White Paper

The Legal and Data Protection Concept for the Operation of the German Human Genome-Phenome Archive (GHGA)

Authors: Simon Parker[1,2*], Katharina Deschler[1], Jan Eufinger[1], Koray Kirli[1], Oliver Kohlbacher[3,4,5*], Lèon Kuchenbecker[6], Fruzsina Molnár-Gábor[2*], Mirjam Schaffer[7], Nicole Schatlowski[3], Oliver Stegle[8,9*], Matthias Struck[10], and the GHGA Consortium

* Corresponding authors: simon.parker@dkfz-heidelberg.de, oliver.kohlbacher@uni-tuebingen.de, fruzsina.molnar-gabor@uni-heidelberg.de, o.stegle@dkfz-heidelberg.de

Version: 1.0 – September 28th, 2023

1 German Human Genome-Phenome Archive (GHGA, W620), German Cancer Research Center (DKFZ), Heidelberg, Germany

2 Faculty of Law, BioQuant Centre, Heidelberg University, Heidelberg

3 Applied Bioinformatics, Department of Computer Science, University of Tübingen, Tübingen, Germany

4 Institute for Bioinformatics and Medical Informatics, University of Tübingen, Tübingen, Germany

5 Institute for Translational Bioinformatics, University Hospital Tübingen, Tübingen, Germany

6 High Performance and Cloud Computing Group and Applied Bioinformatics, Eberhard Karls Universität Tübingen, Tübingen, Germany

7 Legal Department (M280), German Cancer Research Center (DKFZ), Heidelberg, Germany

8 Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany

9 European Molecular Biology Laboratory (EMBL), Heidelberg, Germany

10 Struck Datenschutz, Frechen, Germany

Contents

1. Introduction

The German Human Genome Phenome Archive (GHGA, https://ghga.de) is the German national research data infrastructure for the secure archival and sharing of human omics data. GHGA is being funded within the National Research Data Infrastructure Germany (NFDI) that aims to create national research data infrastructures across all fields of science. GHGA will contribute to the advancement of scientific research, and thereby the development of new preventive, diagnostic and treatment options, by enabling the secure storage and use of human DNA sequences and accompanying phenomic data. To this end, the GHGA Project is developing a data infrastructure to enable the archiving and use of Omics Data for scientific research while meeting the requirements for the protection and security of personal and sensitive data. The GHGA Data Infrastructure will build upon, and promote, international structures and standards, such as technologies and standards developed within the federated European Genome-Phenome Archive EGA (fEGA)[1], the European Genomics Data Infrastructure (GDI)[2], and the Global Alliance for Genomics and Health (GA4GH)[3]. In this context, the GHGA Consortium will also address the legal requirements specific to Germany and make omics data generated or archived in Germany findable to researchers across the world. This engagement will also enable German researchers to help shape future international standards for data exchange and take on leading roles in international research consortia.

GHGA has been designed to operate in a federated manner, with a number of public institutions contributing to the project. They provide the underlying secure data infrastructure and the expertise required to operate the service. A decentralised structure is also the approach preferred by the German federal data protection officer (Bundesbeauftragter für den Datenschutz und die Informationsfreiheit, BfDI) in a recent report[4], as it simplifies the true geo-redundant storage of data and reduces the risk that a data intruder at one site gains access to all the personal data that has been archived.

In this white paper we will describe the legal and data protection concept for GHGA, and how its development was shaped by the legal and data protection obligations under which GHGA operates. In addition to these obligations, there were a number of ambitions for the service offered by GHGA that we sought to achieve when designing the operation of GHGA.

2. Challenges

A key challenge for GHGA is the requirement to process a large volume of omics data, which is considered to be special category of personal data according to Art. 9 (1) GDPR. This processing is being conducted for the purpose of supporting the omics research community. When processing any form of personal data, the GDPR obliges parties to implement a number of protections, and the extent of these are in part shaped by risks that processing has to the rights and freedom of the Data Subjects.

Alongside omics data, there are also a number of other forms of data, both non-personal and personal, that GHGA will process in order to operate a national genome data infrastructure. As such, differing protections may be necessary for the different data types to be processed. Similarly, the different

[1] European Genome-phenome Archive, Federated EGA, https://ega-archive.org/about/projects-and-funders/federated-ega/, last accessed 19.09.2023
[2] GDI, European Genome Data Infrastructure, https://gdi.onemilliongenomes.eu/, last accessed 19.09.2023
[3] Global Alliance 4 Genomics and Health, GA4GH, https://www.ga4gh.org/, last accessed 19.09.2023
[4] BfDI, Tätigkeitsbericht für den Datenschutz und die Informationsfreiheit 2021, https://www.bfdi.bund.de/SharedDocs/Downloads/DE/Taetigkeitsberichte/30TB_21.html, last accessed 19.09.2023

parties involved with GHGA, should have different data processing rights based on the functions they perform.

