoverability, data reception, storage and interfaces, data access management tools, and processing. Each of these functionalities was considered within the 1+MG infrastructure to be supplied at either national or transnational level, and from further discussion with member states the 3 recommendations described above were proposed. The PoC (Overview with an adjusted workflow based on WP2 feedback which follows the architecture indicated in Figure 1 to be distributed separately, Full length video[4] from July 2021, Presentation[5]), based on the rare disease use case, defined a set of services, standards and software components that could be brought together to perform the five functionalities and form the basis of the proposed infrastructure.

Infrastructure

Outline

As genetic and health related data are a special category of data under the General Data Protection Regulation (GDPR), and to ensure the use of data respects the privacy of the

[2] https://digitalhealtheurope.eu/wp-content/uploads/2020/10/1MillionGenomesRoadmap2020-2022.pdf
[3] https://docs.google.com/document/d/1L5imuKcL0wZNQQ1vQmPAXpR8SG42EL4iq1HZJH87Yv0/edit
[4] https://bit.ly/3jd22MA
[5] https://bit.ly/3aSy0sQ

individual as defined in the 1+MG roadmap, it is required to ensure that access is restricted to both specific individuals and used for specific purposes as processing of these data is prohibited in general. It is the responsibility of the DAC to ensure that the requirements for accessing the data are met, for example conform to the appropriate legal basis or consent under GDPR. These restrictions on data use can extend to geographic location or jurisdiction, research purpose (or preventing research at all), and for-profit use along with other data use conditions, such as requirements not to share the data, to ensure the data is held and processed securely, or in the case of pseudonymised data not to attempt to re-identify the individual. A core component of the proposed infrastructure, and the service on which the WP4 2021 scoping paper was based, is Federated EGA[6], which provides a means of sharing sensitive genetic and phenotypic data in a European context, as described in the short video here[7].

Standards

To prevent siloing of the data held in federated locations, a set of technical standards and policies agreed across the federated network are required. The Global Alliance for Genomics and Health[8] (GA4GH) is a policy-framing and technical standards setting organisation that aims to enable responsible and secure genomic data sharing within a human rights framework as outlined in this video[9] . To try and achieve these aims, the GA4GH is working with the research and healthcare community to develop the standards that allow discovery, access, and analysis of datasets across the world, which can enable scientific advancements, as described here[10], but also support disease diagnosis, for example rare disease as described here[11]. GA4GH also has an extensive approval process; standards must be approved and implemented by driver projects, as well as gaining approval from both the Data Security and Regulatory and Ethics workstreams. By ensuring that the proposed infrastructure for 1+MG is based on these standards, which are the same as the standards utilised by Federated EGA, the approval process for GA4GH standards can be leveraged while ensuring the 1+MG infrastructure is globally interoperable.

These GA4GH standards support the five functionalities from the scoping paper, and the revised deliverable, D4.1[12] (Secure cross-border data access roadmap) gives extended descriptions of the specific proposed software components, services, and standards which provide these functionalities within the infrastructure, as applied to the rare disease use case. Figure 1 shows a generalised use case for data access, with generic processing or data analysis supported plus data discovery aided via the ‘Accessible Genome Dashboard’ which lists the genomes and aggregated metadata of the data within the infrastructure to aid data discoverability and citizen engagement.

---

[6] https://ega-archive.org/federated
[7] https://youtu.be/selSbfNmOyw
[8] https://www.ga4gh.org/
[9] https://www.youtube.com/watch?v=0xfIqBRGqX4
[10] https://www.youtube.com/watch?v=9F7u0rvP9kw
[11] https://www.ga4gh.org/news/ejp-rd-promotes-global-data-sharing-to-bring-hope-to-rare-disease-patients/
[12] https://docs.google.com/document/d/1TamgZTe_yltq0yQ72MY0T4hBFs-EvG5QfKMRsScNoEE/edit

Components

The GA4GH standards underpin and support separate software components that provide the five functionalities. The proposed components are Beacon, ELIXIR AAI and the Resource Entitlement Management System, Federated EGA, htsget, and the Genome Phenome Analysis platform as an example of a containerised software analysis tool for the rare disease use case. Figure 1 shows how the components are envisaged to work together with GA4GH and other global standards to provide the five functionalities.

