Developing the German Human Genome-Phenome Archive

A FAIR Data Portal for Human Genomics Data with GDPR compliant Access Control

Jordy D. Orellana Figueroa, Moritz Hahn,
Jens Krüger*, Léon Kuchenbecker, Zehra Hazal Sezer
High Performance and Cloud Computing Group
and Applied Bioinformatics
Eberhard Karls Universität Tübingen
Tübingen, Germany
*jens.krueger@uni-tuebingen.de

Kersten Breuer, Koray Kırlı,
Manikandan Ravichandran
GHGA Office
DKFZ German Cancer Research Center
Heidelberg, Germany

***Abstract*—With recent advances in omics data collection, generating very large datasets of patient omics data for health research has become common, and several data sharing platforms have emerged to allow for researchers to more effectively access broader collections of data. Lacking an existing platform in Germany, the German Human Genome-Phenome Archive (GHGA) project seeks to contribute to the advances in research possibilities by building an archive for omics data using FAIR principles as part of the federated European Genome-Phenome Archive (fEGA), and seeks to allow for data sharing across international borders whilst strictly protecting the privacy of individuals' sensitive data. We present here the basis for the development of GHGA, as well as specific details on the standards used, the organization of the work areas, the software development workflow for both the backend and frontend, as well as the technology used in the project. We also present some of the progress already achieved, as well as provide an overview of the upcoming stages of development and our goals for the future functionality of the GHGA platform.**

Keywords—human genome data, omics, science gateway, FAIR, sensitive data, NFDI

I. Introduction

The German Human Genome-Phenome Archive (GHGA; https://www.ghga.de) is part of the German National Research Data Infrastructure (NFDI). The NFDI initiative aims to establish modern research data management for researchers in all academic disciplines by bringing systematic data management to scientific and research data, providing long-term data storage, backup and accessibility, and network the data across borders. All individual NFDI consortia adhere to FAIR principles (findable, accessible, interoperable, reusable data) and strive to generate additional value for their individual communities through well-structured data management [1]. Specifically, GHGA is committed to supporting researchers from the broader field of biomedical research, dealing with human omics data.

An important part of biomedical research depends on omics data collection from patients, with many applications in biology, translational research, and medicine. Every day, the number of real-life examples of high-volume data generation for personalized therapies are increasing, expanding the toolset for precision diagnostics. The fast-growing amount of data presents both an opportunity for research as well as challenges for handling the infrastructure needs for the data. A growing number of research facilities across Germany and beyond produce an ever-increasing number of extensively characterized data sets. Integrating these local data sets in a broader context is essential to generate additional scientific value. Modern approaches such as machine learning or artificial intelligence can only manifest their full potential with well-annotated, and well-curated, data sets.

As omics data represents the most personal data any individual may share, the balance between our aim to make data open and FAIR for the research community must also be weighed against the protection of the individual's privacy. GHGA strives to build a national infrastructure for the storage and processing of human omics data. It will allow for sensitive omics data to be merged, saved, and analyzed in a uniform, data protection compliant framework.

As a national node of the federated European Genome-Phenome Archive (fEGA), GHGA can bridge the gap between the necessity of adhering to national regulations on data protection and the need to provide a link to international data infrastructures. Consequently, we aim for datasets to be accessible and optimally usable for national and international research. GHGA will also address other practical needs of the research community by providing an efficient, easy-to-use, large-scale analysis infrastructure for biomedical research.

II. Related Work

There are various national programs for the generation, processing, storage, and sharing of omics data: the Genomics England program (https://www.genomicsengland.co.uk/), the Estonian Biobank (https://genomics.ut.ee/en/research/estonian-biobank), the deCODE Genetics program in Iceland (https://www.decode.com/), MASH Data Portal in Vietnam (https://genome.vinbigdata.org/) or Health Data Research Innovation Gateway in the UK (https://www.healthdatagateway.org/). The Health Data Gateway's goal is to help researchers and innovators find and request access to UK health-related datasets. There are also broader data portals for archiving and sharing human genomics data produced for research. The European Genome-Phenome Archive (EGA, https://ega-archive.org/), the database of Genotypes and Phenotypes (dbGaP, https://www.ncbi.nlm.nih.gov/gap/) and The National Center for Biotechnology Information

(https://www.ncbi.nlm.nih.gov/) from the USA are three prominent examples. The NCBI was established by the National Institutes of Health (NIH) to afford for access to biomedical and genomic information. An increasing number of funding agencies and journals require deposition of data to these data portals before accepting and releasing the publications.