A further challenge in developing the legal concept emerged from the federated structure of GHGA. Although a decentralised approach brings a number of data protection benefits, because GHGA is not a single legal entity but a consortium of institutions that work together to provide a service, a legal structure that would enable the institutions to work closely together was required. This legal structure also needed to define how data would be processed by the institutions involved in the project and what their roles, within the definitions given in the GDPR, would be. Whilst there is a certain complexity in separate institutions processing special category personal data collectively, we wanted this complex structure to be outwardly simplified for our stakeholders, particularly those depositing or requesting data, to reduce any barriers that may dissuade them from utilising GHGA's services. For example, depositors may be unwilling to share data with GHGA if in doing so there was always a need for multi-party agreements that would require considerable effort to agree and could considerably slow down access to GHGA's services.

a. Data Types

One of the initial tasks in developing a legal and data protection concept for GHGA was to define the different data types that would be processed. Whilst Omics Data is the most important form of data to be processed, it was necessary to consider, in an all-encompassing manner, other relevant forms of data. By doing so, we conceptualised the differing levels of protection that the different data types require and from this starting point, build a structure for the project.

Through discussion within the GHGA team and relevant stakeholders we decided that three broadly-defined data types would be processed as part of the GHGA Project (see Table 1).

Administrative Data	Administrative Data is generated through the operation of GHGA. This category includes, but is not limited to, three main sources of personal data: 1. Data relating to people who deposit Research Data (omics data) with GHGA and those who want to access Research Data. For example, personal data used to verify a Data Submitter's identity within the GHGA Data Infrastructure. 2. Information regarding staff working on behalf of GHGA. For example, personal data generated when they are using the GHGA Helpdesk to respond to users. 3. Information relating to researchers within the wider scientific community. For example, contact information relating to researchers who have answered a survey and who would like to be contacted in the future. Administrative Data is therefore likely to be deemed personal within the meaning of Art. 4 (1) GDPR.
Metadata **(Personal and Non-personal)**	Metadata is defined as any data or information that describes or explains the Omics Data. This is also used by researchers to find and identify datasets that may be useful for their research purposes. Metadata can be both personal and non-personal within the meaning of Art. 4 (1) GDPR.

	Non-Personal Metadata is publicly shared by GHGA to support the FAIR-ness[5] of the Research Data it describes. Non-Personal Metadata is primarily used within the GHGA Data Portal, a searchable catalogue of Metadata that enables researchers to identify Research Data suitable for their research purposes. It may also be shared with the EGA as part of GHGA's commitment to the fEGA network. Non-Personal Metadata are those parts of the metadata that are considered non-personal according to the principles of Art. 4 (1) GDPR in conjunction with Recital 26, sentence 3-4, GDPR. Personal Metadata may include special category personal data, in particular data related to health. For example, Personal Metadata may include detailed phenotypic or clinical information that relates to the person from whom the Research Data was collected. Functionally, Personal Metadata is handled by GHGA as if it were Research Data and is subject to the same level of protection.
Research Data	Research Data refers to any form of human Omics Data that is deposited at GHGA in order to be shared for secondary research purposes and as part of the usual research publication process. This data may include whole genome sequences, proteomics, transcriptomics, or other forms of Omics Data. Research Data is always considered to be a special category of personal data as defined by Art. 9 GDPR.

Table 1: Data Types within GHGA

b. Classifying the institutions

Having defined the different data types to be processed, it was necessary to also classify the institutions that would be contributing to the GHGA Project. This classification was based on the responsibilities that the institution would have in operating the GHGA Data Infrastructure and how they will be contributing to GHGA. As for the data types, three types of institutions were defined (see Table 2).

GHGA Central	As GHGA is not a legal entity, it is not possible for GHGA to enter into any form of legal agreement. It was therefore decided that one of the institutions involved should represent GHGA and act as a legal entity on behalf of the GHGA Project. This key position is referred to as GHGA Central and is currently being fulfilled by DKFZ since it is also the coordinating institution for the GHGA Consortium. DKFZ has the responsibility for signing contracts with external parties on behalf of GHGA.
GHGA Data Hubs	Institutions that are part of the GHGA Project that will be providing storage and archiving services for Research Data and Personal Metadata are referred to as GHGA Data Hubs.

[5] M. D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," Scientific Data, vol. 3, no. 1, Art. no. 1, Mar. 2016, doi: 10.1038/sdata.2016.18.

	Staff based at the GHGA Data Hubs who are working on behalf of GHGA will also contribute significantly to supporting users of GHGA via the Helpdesk.
GHGA Institutions	Any institution that is part of the GHGA Project but who are not processing Research Data and Personal Metadata, but who are contributing to the GHGA Project in some capacity, are referred to as GHGA Institutions. For example, GHGA Institutions may be contributing to outreach or communication activities.

Table 2: Classification of institutions within GHGA

c. Roles according to the GDPR

Having defined the data types to be processed and classifying the institutions that are part of the GHGA Project, we were able to define the roles according to the GDPR that each type of institution would have for each data type (Table 3).

The GDPR considers two roles: Data Controllers, who define the essential means and purposes of the processing of personal data, and Processors, who process personal data under the instruction of a Data Controller. It is also possible for Processors to engage other parties as Processors to support them fulfil the instructions given to them by a Data Controller; within GHGA we refer to such parties as Sub-processors so that the hierarchy of responsibility from Data Controller to Processor to Sub-processor is more clearly understood.

