European Genomic Data Infrastructure

Deliverable D6.1

Draft data management policy published including ELSI best practice

Project Title Grant agreement no	**Genomic Data Infrastructure** Grant agreement 101081813		
Project Acronym (EC Call)	GDI		
WP No & Title	WP6: Data Management		
WP Leaders	Rob Hooft (21. HRI)		
Deliverable Lead Beneficiary	8. VIB		
Contractual delivery date	30/04/2023	**Actual delivery date**	22/05/2023
Delayed	Yes, due to review.		
Partner(s) contributing to deliverable	VIB, HRI, EMBL-EBI		
Authors	Dilza Campos (VIB)		
Contributors	Frederik Coppens (8. VIB), Rob Hooft (21. HRI), Jeroen Belien (21. HRI)		
Acknowledgements			
Reviewers	Thomas Keane (1. EMBL-EBI), Mallory Freeberg (1. EMBL-EBI)		

Log of changes

Date	Mvm	Who	Description
02/05/2023	0v1	Dilza Campos (VIB)	First draft of complete document.

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

09/05/2023	0v3	Nikki Coutts (ELIXIR Hub)	Copy circulated to the GDI-MB and 1+MG Stakeholders for review
xx/05/2023	1v0		Final version submitted to the EC portal

Table of contents

Contents

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

1. Executive Summary

Defining the data management policy is an important step to ensure genomic data in GDI is protected from unauthorised access, use, or disclosure, and that it complies to local legal and regulatory requirements. This is the draft version of the data management policy and it will evolve as the discussions around data governance, ELSI requirements and data management plans are taken into account in WP6 and in GDI as a whole. This report describes the main elements of the data management plan that need to be taken into account and as such this should be used as a guide for each node in GDI. Over time, we will describe good common practices for each of these elements, and individual nodes will be able to add any deviations from these common practices to represent the way data management is actually performed. We plan to collect information using an instance of the Data Stewardship Wizard that will be deployed with a GDI-specific knowledge model that will be continuously updated to encode the common practices, provide integrations to more in depth information, and will have filled questionnaires for each node to document their choices.

2. Contribution towards project outcomes

With this deliverable, the project has reached or the deliverable has contributed to the following project outcomes:

	Contributed
Outcome 1 Secure federated infrastructure and data governance needed to enable sustainable and secure cross border linkage of genomic data sets in compliance with the relevant and agreed legal, ethical, quality and interoperability requirements and standards based on the progress achieved by the 1+MG initiative.	Yes

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

Outcome 2 Platform performing distributed analysis of genetic/genomic data and any linked clinical/phenotypic information; it should be based on the principle of federated access to data sources, include a federated/multi party authorisation and authentication system, and enable application of appropriate secure multi-party and/or high-end computing, AI and simulation techniques and resources.	No
Outcome 3 Clear description of the roles and responsibilities related to personal data and privacy protection, for humans and computers, applicable during project lifetime and after its finalisation.	Yes
Outcome 4 Business model including an uptake strategy explaining the motivation, patient incentives and conditions for all stakeholders at the different levels (national, European, global) to support the GDI towards its sustainability, including data controllers, patients, citizens, data users, service providers (e.g., IT and biotech companies), healthcare systems and public authorities at large.	No
Outcome 5 Sustained coordination mechanism for the GDI and for the GoE multi-country project launched in the context of the 1+MG initiative.	No
Outcome 6 Communication strategy – to be designed and implemented at the European and national levels.	No
Outcome 7 Capacity building measures necessary to ensure the establishment, sustainable operation, and successful uptake of the infrastructure.	Yes

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

Outcome 8 Financial support to the relevant stakeholders to enable extension, upgrade, creation and/or physical connection of further data sources beyond the project consortium or to implement the communication strategy and for capacity-building.	No

3. Methods

The scope of this task is to draft the data management plan to support the nodes with overall data management including best practice with ELSI compliance. It interfaces with WP4 (SOPs covering areas such as data protection implementation, helpdesk and daily operations, data management, and security) from Pillar II and builds upon the work accomplished by B1MG WP2 (Ethical, Legal and Social Issues [ELSI]) as well as the 1+MG ELSI Working Group and the 1+MG Trust Framework. This draft was based on the discussions from monthly meetings with the individuals assigned as responsible for this task and the WP leads.