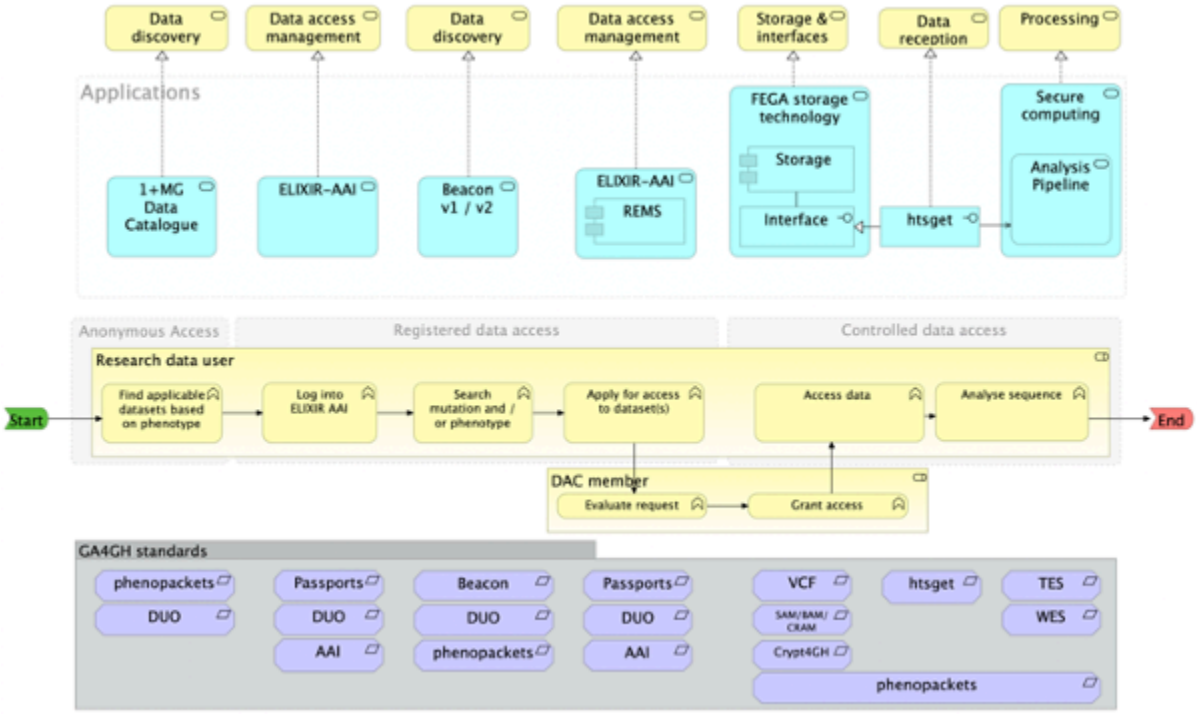

Figure 1: A user story for a researcher who utilises the five functionalities, and their supporting standards and services to discover, access, and analyse genomic data within the 1+MG infrastructure.

Beacon

Beacon[13] is a data discovery tool that allows a user to find genomic data of interest by querying specific areas of the genome. The type of query, and the response given, can be tailored to the user to ensure the query and response comply with GDPR and ELSI. Additionally, developments within the Beacon standard are allowing users to query phenotype or disease data as well as genomic data. Beacon has a variety of implementations, based on the GA4GH Beacon standard.

ELIXIR AAI

The ELIXIR AAI[14] is an infrastructure that allows a user to access multiple diverse resources using a single sign on or identity, such as the user's institutional identity. Such an infrastructure is crucial to allowing a user to access federated and distributed resources without having to log into each resource separately. It also ensures that the data the user has access to are maintained and consistent across a federated network. As the ELIXIR AAI is built on international standards for authentication, authorisation, and identity management, it is interoperable with other infrastructures, such as eduGAIN[15].

Resource Entitlement Management System

The Resource Entitlement Management Systems[16] (REMS) is an access management tool that allows users to apply for, and data controllers or data access committees to administer, access rights and permissions to data. The data may be genomic, clinical, or controlled access data. The data may be divided by individuals or cohorts, and such cohorts may be virtual. REMS support customised application workflows to adapt to any different application procedures required by diverse ELSI data governance recommendations.

Federated EGA

As described in the scoping paper which was based on Federated EGA technology[17], Federated EGA (FEGA) provides the core infrastructure for storing and accessing controlled access data.