Data Type	Data Controller	Data Processor	Data Sub-processor
Administrative Data	GHGA Central (jointly) GHGA Data Hubs (jointly) GHGA Institutions (jointly) *	-	-
Non-Personal Metadata	Not applicable as this form of Metadata are non-personal.		
Personal Metadata	Data Submitters / Data Controllers†	GHGA Central	GHGA Data Hubs
Research Data	Data Submitters / Data Controllers†	GHGA Central	GHGA Data Hubs

* Following the GDPR principle of data minimisation, only certain forms of Administrative Data are jointly-controlled by GHGA Institutions. GHGA Institutions do not have access to certain types of Administrative Data as they do not need to process it in order to fulfil their obligations within the GHGA Project.

† The term Data Submitter is used within the GHGA project to refer to the natural persons who archive Research Data and Personal Metadata with GHGA. This is usually done on behalf of an institution, a legal person, that is the Data Controller for the Research Data and the Personal Metadata

Table 3: Roles according to GDPR with relation to the data types processed in GHGA

It should be noted that due to the structure of the GHGA Consortium, the institutions operating the GHGA Data Hubs will often also be institutions who submit data into GHGA. In this case they will have a dual role as Data (sub-) Processor and Data Controller.

3. Objectives

When designing our legal concept, in addition to our obligations under data protection law, there were five other objectives that we wanted our concept to support:

i. GHGA *must* be able to take in Research Data and Personal Metadata and store it securely at *any* of the Data Hubs.
ii. *Only* Non-Personal Metadata can be published freely on the GHGA Data Portal.
iii. Institutions operating GHGA Data Hubs *must be able* to share personal Administrative Data.
iv. Institutions not operating GHGA Data Hubs *should only have* access to certain types of personal Administrative Data.
v. Data Controllers *should only* have to sign one contract to submit data to GHGA even though GHGA is not a legal entity.

4. Consortium Structure

To accommodate for the different objectives of the GHGA Project a concentrically organisational structure has been formed (see Figure 1), consisting of two groupings:

(1) The GHGA Consortium is formed by all GHGA Institutions (N=19), as regulated by the *GHGA Cooperation Contract*.
(2) As a subset of (1), institutions providing data infrastructure (storage and compute) act as GHGA Data Hubs (N=7) within the GHGA Operations Consortium. The relationship is regulated by the *GHGA Data Hub Cooperation Contract* and in addition through separate *Central-to-Data Hub Bilateral Contracts*.

DKFZ acts as the coordinating institution for both the GHGA Consortium and for the GHGA Operations Consortium in its role as GHGA Central.

In the following we describe the developed contractual agreements in further detail.