GHGA will fill the need for these activities in Germany by building on the existing European network of EGA. EGA has been providing the service for permanent storage, archiving and sharing of various personally identifiable genetic and phenotypic datasets produced by biomedical research projects. Expanding resource needs and customized data privacy laws for different regions necessitated structural changes to this central operation. EGA is currently transitioning from the centralized infrastructure (main node at the EMBL-EBI in Hinxton and a mirror node at the CRG in Barcelona) to a federated setup (fEGA) with national nodes. Many European countries are taking part in this federation, Finland, Sweden, Norway and Denmark along with Germany are the initial implementation nodes.

III. GHGA Community and Standards

By becoming a node of the fEGA, GHGA will adhere to the national regulations on data protection and, at the same time, be intricately linked to international data infrastructures. As such, GHGA datasets will be easy to find, access, and use for research. Moreover, German researchers will take part in shaping the future of international standards for data storing, archiving, sharing, and processing. In close coordination with European partners, we will adapt and extend the fEGA solutions to account for special requirements linked to German legislation. Thereby, we will take on important responsibilities in international research consortia such as the 1+ Million European Genomes Initiative or The Global Alliance for Genomics and Health (GA4GH, https://www.ga4gh.org/).

In a similar fashion as to how fEGA federates across national nodes, GHGA will federate across a growing list of German health-related institutions. This will allow data to be stored and controlled by the producing institution while still being available to other members of the research community. Moreover, this federated architecture will allow GHGA to distribute the burden of funding and operating storage and compute infrastructure and enable georedundancy.

GA4GH focuses on policies and technical standards for responsible genomic data sharing within a personal rights framework. GA4GH has already established a range of standards, of which GHGA would like to utilize the following: Data Use Ontology (DUO) for standardized consent information, Passports for data access policy, Crypt4GH for data encryption, Data Repository Service (DRS) for data storage. Besides aligning to set international standards, GHGA aims to actively engage and help define new standards.

The GA4GH Passport standard covers use cases around digital identity and access permissions [2]. With GA4GH Passports, a machine-readable digital identity is used to convey both the user roles and the data access permissions. These permissions are referred to as visas. They are issued by the data stewards under the control and guidance of the data access committees who are responsible for data sharing decisions of human biomedical data, and they are checked for validity when data is requested. GHGA will use the ELIXIR AAI (https://elixir-europe.org/services/compute/aai, which will soon be migrated to the Life Science Login: https://lifescience-ri.eu/ls-login/elixir-migration-news.html) for identity and permission management, which also incorporates the Passport standard.

The Data Repository Service (DRS) API (https://www.ga4gh.org/news/drs-api-enabling-cloud-based-data-access-and-retrieval/) is a standard released by GA4GH in 2019. It provides standardized access to data independent from the architecture or technology stack of the storage repository. It essentially acts as a data catalog that lists access metadata in a standardized way. Another GA4GH standard being used by GHGA is the encryption file format Crypt4GH [3]. It acts as a container around existing file formats, encrypting files with a symmetric stream cipher to allow for random access of the encrypted data. The symmetric key is encrypted separately via an asynchronous encryption. Ideally, this means that neither the secret, nor the data itself will be stored on disk in a decrypted state throughout the file life cycle. Crypt4GH thus provides a solution to both encryption at rest as well as transfer encryption. It is meant as a file format for bioinformatics research, but in theory can be used for any kind of file formats, and some bioinformatic toolsets, like SAMtools [4], already provide support for Crypt4GH.

Data Use Ontology (DUO) is a standard released by the fEGA, which grants researchers the ability to use human biomedical datasets for controlled-access datasets depending on their research purpose and permissions. Users can tag the datasets with specific usage constraints, which allows them to be discovered based on the permissions granted to health, clinical, and biological researchers. The DUO standard has already been used to annotate over 200,000 datasets throughout the world. For instance, a rare disease researcher can access any dataset that is authorized for commercial and rare disease use cases. The DUO standard contains human-readable explanations and terms, generated by the corresponding data access committees (DAC). The DUO standard is structured with 25 terms that reflect two types of data use: permission and modifier terms. Permission terms contain, for instance: general research use (GRU), health or medical or biomedical (HMB) disease-specific (DS), and population origins or ancestry (POA), which are all clearly approved applications or specialized fields of study. Modifier terms include limitations and requirements within controlled access such as non-commercial use only (NCU), ethics approval required (IRB) and genetics studies only (GSO). DUO will allow GHGA users to have unified standards of data use and make it much easier for researchers to request and obtain access to specific databases.