4. Description of work accomplished

4.1 Data management plan

A data management plan (DMP) is a document that outlines the policies and procedures for managing data throughout its lifecycle. It documents the choices made in a project or infrastructure that ensure consistency in data handling, aiming for maximal data value and impact and minimal data mishaps, while also allowing to meet ethical and legal requirements for data handling and sharing. Since the majority of research funders now require a DMP as part of the grant proposal process, there are a number of DMP models and tools available for researchers. To ensure that all nodes describe their data (and also metadata) in a standardised way, the goal of task 6.2 is to establish a DMP model that may be used by every node and their data providers. As such, this model also serves as a checklist to ensure all relevant aspects are considered for the management of sensitive data.

Another goal of WP6 is to ensure that the data used to test the nodes is representative of the data the node manages and that the data therefore, the testing of the node using the starter kit components is valid and relevant. To accomplish this objective, we will gather information from each node`s DMPs and we will report on common and deviating practices along the data lifecycle. Among the available tools and guidelines for writing DMPs, we have decided to use the Data Stewardship

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

Wizard (DSW)[1] for this task, because of its functionalities that can support the evolving nature of the DMP and the commonalities and differences between processes followed by all parties involved in the project.

This tool will allow us to establish a set of questions to accurately gather information from each node and at the same time we will be able to update those questions and answers based on feedback from data managers that complete the questionnaire as well as improving insights over time. After completing any version of the questionnaire, the tool will allow the node representative to export those answers as a complete DMP (Figure 1). The set of questions will be tailored to the specific information relevant to the project and we will focus on closed questions instead of open ones to be able to minimise the effort involved in keeping the plans updated over time, and to effectively collect information in a machine actionable way.

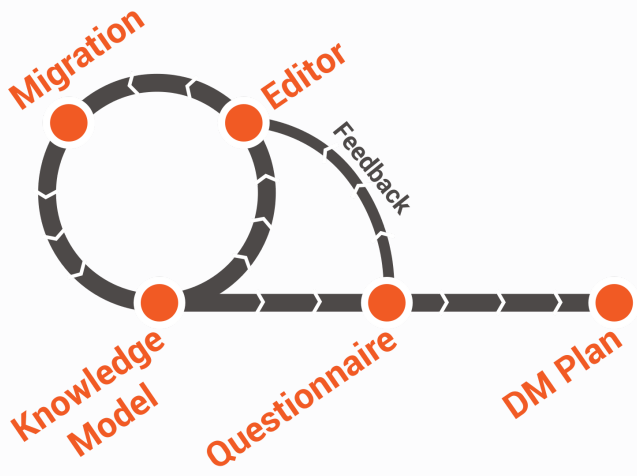

Figure 1. Overview of the proposed framework: the knowledge model provides a questionnaire that can be adapted through feedback. The feedback is incorporated into a new knowledge model that originates a new updated questionnaire. The final version of the answered questionnaire can be exported as a full data management plan.

In this work we have set up a DSW instance[2] for GDI with two generic knowledge models, namely: Common DSW Knowledge Model and Life Sciences DSW Knowledge Model. These knowledge models are in the process of being adapted and reviewed by members of the WP. The main advantage of using the preexisting generic knowledge model instead of writing a new one is that the existing ones already have links to external resources that can guide the data manager through the answering process. This enables us to provide guidance from the beginning, supporting the nodes in managing their data, while we iteratively finetune the resource with one new release mostly every six

[1] https://ds-wizard.org/
[2] https://gdi.ds-wizard.org/

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

months. The 1+MG Trust Framework (to be published) will be based on technologies that have existing integrations in DSW, enabling fast integration of this resource. For the reasons listed here, we decided to edit the Common DSW Knowledge Model for GDI and in the next sections we will briefly describe its structure and how it relates to GDI main outcomes and objectives.

4.1.1 Common life science knowledge model

In WP6 we are currently customising the Common DSW Knowledge Model, 2.4.4[3] for GDI, adding specific questions pertinent to genomic data and infrastructure. This data model is structured in seven chapters with a few hundred questions in total that are distributed unevenly across the chapters described briefly in the following subsections.

4.1.1.1 Administrative information

This chapter will capture how the node is structured and will guide it through some governance questions to support long-term sustainability of the infrastructure developed by the node. In this chapter, there will be questions regarding the description of the roles and responsibilities of those involved in managing the data, including any data management training or support that will be provided within the node. It also includes questions about how the node is structured regarding its data and infrastructure providers, how they legally interface with each other, if data is stored centrally or not, mechanisms of communicating with a Data Access Committee (DAC), among other questions.

4.1.1.2 Data reuse

This chapter focuses on questions about the mapping of the genomic datasets available to GDI. This chapter will contribute directly to the deliverable 6.2 (Report on European data/resources suitable for inclusion into the GDI) and will help map the volume of genomic data identified through the project - a proposed key indicator of GDI.