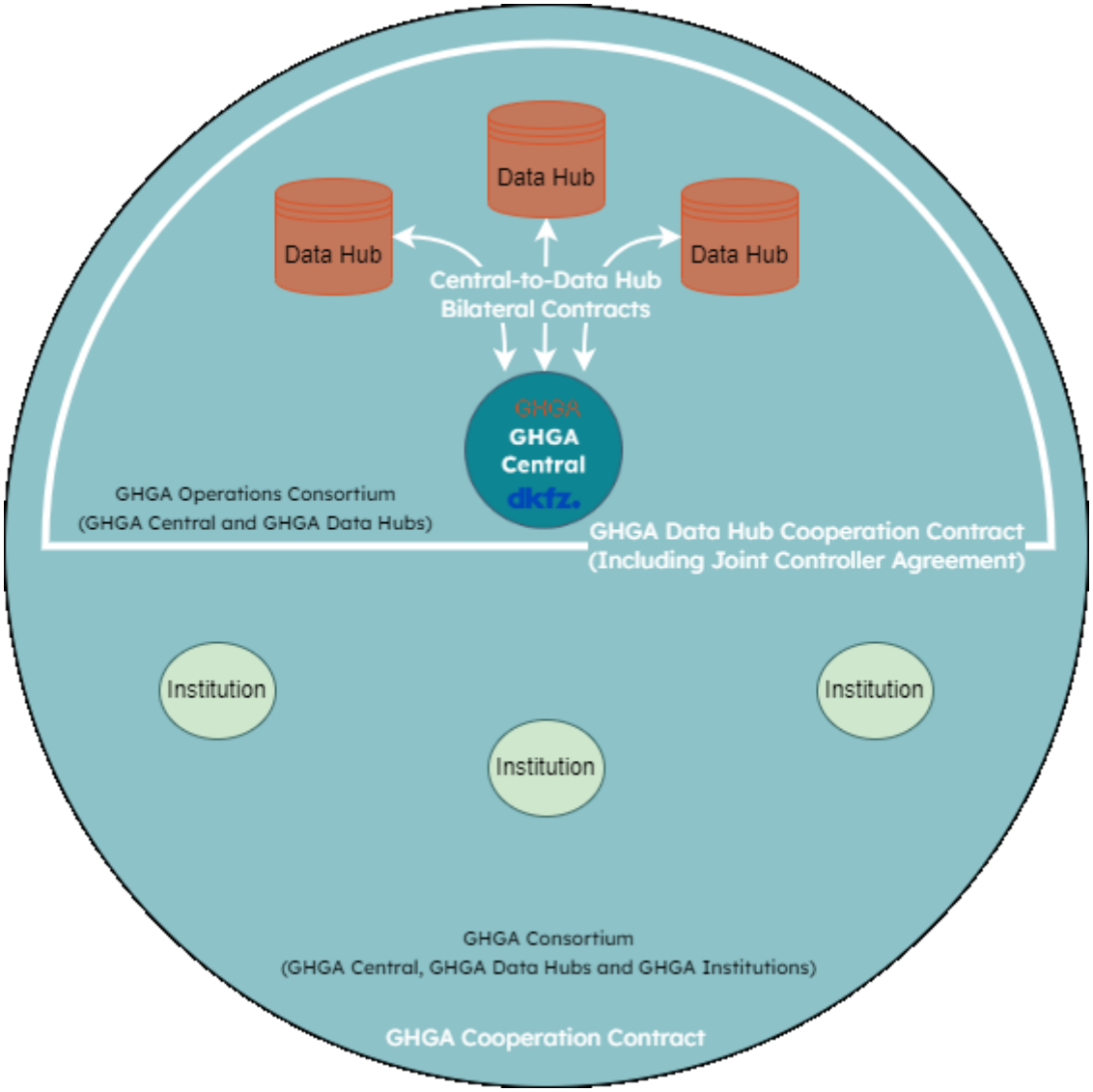

Figure 1 - Structure of the GHGA Project

The broadest grouping is the GHGA Consortium, which comprises of GHGA Central, all of the GHGA Data Hubs, and all of the GHGA Institutions. The relationship between these parties is defined in the GHGA Cooperation Contract which was agreed at the beginning of the GHGA Project. This contract defines the governance of the GHGA Consortium, with a particular focus upon how the parties will work together with respect to project governance as well as internal reporting. It also covers the distribution of funds received from the DFG within the GHGA Project. Any institution wishing to contribute formally to the development of GHGA is required to sign the GHGA Cooperation Contract. It should be noted that the services developed and operated by GHGA will be available to the German research community as a whole and is not limited to an integration in the GHGA Consortium.

The GHGA Operations Consortium is a subset of the GHGA Consortium, and is used to distinguish GHGA Central and the GHGA Data Hubs from other institutions involved in the project (Figure 1, upper half). The distinction between the GHGA Consortium and the GHGA Operations Consortium was considered necessary due to the additional responsibilities that these institutions have within the GHGA Project, particularly with regards to data protection. In addition to agreeing the aforementioned GHGA Cooperation Contract, institutions wishing to operate as a GHGA Data Hub are required to sign the GHGA Data Hub Cooperation Contract, which sets out the governance of matters within the GHGA

Operations Consortium. This covers topics such as intellectual property, liability, decision-making, and the on- and off-boarding process of GHGA Data Hubs.

A separate Central-to-Data Hub Bilateral Contract is used to govern the relationship between GHGA Central (represented by the DKFZ) and each Data Hub. This contract describes how the institution wishing to operate a GHGA Data Hub will do so. It also regulates the requirements and assurances GHGA Data Hubs need to make towards GHGA Central so that they can act as a sub-processor for GHGA Central. Each institution that is contributing to the implementation of the GHGA Data Infrastructure is doing so based on their pre-existing systems which are not standardised across the GHGA Operations Consortium. This was a necessary pre-requisite in the creation of GHGA as the NFDI funding currently does not fund investments into (IT) infrastructures. As such, it was not possible, or desirable, to have a standard model for how a GHGA Data Hub should operate. The flexibility of separate bilateral contracts enables each institution to outline how they will fulfil the requirements of being a GHGA Data Hub based on their local infrastructure and resources. This flexibility may enable a wider range of institutions to consider joining the GHGA Operations Consortium as a GHGA Data Hub.

5. Contractual Relationship with GHGA Users

a. Data Submitters

Data Submitters/Data Controllers that wish to deposit Research Data and Personal Metadata with GHGA are required to agree a Data Processing Contract (DPC) with GHGA Central. This contract sets out the Controller to Processor relationship between the Data Submitter/Data Controller and GHGA Central. With regards to the processing of the processing of Research Data and Personal Metadata, the Data Processing Contract utilises the Standard Contractual Clauses template for Art. 28 (7) GDPR published by the European Commission. The Data Processing Contract also defines the processing of Non-personal Metadata that is to be published via the GHGA Data Portal, and the personal Administrative Data of the Data Submitters that is required for GHGA Central to provide an ongoing service to them.

The DPC permits GHGA Central to process Research Data and Personal Metadata only in order to offer long-term storage of it and for its dissemination to approved Data Requesters. The Data Controller remains entirely responsible for approving requests to access the Research Data and Personal Metadata that they control. Through the DPC, the Data Controller also authorises GHGA Central to use the GHGA Data Hubs as sub-processors to support the processing of Research Data and Personal Metadata.