IV. GHGA Team and Task Organization

The work packages of the GHGA project have been organized into workstreams. Workstreams bring together the experts (team and principal investigators), the goals, and tasks that need to be achieved (see Fig. 1).

Workstream *Architecture*

The architecture workstream focuses on providing the technical architecture and technology for envisioned GHGA functionality by employing community standards and investing in best practices for software development for sustainability. The expected functionality is grouped into high-level milestones and active milestones are broken into finer work

packages and tasks in an agile methodology. The three big GHGA milestones are Archive, Atlas, and Cloud.

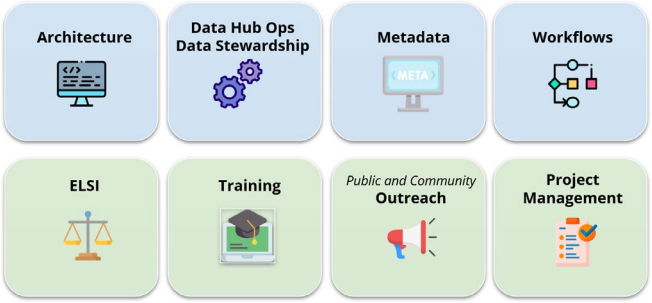

Fig. 1: The GHGA team is organized through the workstreams: *Architecture*, *Data Hub Operations*, *Metadata, Workflows*, *ELSI*, *Training*, *Outreach*, and *Project Management.*

With the Archive phase, GHGA will provide EGA functionality in a federated structure in Germany. This functionality will be expanded with data processing capability and provide insight into portal content by overarching statistics and aggregations with the GHGA Atlas phase. Finally, GHGA will explore becoming a Platform as a Service (PaaS) by providing researchers with tools to use GHGA data and hardware resources for their customized data processing pipelines within the GHGA Cloud phase. The architecture team will be responsible for providing the technical basis for all functionalities.

Workstream *DataHub Operations*

As GHGA transitions from assembly to operation, the GHGA DataHub workstream will provide the support needed for day-to-day activities at the data hubs. These support tasks will be in two domains: data stewardship and DevOps activities. The DevOps team is currently being formed under the Architecture workstream, and when GHGA goes operational, they will be supporting the data hubs for tasks such as resource allocation and software updates, ensuring high fidelity operation. The data steward team will be the user-facing component of GHGA and will provide the support needed for data upload and download requests, covering legal and technical steps. GHGA currently encompasses six data hubs located at the DKFZ (Heidelberg), University of Tübingen, Technical University Munich, University of Cologne, Technical University of Dresden, and the University of Kiel. They are connected to local omics centers (sequencing centers, but also proteomics facilities), which generate a substantial portion of the research omics data and associated technical metadata in Germany. The data hubs themselves host and support the data stewards by providing the storage and compute infrastructure for data deposition, staging, access, and cloud computing. In close collaboration with the Architecture and ELSI workstreams, the operations team will also explore, document, and implement best practices for data protection and security tailor-made for working with human genomics data.

Workstream *ELSI*

The ELSI (Ethical, Legal and Social Implications) workstream develops the ethical and legal context of GHGA. It provides documents, formulates recommendations, and organizes events regarding ELSI topics such as informed consent, patient engagement, FAIR data access, governance, data protection, and legal implementation and interoperability. The ethics team works on informed consent modules for prospective data collection, ethical guidelines for data use and access, and patient engagement strategies. The legal team is contributing to the data protection framework for GHGA and is developing its diverse elements, as well as an international data governance framework for its international embedding. Developing the legal framework for GHGA will be an important challenge for the legal team, especially with the envisioned distributed system of data storage across GHGA nodes. Drafting the contracts outlining the relationship between the central and local GHGA nodes, as well as the latter's compliance with regulations and their obligations as member nodes of GHGA will be an important milestone in the development of GHGA, and will require close cooperation with the DataHubs Operations team.