[13] https://beacon-project.io/
[14] https://elixir-europe.org/services/compute/aai
[15] https://edugain.org/
[16] https://github.com/CSCfi/rems
[17] https://github.com/federated-ega

All data stored using FEGA is encrypted using the GA4GH Crypt4GH standard, and access requests or attempts, as well as distribution locations are logged and tracked to ensure data security. Additionally FEGA stores the metadata associated with the genomic of phenotypic data stored within the system, helping ensure the confidentiality, integrity, and availability of the data. Federated EGA technology is being used to support the sharing of host genetic data related to the Covid-19 pandemic via the H2020 CONVERGE project[18].

htsget

htsget[19] provides a secure data transfer protocol for genomic data which minimises the data transferred by only transferring the required genomic regions, and ensuring security by transferring the data using the international https standard. This allows data analysis tools to directly access authorised data increasing security.

Genome Phenome Analysis Platform / Secure containers

For the rare disease use case the Genome Phenome Analysis Platform[20] (GPAP) has been used as an example of the processing function, which due to the use of htsget can exist either within the source data hub or in a remote location. A broad range of software for heterogeneous analysis requirements will be needed by the different 1+MG use case working groups. Hardware infrastructures on distributed geographic locations require that data processing functionality must be portable, and hence use existing technical standards, such as the Open Container Initiative[21] (OCI), to help enable portability. A container provides a lightweight immutable infrastructure for application packaging and deployment. An application can be deployed on a range of different environments with little or no modification; if the container runs a well-constructed data analysis, for example calculating the Tumor Mutational Burden (TMB) for the WG9 Cancer use case, or an instance of the GPAP for the rare diseases (WG8) use case, these analysis can be run on different geographical locations on heterogeneous compute or cloud platforms while returning comparable results.

Risk Mitigation

Table 1 lists 8 identified risks and how the proposed infrastructure mitigates them. Overall the proposed infrastructure uses global standards, for example GA4GH standards, and existing

[18] https://inb-elixir.es/news/elixir-converge-receives-further-funding-accelerate-federated-ega-amid-covid-19-data-sharing
[19] http://samtools.github.io/hts-specs/htsget.html
[20] https://platform.rd-connect.eu/
[21] https://opencontainers.org/

infrastructure components to provide the functionalities. This ensures the infrastructure is not monolithic, and can adapt to changing requirements and use cases.

Table 1: Identified risks for the proposed infrastructure and how the infrastructure mitigates these risks.

Identified Risk		Mitigation
1	Deployed infrastructure is not interoperable within Europe	Utilising existing global standards for genomic data and existing repositories.
2	Deployed infrastructure is not interoperable with data and standards on a global scale, or global standards develop away from European standards	Utilise existing and developing global standards, which are supported by a wide range of experts and institutions. Ensure engagement with the development of the standards in future as these standards develop. Continue to engage and collaborate with external projects, such as CINECA, EJP-RD, and EUCanCan for example, plus the EHDS and TEHDAS
3	Deployed 1+MG infrastructure is not interoperable with existing repositories of genomic and phenotypic data	Utilise the standards already used by existing resources, and where possible, the implementations that these resources already deploy.
4	Advancements in the scientific use cases, genomics, or information technology require changes to the infrastructure.	Ensuring the infrastructure is based on global standards helps to ensure that developments affecting the infrastructure can be tracked and deployed as they occur. The infrastructure itself can be a voice within the community driveíng developments based on the needs of 1+MG. Extending the scope of the requirement placed on the infrastructure must be managed with appropriate resources.
5	Resource required for development of the infrastructure is not forthcoming	The infrastructure is based on open source components which are supported by other funding streams and used within production environments. This model maximises the re-use of existing technologies, reduces the cost, and mitigates the risk that the infrastructure will become outdated due to lack of development resources.
6	Different member states have different infrastructure requirements	The design of the infrastructure with separate components communicating via international standards allows components to be added, or replaced, adding or changing the way in which the five functionalities are provided.
7	Software developments in the	The use of containerisation technology helps

	‘processing’ functionality changes over time and between use cases, and ensuring these are used across the federated infrastructure	maximise the portability of software tools used by the different use cases across different hardware infrastructures and hence different data hubs, and hence the ease with which updates can be deployed across the infrastructure
8	Semantic interoperability between use cases maximising data re-use is not supported	The use of standards between the components, such as phenopackets, ensures that the original meaning of the data is maintained across the infrastructure. For example, phenopackets supports the use of ontologies as well as the standards proposed by WP3.
9	Data protection by design and default analyses may identify gaps or data protection risks that need mitigation	Additional R&D may be necessary for some elements, though as described above the infrastructure has been designed to utilise existing services and components. Where required, existing projects and in-kind contributions will be utilised where possible to adapt tools and workflows if necessary.