In the initial 'Catalog' phase of GHGA, Data Submitters may choose to submit only Non-personal Metadata without any associated Research Data and Personal Metadata. To do so, a Metadata Processing Contract is agreed between the Data Controller and GHGA Central; the part of which describes the responsibility of the Data Submitter to ensure that the metadata submitted to GHGA is non-personal according to the principles of Art. 4 (1) GDPR in conjunction with Recital 26, sentence 3-4 GDPR.

b. Data Requesters

Researchers interested in accessing Research Data archived with GHGA are able to contact the relevant Data Controller via the GHGA Data Portal. The Data Controller will then agree a Data Transfer Agreement with the Data Requester, and instruct GHGA to share the Research Data and Personal Metadata with the approved Data Requester. It is the responsibility of the Data Controller to ensure

that there is a suitable legal framework for the transfer to take place, including whether the conditions required for any international transfers have been met.

6. Data Protection

Data protection is a key aspect of the work of GHGA and how data can be processed safely has been at the core of the development of the legal concept. As previously mentioned, the different data types processed by GHGA have differing protection needs and we have tailored our approach to reflect this heterogeneity. We will now describe in turn for each data type, the data protection aspects of the GHGA Project,. A fuller picture of data protection across the GHGA Project can be found in the GHGA Data Protection Framework[6].

a. Administrative Data

The institutions which are part of the GHGA Operations Consortium jointly controls the personal Administrative Data that is processed within the GHGA Project. Joint controllership was considered to be the most appropriate structure for the GHGA Operations Consortium as the institutions within have jointly decided the essential means and purposes of processing of the Administrative Data.

The joint controllership of the Administrative Data is governed by the *GHGA Joint Controller Agreement* which is an annex to the GHGA Data Hub Cooperation Contract, that all institutions wishing to operate a GHGA Data Hub are required to sign. This Agreement has been formulated to be in accordance with Art. 26 GDPR.

Institutions that are not part of the GHGA Operations Consortium were not originally considered to be joint controllers of the personal Administrative Data. However, through the course of GHGA's work thus far, it has become apparent that GHGA Institutions do need to contribute to defining the essential means and purposes for processing certain types of personal Administrative Data. For example, the GHGA Institutions will be contributing to outreach and communication efforts, but are not currently permitted to process personal Administrative Data that is generated through those efforts. For this reason, we have begun to develop a Joint Controller Agreement that covers all GHGA Institutions that are part of the GHGA Consortium.

GHGA Institutions wishing to process personal Administrative Data are required to confirm that they have adequate protections in place before they process the data. This includes both the implementation and then the documentation of a variety of measures. For example, the processing activity must be recorded, in accordance with Art. 30 GDPR and supported by appropriate Technical and Organisational Measures (TOMs). These TOMs are supported by documents such as an Authorisation Concept, a Data Deletion Concept, a Data Breach Procedure, and a Data Subjects' Right Procedure, developed by GHGA Central and adapted at the Institutions, which are available upon request. These documents have been implemented through documented Standard Operating Procedures for staff to follow.

A threshold analysis, based on a DSFA template developed by Matthias Struck based on official recommendations, was performed to ascertain if a data protection impact assessment (DPIA) was required for the processing of Administrative Data; the assessment concluded that this was not required.

[6] Available at: https://www.ghga.de/resources/dataprotection

b. Non-Personal Metadata

Non-Personal Metadata is processed by GHGA Central and published via the GHGA Data Portal to aid in the discovery of Research Data. This processing is authorised by the Data Controller of the Research Data it describes under either a Metadata Processing Contract or under a Data Processing Contract. These contracts set out how the Non-Personal Metadata is to be used, and importantly that it is the responsibility of the Data Controller to ensure that the Metadata submitted to GHGA Central for this purpose are non-personal.

During the development of the GHGA Catalog, a Privacy Impact Assessment was performed to ascertain if we were correct to consider the Metadata publicly shared by GHGA as non-personal. This PIA focused on the risk that a Data Subject could be identified from the metadata that will be made public via the GHGA Data Portal.

The metadata schema developed by GHGA[7] does not ask for any direct identifiers and as such any possible route to identification would rely on the capability to link the Non-Personal Metadata to other data sources. One potential avenue to do so identified in the course of the PIA, would be by combining pseudonyms (for example, a sample ID used as a unique reference for a Data Subject) with matching files that contain both the pseudonyms and directly identifiable personal data. Whilst this is a non-zero risk, adequately implemented Technical and Organisational Measures protecting such matching files make such a scenario unlikely. Parties submitting Non-Personal Metadata are therefore asked to confirm within the contract they agree with GHGA Central, that adequate protections are in place and other potential sources of disclosure, such as breaches of k-anonymity, have been addressed prior to submission. This was done to be consistent with Recital 26 of the General Data Protection Regulation.

As non-personal data is outside of the scope of the GDPR, no further data protection document was required.

c. Research Data and Personal Metadata

GHGA Central is considered to be a Data Processor for the Research Data and Personal Metadata that is submitted to GHGA, with the GHGA Data Hubs acting as sub-processors supporting GHGA Central to fulfil the Data Controller's instructions. The Data Controller remains in full control of the Research Data and Personal Metadata after it has been submitted to GHGA. These relationships are described in two contracts.