Workstream *Metadata*

The metadata concept for GHGA includes the definition and implementation of the metadata model, which is aimed at capturing relevant information to accommodate different fields of omics research. The needs of different research communities are identified and reflected in the metadata model, so data submitters and requesters have access to a holistic model that provides all the required and desired information. The metadata concept for GHGA also includes the provision of a submission media, which reflects the implemented model, and provides data submitters with a facilitated way to provide their data. All this is bundled in a technical implementation of the model that can be shared across different workstreams at GHGA. Thereby, GHGA is committed to ensuring compatibility and interoperability with other national and international metadata resources and databases.

Workstream *Workflows*

The workflows workstream will be the gateway for connecting GHGA to the relevant research communities with the help of GHGA bioinformaticians and data stewards. This workstream will identify highly relevant reference datasets by actively working with these communities and create standardized reference datasets for which the underlying data will be processed using state-of-the-art processing workflows. The workstream aims to augment GHGA data by delivering consistently processed community reference datasets, such as aggregated genetic variants data from the German population.

Workstream *Outreach*

The outreach workstream comprises the GHGA communication strategy, both towards scientists and clinicians as potential users of GHGA and toward the public. This includes representation of GHGA at scientific community meetings, holding workshops at conferences both to promote GHGA and the idea of FAIR data sharing, as well as engaging and developing new communities. GHGA also aims to educate the interested public on omics research and FAIR data sharing. The team explores different channels to convey these messages such as brochures, podcasts, and local hub events. The workstream further promotes any products from the other workstreams to the relevant stakeholders.

Workstream *Training*

To support our users, GHGA aims to provide guidance and training on all portal functions, and plans to develop material in various formats, ranging from text-based material to on-demand instruction videos and webinars, as well as in-person seminars. In addition, the workstream plans to offer training for specific communities (e.g., for single-cell data, cancer, or rare

diseases researchers). Furthermore, training material of different formats will be developed around topics relevant to GHGA (e.g., GDPR, consent, and bioinformatics workflows).

Workstream *Project Management*

The workstream encompasses all measures directly related to the management of the project, in particular the internal coordination and communication, management of finances and contractual arrangements within the consortium, monitoring of project progress, project reporting, establishing and supporting the governance structures, and – partially – outreach activities of GHGA.

V. Management and Standard Operating Procedures

An important achievement for GHGA was setting up the software development team and work culture. Currently, we have three subgroups working on backend development, UI/UX design and frontend development, and DevOps tasks. The team culture is built on agile principles, where large scale goals and functionalities are broken down into epics, which are analyzed for dependencies and time allocations. All epics – following the GHGA Epic SOP – scope user journeys and produce technical specifications before the epic tasks start. To implement the agile methodology, the GHGA dev team is using the Scrum framework. We iteratively develop in sprints lasting two weeks each. The Jira project is linked to the GitHub repositories for the individual microservices; this allows for an automatic workflow transition, as well as linking repository branches to individual Jira tasks.

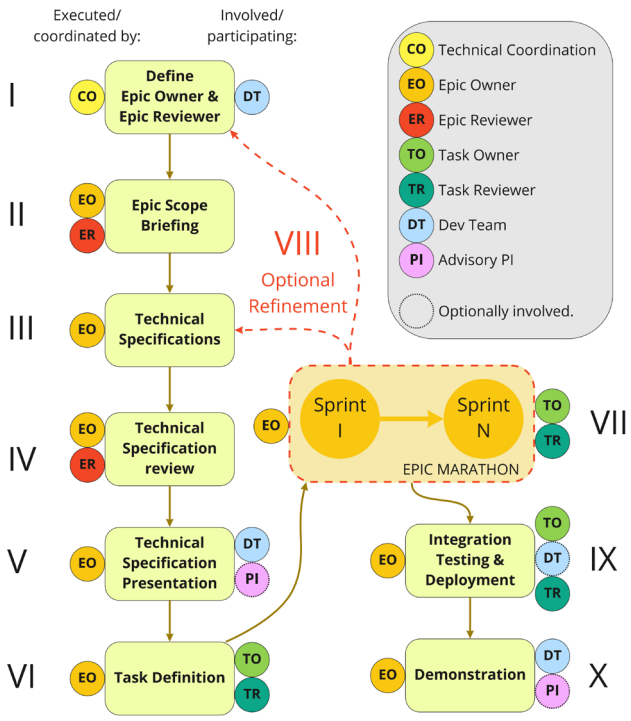

Fig. 2: The GHGA backend development standard operating procedure described in the paragraph below.