4.1.1.3 Creating and collecting data

In this chapter, there will be questions regarding the genomic and associated phenotypic data stored in the node. This chapter is the bigger one and will allow a description of the types of data that will be generated or collected, including the formats, quality of data, data organisation, identifiers, encryption, metadata generation and storage. Other topics included in this chapter are the tools, software, or hardware used by the node to collect genomic and/or clinical or phenotypic data, as well as a well-documented data provenance framework. We will also document how each node links genomic data with other sources of information, like clinical data. This chapter focused on GDI's main objective of deploying infrastructure, standards and services to an operational level.

[3] https://registry.ds-wizard.org/knowledge-models/dsw:root:2.4.4

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

4.1.1.4 Processing data

Questions about data volume, versioning, compute capacity, workflows and reproducibility will be addressed in this chapter. These questions are related to the project work towards assessing the technical readiness and interoperability of national nodes.

4.1.1.5 Interpreting data

The members of the WP have considered this chapter the least relevant to a genomic infrastructure DMP, especially in the current, early phase of the project. However, we decided to keep it on the template because of the federated analysis questions present on the questionnaire that may be relevant to the project.

4.1.1.6 Preserving data

In this chapter we will cover questions regarding cold storage, costs related to storage, backup and contingency plans, and infrastructure security policies. These questions may help the node build its business/financial plan to ensure the continued operation of the infrastructure - one of the objectives of Pillar 1.

4.1.1.7 Giving access to data

Questions about national data catalogue, how data access committees are contacted and organised inside the node will be addressed in this chapter. These questions are important to allow the federated access to genomic data envisioned by GDI.

4.2. ELSI best practices

For genomic data to be suitable for interchange, it will have to meet GDPR criteria and also broader ELSI requirements, to cover the issues that arise from genomic research, and the personal, social, and cultural factors related to the use of individuals' genomic information. We envision that the questions from the questionnaire will be able to cover these topics and will guide the respondents to look for further guidance. The questions will have guidance including links to direct the respondents to the 1+MG Trust Framework that is being developed and to be published soon. It will contain the legal and ethical requirements identified by that project on making nationally compliant datasets for pan-European access. Also, additional questions and orientation can be added to the DMP as ELSI discussions are held during the project duration and those will be embedded in the chapters described previously. Data protection by design and default principles are addressed in every chapter described here.

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

5. Results

The working group for this deliverable established effective communication and have held monthly meetings on this topic since February 2023. We have also established communication channels with other WPs within the project, as well as relevant other initiatives (e.g. 1+MG Trust Framework). We have also collectively decided on using DSW as a DMP platform and created the DSW instance for GDI[4] that will later be used by the other members of the project to accomplish the objectives described previously in this document. The general knowledge model of DSW and its functionalities can be viewed and explored by any user that registers on the website.[5]

6. Discussion

A DMP is an evolving document that can allow finer levels of granularity as the project progresses and when changes or updates occur. This is the case with the approach taken here. This draft will be updated as the discussions and the implementation of the starter kit by nodes develop through every maturity stage. We plan to release these updates on the DMP knowledge model every six months during the project. Using the proposed tool, the nodes can clone the DMP to their own DSW instances and keep using it to build their financial plan and to ensure the continued operation of the infrastructure, even after the GDI project is over. One aspect that will be further discussed is the development of a project specific DMP to describe how data is made available through the GDI user portal to the end users, the relevant technical aspects of this and the mechanisms in place to ensure that the data displayed at the portal is ELSI compliant.

At the moment of writing, this first version DMP model is being reviewed by the WP members and in the coming weeks we will proceed to the first presentation of the questionnaire to improve its content.

7. Conclusions & Impact

Defining the data management policy is of paramount importance to ensure genomic data in GDI is protected and that it complies to national and European legal and regulatory requirements. In this task we have raised awareness about the importance of establishing good data management practices from the early stages of the project and we have established a regularly meeting group to discuss and improve this draft version of the data management policy as the project evolves.

[4] https://gdi.ds-wizard.org/
[5] https://researchers.ds-wizard.org/

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.

European Genomic Data Infrastructure

8. Next steps

First adaptations of the DSW knowledge model for GDI have commenced, this work will continue until we have reached a minimal workable model encoding currently known good practice suggestions. We envision that the questionnaire will have links to the 1+MG Trust Framework standards and rules website (under construction) to offer further guidance to the nodes to ensure their data complies with ELSI requirements. Subsequently, we will ask a small set of (vanguard) nodes to fill in their first questionnaires based on the adapted knowledge model. Feedback from these first nodes will be incorporated before the first general rollout.

GDI project receives funding from the European Union's Digital Europe Programme under grant agreement number 101081813.