When Research Data and Personal Metadata are submitted to GHGA, a DPC is agreed between GHGA Central and the Data Controller. This Contract forms the basis of the instructions given by the Data Controller to GHGA Central, and through which GHGA Central operates as a Data Processor.

In addition to describing how a Data Hub will operate, the Central-to-Data Hub Bilateral Contract between GHGA Central and each GHGA Data Hub formalises a processor to sub-processor relationship for the Research Data and Personal Metadata processed by GHGA. The Central-to-Data Hub Bilateral Contract confers the same responsibilities and obligations on the GHGA Data Hub towards GHGA Central and the Data Controller that GHGA Central has, through the Data Processing Contract, towards the Data Controller. Similarly to the Data Processing Contract, Standard Contractual Clauses have been used to define the data protection obligations of the GHGA Data Hub.

As for personal Administrative Data, GHGA Central and the GHGA Data Hubs are required to document and implement appropriate protections. As before, the processing activity must be recorded, in accordance with Art. 30 GDPR and supported by appropriate Technical and Organisational Measures

[7] https://github.com/ghga-de/ghga-metadata-schema

(TOMs). These TOMs are supported by documents such as an Authorisation Concept, a Data Deletion Concept, a Data Breach Procedure, and a Data Subjects' Right Procedure which are available upon request. These documents have been implemented through documented Standard Operating Procedures for staff to follow.

As a Data Processor and Data Sub-processors, GHGA Central and the GHGA Data Hubs are not obliged to produce a DPIA. However, the submission of large volumes of genetic data into an archive for resharing is likely to meet the necessity threshold for the Data Controller to require a DPIA. In order to make this process easier for Data Controllers that wish to support data to GHGA, GHGA is finalising a shortened risk assessment and report to document the risks that may exist within GHGA's processing scope. Along with the risks, our implemented mitigations are described through TOMs and within other related documents are also stated so that Data Controllers can assess whether they wish to share data via GHGA. This work will be made available to Data Controllers so that they can also incorporate it into their existing documents if they decide that GHGA Central is an appropriate processor for the data.

7. Core functions offered by GHGA

As a data archive, the core function GHGA offers to the research community is the ability to submit Research Data, Personal Metadata, and Non-Personal Metadata that they have generated so that it is appropriately archived and may be used by others. We have therefore sought to make the processes of submitting and accessing data as efficient as possible whilst being mindful of the data protection needs of the Data Subjects. These processes are described below.

a. Data Submission

i. A Data Submitter (the party that acts on behalf of a Data Controller) expresses to GHGA that they would like to submit data. A Data Steward advises the Data Submitter with regards to the process, and provides them with a DPC to be completed by the Data Controller.
ii. The Data Controller signs the DPC and returns it to GHGA Central.
iii. The GHGA Operations Consortium decides which GHGA Data Hub will be responsible for processing the data. This decision is taken on a number of factors including the balancing of resources across the GHGA Operations Consortium and whether there are any specific legal requirements that may impact the choice of GHGA Data Hub.
iv. Once GHGA Central has countersigned the DPC, the Data Submitter will be invited to submit Non-Personal Metadata with the support of a Data Steward from the designated GHGA Data Hub.
v. The Non-Personal Metadata are validated upon receival to check that the schema has been completed as required. After successful validation, the Data Submitter is invited to submit Research Data and Personal Metadata.
vi. The Data Submitter uses the command line interface tool developed by GHGA to submit encrypted Research Data and Personal Metadata. The data is automatically validated by the tool to confirm that the upload has been successfully completed. The Research Data and Personal Metadata are re-encrypted using a different key and stored in the secure storage area.
vii. The Data Submitter has the opportunity to approve that the submission has been completed before the Non-Personal Metadata are made public via the GHGA Data Portal.

b. Data Access

i. A Data Requester uses the Non-Personal Metadata within the GHGA Data Portal to identify a previously submitted dataset that would be suitable for their needs.
ii. Having located a suitable dataset, the Data Requester logs into the GHGA Data Portal so that they can send a notification to the Data Controller that they wish to request access to their data.
iii. The Data Controller and Data Requester agree a Data Transfer Agreement or similar legal contract by which the data can be transferred to the Data Requester. Usually, the Data Controller will utilise a Data Access Committee or comparable entity to review applications to access data, but as the Data Controller, they can define the process however they wish. GHGA has no role in this decision at all in the role of a Data Processor.
iv. Having approved access to the data, the Data Controller notifies GHGA via the GHGA Data Portal that the data can be shared with the Data Requester. The GHGA Data Portal re-encrypts a copy of the data using a public key shared by the Data Requester.
v. The Data Requester can now log into the GHGA Data Portal and download a copy of the data they have been approved to access. As it is encrypted using a key that they have provided, they are able to decrypt the data locally once it has downloaded. After a period of time, the Data Requester is no longer able to download the data without re-approval.

8. Conclusion and Outlook

The legal concept for the GHGA Project has had to account for a number of challenges as well as our own ambitions to create a service which is both beneficial to the research community and provides Data Subjects with the level of protection they would expect. We have developed a framework that adequately addresses the decentralised structure of the GHGA Project without compromising on the level of protection offered.