In backend development, we subdivide epics into two types: exploratory and implementation epics. Implementation epics should result in a production-ready implementation. Exploratory epics seek to obtain a better understanding of a specific technology or concept and often prepare the ground for the technical specifications of an implementation epic. An epic under development runs through 10 stages, as shown in Fig. 2.

(I): GHGA's Technical Coordinator (CO) together with the Development Team (DT) defines an Epic Owner (EO) and one or more Epic Reviewers (ER). (II): The EO briefs the ERs on the scope of the Epic. Epic scopes are defined ahead of time by the CO. (III): The EO writes the Technical Specifications (TS). Throughout the whole implementation process, the TS stay the sole source of truth. (IV): The ERs provide a review of the TS. (V): The EO presents the TS to the DT. Smaller fixes can be made on the spot, bigger issues lead to the epic going back to (III) for refinements. (VI): The EO, together with all developers that will be part of the epic, define and prioritize tasks. Each task gets assigned a Task Owner (TO) responsible. (VII): The tasks are worked on as part of sprints. Each epic should be finished in a maximum of four Sprints. (VIII): If during the sprint issues are discovered that require modifications to the TS, those must be done via refining the TS and having the ER approve. (IX): Integration testing and deployment are done continuously throughout the epic. (X): Once the implementation has been completed, the work is presented to the broader development team. This presentation includes an overview of the epic scope, a summary of the implementation details and, if applicable, a live demo.

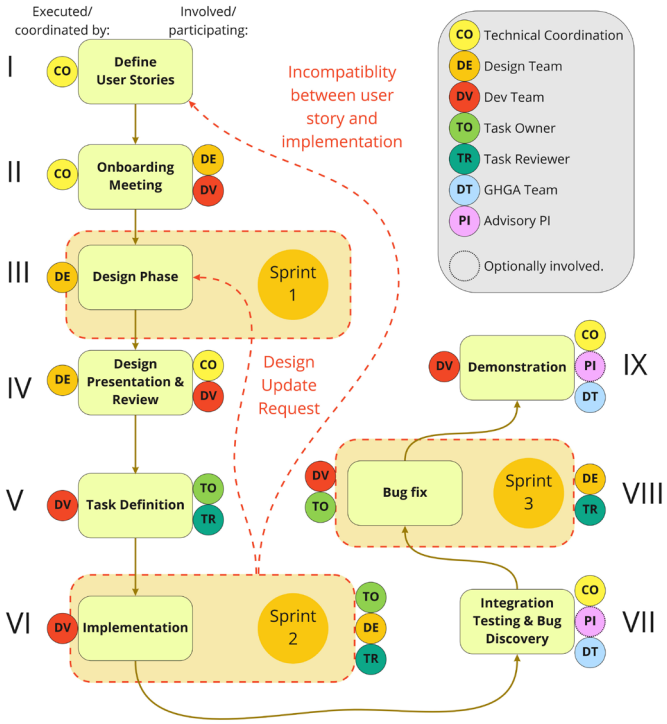

Fig. 3: The GHGA UI/UX development SOPs described in the paragraph above.

In UI/UX development, we follow a similar approach as in backend development but only use implementation epics (see Fig. 3). Moreover, we replaced the writing of the technical specifications with a design phase. The design is done by a dedicated Design Team (DE), while implementation is done by the UI/UX developers (DV). Like the technical specifications in the backend, the design remains the only source of truth during implementation. Due to this separation, a rhythm has been established in which features designed in one sprint are implemented in the next sprint. Following the second sprint, an optional third sprint is added if bugs have been discovered

during integration testing. Integration testing and the subsequent demonstration of the final product is done not only by the developers, but rather, by the broader GHGA team.

VI. Interface and Functionality

Recent advances in the production of substantial amounts of omics data have led to the formation of new platforms on which to store and access them. As building such an accessible platform is one of the primary goals of GHGA, our objective is to develop a user experience (UX) that is at least comparable to that of other omics data sharing platforms.

Within the Archive phase, GHGA will achieve national EGA functionality i.e., serving as the German national node for the federated European Genome-Phenome Archive. GHGA will not only provide EGA functionality, but also build the architecture in which all these features and new integrations can be embedded. To satisfy these requirements, GHGA is building an agile software development team that works with low-detail long-term plans, and highly detailed short-term work packages.