Interactions with other data spaces

The infrastructure described here supports the technical interoperability of genomic data, and has been primarily focussed on the research context. As shown above via GA4GH video, data made available via this infrastructure can support the diagnosis of rare disease and breast cancer. This is currently demonstrated within the H2020 Solve-RD project, where the GPAP, the European Genome-phenome Archive, and a set of 4 core European Research Networks (ERNs) supported by Solve-RD have diagnosed over 250[22] rare disease patients whose data was made available using the same technologies and standards as described here. But linkage needs to be defined here between the clinical or electronic health record (EHR) which would remain in the European Health Data Space[23] (EHDS), and the GA4GH phenopacket[24] standard, which takes the relevant pseudonymised phenotypic and clinical data into the research or 1+MG domain. Close collaboration with the EHDS or healthcare institutions will therefore be required to ensure that the data is linked, but respects the privacy of the patient, for example via the outline in Figure 2. The core infrastructure for genomic data can be shared between the data spaces, which would drive efficiencies, but the personal clinical data would remain within the health data space.

[22] https://www.nature.com/articles/s41431-021-00859-0
[23] https://ec.europa.eu/health/ehealth/dataspace_en
[24] https://phenopacket-schema.readthedocs.io/en/v2/

The use of international standards for containerisation supports the federation of data analysis, so where the requirements on the data required it remain within the health data space, the analysis could be federated to the data, similar to the Genomics England research environment[25]. Additionally this would address access requirements for clinicians, who need and have the permission to access data without formal DAC approval.

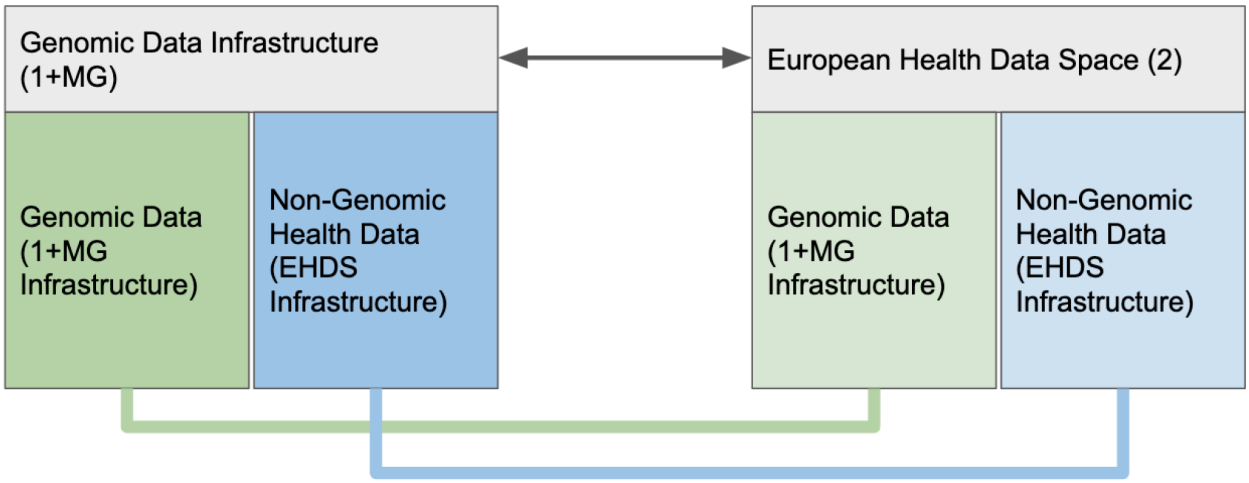

Figure 2: Example of the interaction between the Genomic Data Infrastructure and the European Health Data Space (ELIXIR Europe).

Conclusion

The PoC demonstrates how the proposed infrastructure supports the five functionalities from the scoping paper for the rare disease use case. The use of global standards, not only genomic and health standards such as those provided by the GA4GH, but also international technical standards, maximise the interoperability of the proposed infrastructure with other data spaces and non EU resources. Using existing services or software components and arranging these into a non-monolithic infrastructure maximises the the re-use of resources, minimises cost, and helps to ensure that the infrastructure remains relevant and continues to develop to support the needs of the research and clinical communities, and provide secure federated access to genomic data as required by the 1+MG roadmap.

[25] https://research-help.genomicsengland.co.uk/display/GERE/Using+containers+within+the+Research+Environment

Appendix I: Videos

Proof of Concept Full Version
Secure human data management
The Global Alliance for Genomics and Health
Enabling Federated Discovery, Access, and Analysis of Global Datasets