Returning to our ambitions we can see that this legal concept delivers on our goals:

i. GHGA *must* be able to take in Research Data and Personal Metadata and store it securely at *any* of the GHGA Data Hubs.

   The Data Processing Contract enables Data Controllers to submit Research Data and Personal Metadata to GHGA Central. It also gives GHGA Central the ability to utilise any suitable sub-processor to support this task.

ii. *Only* Non-Personal Metadata *can be* published on the GHGA website.

   The metadata schema developed by GHGA takes into account that Metadata can also be personal data and has been designed such that Metadata which are likely to be personal are not included.

   Our PIA work demonstrates that with the appropriate protections in place, the risk that the Non-Personal Metadata could be used in a disclosive manner is very low and it is not ‘reasonably likely’ that the means required to disclose the identity of a Data Subject would be used.

The Data Processing Contract and Metadata Processing Contract set out the Data Controller's responsibility to only submit Non-Personal Metadata for inclusion in the GHGA Data Portal.

iii. Institutions operating GHGA Data Hubs *must be able* to share personal Administrative Data.

The Joint Controller Agreements used by GHGA enable us to share personal Administrative Data.

This sharing is supported by requirements around the implementation of appropriate protections and documentation.

iv. Institutions not operating GHGA Data Hubs *should only have* access to certain types of personal Administrative Data.

The Joint Controller Agreements are tailored such that institutions do not have more access than they require. This is one of the benefits of classifying the institutions involved in the GHGA Project, in part, by the data processing they are required to do.

Even though the scope of their processing is reduced, GHGA Institutions are still required to have appropriate protections in place for the personal Administrative Data they process.

v. Data Controllers *should only* have to sign one contract to submit data to GHGA even though GHGA is not a legal entity.

Through the use of the Data Processing Contract and the Central-to-Data Hub Bilateral Contracts, GHGA Central is able to define which GHGA Data Hub(s) are to process the Research Data and Personal Metadata that is submitted. This means that Data Controllers do not need to sign contracts with multiple institutions to submit data with GHGA, thereby reducing the administrative burden to them.

We believe, therefore, that the legal concept we have developed for GHGA achieves the aims of the project and meets our obligations under data protection legislation. That is not however to say that we are not continuing to develop and improve our approach. In the future, GHGA will be exploring how we can provide researchers with an environment within which they can work on the data they are approved to access, thereby avoiding the requirement to download large Omics Data sets. This will speed up their ability to begin their research and may also help to improve data protection standards by limiting the settings in which sensitive data are processed. Although this will not fundamentally change our structure, it will place an additional burden upon GHGA to ensure that the environment provided can achieve the greater security hoped for whilst remaining useful.

It is also important to remember that data protection continues to evolve, both in terms of the threats faced but also in the sense that organisations should always strive to improve their processes and not consider their data protection obligations to be 'done'. As we face up to this ongoing challenge, we aim to continue ensuring that GHGA is transparent to stakeholders about what we do and how we do it, we are reflective of our work and seek to improve, and that we have respect for the Data Subjects

who have made their data available to the research community without whose trust none of this would be possible.

9. References

External References

BfDI, Tätigkeitsbericht für den Datenschutz und die Informationsfreiheit 2021, https://www.bfdi.bund.de/SharedDocs/Downloads/DE/Taetigkeitsberichte/30TB_21.html, last accessed 19.09.2023

M. D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," Scientific Data, vol. 3, no. 1, Art. no. 1, Mar. 2016, doi: 10.1038/sdata.2016.18.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

Publicly-available GHGA References

GHGA Data Protection Framework

GHGA Data Processing Contract

GHGA Metadata Processing Contract

(The documents above can be found at https://www.ghga.de/resources/dataprotection).

10. Acknowledgement

GHGA is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of the National Research Data Infrastructure Germany (NFDI) (www.ghga.de, Grant Number 441914366 (NFDI 1/1)).

11. Glossary

Administrative Data: Data which are generated through the operation of GHGA Data Infrastructure. This may include personal data which is directly identifying, such as names and email addresses which are used to communicate with, and support, service users. It may also include personal data and business data which are used internally by staff working on behalf of GHGA Central or GHGA Data Hubs. Personal administrative data is jointly controlled by the GHGA Operations Consortium members according to the Joint Controller Contract.

Central-to-Data Hub Bilateral Contract: Agreement between GHGA Central and a GHGA Data Hub. Based on the GHGA Data Hub Cooperation Contract it regulates the relationship between GHGA Central and the Data Hub and the corresponding rights and responsibilities in full detail. In particular it defines the processor to sub-processor relationship between GHGA Central and the individual Data Hub. It also enables adjustments with respect to local infrastructures and federal data protection law where required.

Data Controller: Pursuant to Art. 4 [7] GDPR, Data Controller "means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law".

Data Processing Contract: Bilateral agreement signed by GHGA Central and a Research Data Controller who wishes to deposit data in the archive. The agreement regulates the rights and duties of the controller and GHGA Central in processing the deposited data.

Data Processor: Pursuant to Art. 4 [8] GDPR, Data Processor "means a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller".

Data Subject: Pursuant to Art. 4 [1] GDPR, the Data Subject is "an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person".