The backend team initially focused on creating an architecture setup that would support GHGA through its lifetime. All software is built by breaking the large-scale domains into separated services (domain-driven design) and sequestering these functions into separate microservices that communicate over a messaging system. This is useful, as functionality can be expanded over time, and services will support a federated structure where central and local functionality will be different. Microservices are also useful since different microservices can be subject to various levels of security layers or scaling.

To kickstart the UI and UX design, we first compiled and reviewed other well-known omics data portals, determining which features functioned best from a user perspective. Through a discussion of the platforms, providing and receiving feedback from other members of the design team on the UI and UX of these platforms, we strove to gain a better knowledge of user behavior, requirements, and motivation.

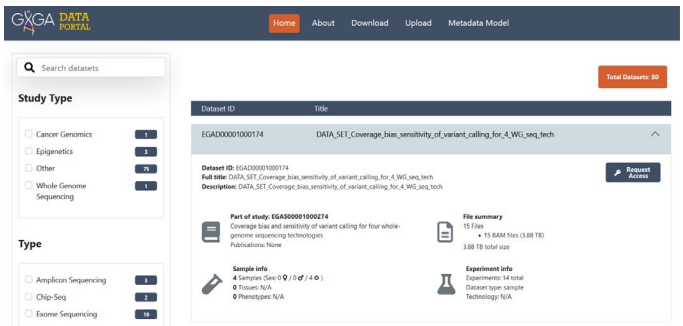

Fig. 4: An image showing a section of the current data portal UI design.

An important part of the design process was incorporating the experience of both users and data stewards using Miro (https://miro.com/), an online collaborative whiteboard platform. Following the completion of this review, we were able to create a template for our desired data portal, including a preliminary list of pages required to fulfil all users. Using Miro, we then began designing a basic wireframe for a dataset browser page with search and filter functionality. The resulting wireframe was brought to a tool built more specifically for UI design and UX prototyping: Figma (https://www.figma.com/). Before any code was written, we discussed the aesthetic components to use in the design of the data portal, such as the layout of elements, fonts, colors, icons, and functional components (see Fig. 4).

The GHGA data portal uses the TypeScript library (version 4, https://github.com/microsoft/TypeScript/releases/tag/v4.6.2), with the TypeScript-based React library (version 17, https://www.npmjs.com/package/react/v/17.0.2); the same major version as the main React library (https://github.com/facebook/react/releases/tag/v17.0.2). For the back-end environment, the data portal uses Node.js version 16 (https://nodejs.org/dist/latest-v16.x/docs/api/), along with its TypeScript implementation library (version 17, https://www.npmjs.com/package/@types/node/v/17.0.21).

In addition, the data portal uses a React-based implementation for Bootstrap (react-bootstrap version 2, https://github.com/react-bootstrap/react-bootstrap/releases/tag/v2.2.1). The appearance of the data portal is based on the default styles from Bootstrap 5 (https://getbootstrap.com/docs/5.1), but for further customization, we use a JavaScript library implementation of SASS ([dart-]sass; version 1, https://github.com/sass/dart-sass/releases/tag/1.49.9). Additional libraries may be included in the future, especially as functions such as authentication and user accounts remain to be added.

Much like the other GHGA microservices, the front end is set-up in a Docker container (https://docs.docker.com/) to package the contents needed to run the microservice, which is then subsequently deployed into a Kubernetes cluster (https://kubernetes.io/docs/home/) hosted on the participating data hubs in Germany (see Section IV). The UI microservice currently directly communicates with the GHGA metadata storage service and the GHGA metadata search service by using (or rather, consuming) their respective APIs to obtain, display, filter, and search datasets available in the metadata storage microservice. Additional connections to other microservices will be implemented as additional features (e.g., data requests, user accounts, DAC) are added to the front-end functionality.

In addition, we use the React Router library version 6 (https://github.com/remix-run/react-router/releases/tag/v6.2.2). The library also allows us to pass search, filtering, and pagination parameters as URL parameters we can parse with the same library, which will serve to easily save and share specific searches and results while also being able to define the desired behavior of the navigation buttons (both in the browser and on the website) to deliver an ideal user experience. As of this writing, the implementation of the functionality of the dataset browser (and search) page is nearly complete. However, additional pages such as log-in, account information, and account settings – all critical for a usable data portal – will also be added in the upcoming stages of development.