Data Submitter: Users who are depositing data with GHGA Central (and includes the Data Controller, if not the same person(s), as defined in the Data Processing Contract). The Data Controllers will regularly operate one or several Data Access Committees (DACs) facilitating decision-making regarding access to the Research Data shared via the GHGA Data Infrastructure.

EGA: The European Genome-Phenome Archive (EGA) provides archiving and sharing support for personally identifiable genetic and phenotypic data. It is operated jointly by the European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI), an intergovernmental organization, and the Centre for Genomic Regulation in Barcelona (CRG). The EGA is developing a federated model through which national nodes will take on the archiving of genomic-phenomic data for their country; this federated network of institutions will be referred to as the federated EGA (fEGA). EGA will take over the coordinating function as operator of EGA-Central, GHGA is planned as the national node for Germany.

GHGA Central: Deutsches Krebsforschungszentrum (DKFZ) shall be the legal entity responsible for GHGA Data Infrastructure. The responsibilities are outlined in the GHGA Data Hub Cooperation Contract and are to be fulfilled in accordance with the regulations outlined in this contract. These responsibilities will include taking on the role of a Data Processor for the Research Data submitted to the GHGA Data Infrastructure. Within this agreement, DKFZ will be referred to as GHGA Central. DKFZ will also operate a GHGA Data Hub.

GHGA Consortium: The organizations that have signed the GHGA Cooperation Contract in December 2020 and receive (some of) their funding as part of the Nationale Forschungsdateninfrastruktur (NFDI) from the German Research Foundation (DFG).

GHGA Cooperation Contract: Regulates the organization of the GHGA Consortium to carry out the GHGA Project. Does not regulate the exchange of Person-related Data.

GHGA Data Hub: A GHGA Consortium Member that stores and processes Research Data for the purposes of archiving and secondary analysis for scientific research purposes on behalf of GHGA Central. The responsibilities are set out in the GHGA Data Hub Cooperation Contract.

GHGA Data Hub Cooperation Contract: This contract is agreed by GHGA Central and GHGA Operations Consortium members that operate, or wish to operate, a GHGA Data Hub. The contract sets out the structure of the Operations Consortium, including a definition of the roles and responsibilities of members as well as the data governance framework. The appendices define a number of common standards that will be utilised by GHGA, including the Joint Controller Agreement for Personal Administrative Data.

GHGA Data Infrastructure: GHGA Central, together with the GHGA Operations Consortium, will operate an IT infrastructure to enable FAIR (findable, accessible, interoperable and reusable) data sharing of human Omics Data for scientific research purposes as defined by the Joint Controller Contract and the Bilateral Contracts.

GHGA Institution: Any institution that is contributing to the GHGA Project, and has signed the GHGA Cooperation Contact, but serving as neither GHGA Central or a GHGA Data Hub is referred to as a GHGA Institution.

GHGA Operations Consortium: GHGA Central and the GHGA Data Hubs. The GHGA Operations Consortium Board will be responsible for decision-making within the GHGA Operations Consortium.

GHGA Project: The Project refers to the overall GHGA structure. This includes GHGA Central, the GHGA Consortium, and GHGA Partners who are working together to make genomic-phenomic data available.

Metadata: Information that describes or annotates Research Data to aid understanding or to describe the relationship between data items. It may be personal or non-personal.

Metadata Processing Contract: This agreement covers the deposition and sharing of Metadata in the GHGA Metadata Catalog during the GHGA Catalog phase of the project. During GHGA Catalog, only Non-Personal Metadata describing Research Data will be processed and shared; the corresponding Research Data and Personal Metadata remain with the Data Controller and are not stored by the GHGA Operations Consortium.

Non-Personal Metadata: Information that describes or annotates Research Data to facilitate its interpretation or to describe the relationship between data elements. For example, the name of the instrument used to generate the data or information defining a group of Data Subjects. Non-Personal Metadata will be available for public search online within the GHGA Data Infrastructure.

Omics Data: The Research Data collected as part of omics-based research. This research focuses on collecting information regarding the entire set of certain molecules in a sample. Within the context of GHGA, Omics Data linked to genetic information of an individual are of particular interest since in many cases Omics Data would fall under the definition of personal data in Art. 4 Nr. 1 GDPR. Types of Omics Data considered in GHGA are e.g.: genomics – the entirety of the hereditary information in a sample's DNA; transcriptomics – the entirety of the RNA transcribed from DNA; epigenomics – information on epigenetic modifications of the genetic materials.

Personal Metadata: Information that describes or annotates Research Data to facilitate its interpretation or to describe the relationship between data elements. For example, demographic data or information on the ancestry of the Data Subjects of the Research Data that allow conclusions to be drawn about individuals and thus fall within the scope of the GDPR, Art. 4 No. 1 GDPR. Personal Metadata are made available to the Data Requester together with the Research Data only under controlled access after release by the Data Submitter.

Research Data: Omics or other forms of genetic (Art. 4 Nr. 13 GDPR) and health data (Art. 4 Nr. 15 GDPR) that are used for scientific research purposes. This is considered to be special category personal data under Art. 9 (1) in conjunction with Art. 4 Nr. 1 GDPR.