VII. Data Handling

GHGA aims to provide access to human omics data. This does not only include sequenced whole genomes but also related data types such as transcriptomes, and single cell analyses. Depending on data format and compression, raw sequence data for a single human can easily reach 200 GB alone, raising the need not only for large storage infrastructures, but also fast transfer speeds to process the data.

These raw data files are organized via a metadata schema (https://github.com/ghga-de/ghga-metadata-schema) that is under active development. The GHGA metadata schema aims

to be compatible with existing standards such as the EGA metadata model to allow for not only national, but also international exchange and cooperation. Broadly speaking, uploaded sample files are grouped into studies by data submitters. The studies can then be rearranged into datasets that are made available to researchers. Such a dataset can contain files from one or multiple studies.

While human omics data have become important in research and clinical application, the re-identifiability of the individual participants has been a major concern in discussions surrounding data protection. The GDPR "[provides] stringent requirements for specific consent and data storage [...], under a so-called 'research exemption,' allows for some flexibility for the processing of personal data for scientific research [...]" [5]. As a primary measure, in accordance with the GDPR, GHGA adopts the safeguard of pseudonymization. All entities within GHGA—such as files, studies, and datasets—are assigned a unique internal ID, as well as an additional public-facing ID. Data and metadata are stored separately and in an encrypted state, while decryption keys are also stored in separate, secure vaults. GHGA collects publicly available – as well as sensible – metadata but never data linking directly to the data subject such as names or addresses.

GHGA will consist of federated data centers storing the data themselves, as well as one central node, storing all metadata. It is anticipated to partially replicate data among the federated data centers to create redundancy and to speed up access. The architecture in both the data centers and the central node is microservice-based (see Fig. 5). These microservices receive access to data on a per-need basis. The data centers have no access to the metadata, while the central node has no access to the data itself. Microservices communicate with each other by publishing and subscribing to different message topics via the event streaming platform Apache Kafka (Version 3). In those event messages, objects are referred to by their internal ID. The public-facing ID is only known by public-facing microservices. For additional security, public-facing microservices never have access to the permanent storage. Data ingress and egress use their own, separate S3-compatible object storage. Uploads and downloads are performed using pre-signed URLs, which give temporary access to a specific object. Data Stewards will help with the manual curation of uploaded data, while Data Access Committees will be responsible for managing access to the datasets.

After uploading and before downloading, files are moved to the permanent storage, which in most cases is also an S3-compatible object storage, through a microservice currently called the interrogation room. The interrogation room performs file operations such as encryption, decryption, compression, and basic quality control such as FastQC. GHGA will use the GA4GH encryption standard Crpyt4GH for transfer encryption, and likely for encryption at-rest as well. In addition, the usage of hardware-based encryption for the permanent storage is under discussion.

The upload and download of data to the data centers as well as metadata to the central node will work via public-facing, RESTful APIs currently under development. For the data download, the DRS API will be used. A command line client for data upload and download using this API is under development and will likely be expanded into a graphical user interface capable of uploading and downloading larger numbers of files. In addition, large data contributors might choose to use their own client implementations that fit best into their established workstreams.

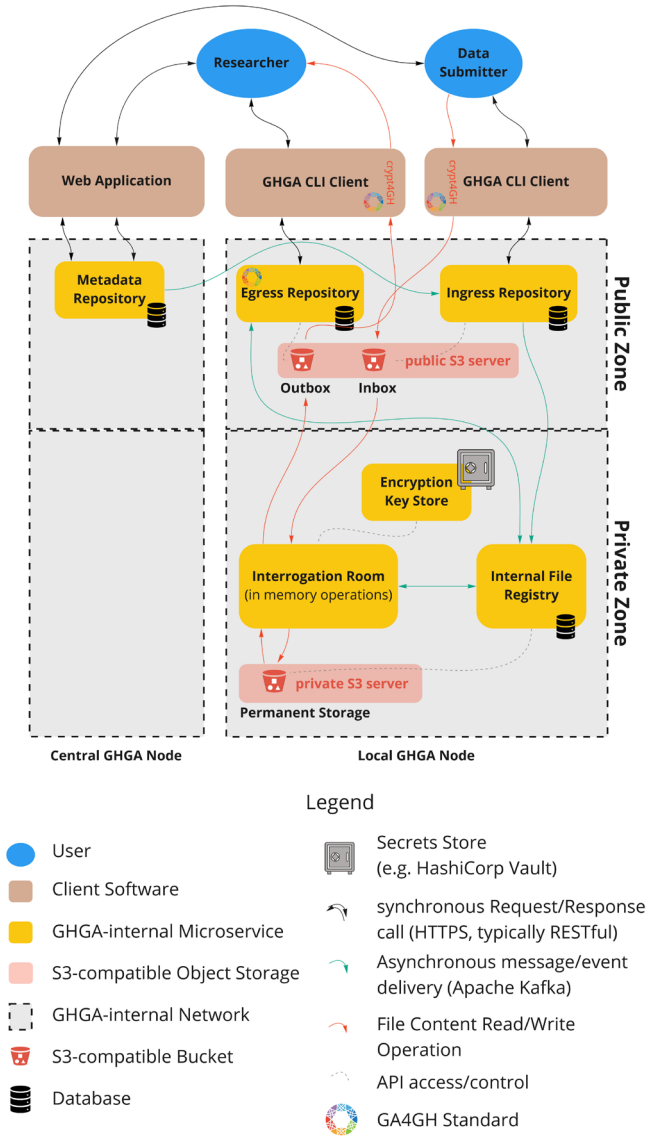

Fig. 5 Current state of backend architecture with the public/private security zones as well as the local and central GHGA nodes.

In 2022, GHGA will roll out the technology that provides GA4GH-compliant archive features that can fulfil EGA functionality. These services will be split into the mentioned central and local services, which will separate the operational flow located at the GHGA Central Hub (public metadata, user management, access control) from the research data management located at the GHGA Data Hub (file transfer, encryption, validation, quality control, archival). To streamline onboarding of new data hubs, deployment support is embedded into the GHGA workflow through widely used technologies like Kubernetes, supported by the GHGA DevOps team.

VIII. Conclusion

GHGA is committed to providing a secure environment for the handling of sensitive human genomics data for the biomedical research community in Germany and beyond. By building on state-of-the-art approaches and technologies, GHGA established a development stream which serves as the basis for

the implementation of GHGAs microservices. In conjunction with a strong infrastructure, these will enable researchers to annotate, analyze, store, publish and share their data in the best sense of FAIR data principles. The roadmap for GHGA's future consists of three phases. The first one, discussed in detail here and currently in development is GHGA Archive, which is planned to have full fEGA functionality, as well as forward metadata to EGA. The second phase will be GHGA Atlas, which will allow for standardized data analysis and visualization. The last phase will be GHGA Cloud, which will explore deploying as a PaaS, as well as provide cloud-based data analysis tools and resources for researchers.

IX. Acknowledegement

We represent only a small fraction of the GHGA consortium and the development team, we thank all colleagues and friends from GHGA for their continuous support. We acknowledge support for GHGA from the German National Research Data Initiative (441914366). The authors also acknowledge the use of the de.NBI Cloud and the support by the Federal Ministry of Education and Research (BMBF) through grant no 031A535A.

References

[1] M. Wilkinson, M. Dumontier, I. Aalbersberg, et al. “The FAIR Guiding Principles for scientific data management and stewardship” Scientific Data 3, 160018, 2016, doi: 10.1038/sdata.2016.18.

[2] C. Voisin, M. Linden, S.O.M. Dyke, et.al. “GA4GH Passport standard for digital identity and access permissions”, Cell Genomics, Volume 1, Issue 2, 2021, https://doi.org/10.1016/j.xgen.2021.100030.

[3] A. Senf, R. Davies, F. Haziza, J. Marshall, J. Troncoso-Pastoriza, O. Hofmann, T.M. Keane, T. M. “Crypt4GH: A file format standard enabling native access to encrypted data”, Bioinformatics, *37*(17), 2753–2754, 2021, https://doi.org/10.1093/bioinformatics/btab087.

[4] P. Danecek, J.K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M.O. Pollard, A. Whitwham, T. Keane, S.A. McCarthy, R.M. Davies, H. Li “Twelve years of SAMtools and BCFtools”, GigaScience, 10(2), https://doi.org/10.1093/gigascience/giab008.

[5] M. Shabani, L. Marelli, “Re-identifiability of Genomic Data and the GDPR”. EMBO Reports, 20(6) 2019, https://doi.org/10.15252/embr.201948316.