European Genomic Data Infrastructure

Deliverable D8.7

Report on semantic interoperability scenarios

Project Title Grant agreement no	**Genomic Data Infrastructure** Grant agreement 101081813		
Project Acronym (EC Call)	GDI		
WP No & Title	WP8: Application and Innovation Solutions		
WP Leaders	Alfonso Valencia (37. BSC) Salva Capella-Gutierrez (37. BSC) Marc Van Den Bulcke (9. SC)		
Deliverable Lead Beneficiary	34. UMCG		
Contractual delivery date	30/04/2024	**Actual delivery date**	30/05/2024
Delayed	Yes		
Partner(s) contributing to deliverable	UMCG, BSC, UT, HRI, UM		
Authors	K. Joeri van der Velde (UMCG) Morris A. Swertz (UMCG) Gerieke Been (UMCG)		
Contributors	Carles Hernandez-Ferrer (BSC) Juan Arenas (ELIXIR Hub) Rob Hooft (HRI) Laura Portell Silva (BSC) Salvador Capella (BSC)		
Acknowledgements	All members of WP8		
Reviewers	Milan Ojsteršek (UM) Priit Kleemann (UT) Jeroen Beliën (HRI)		

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

European Genomic Data Infrastructure

Log of changes

Date	Mvm	Who	Description
20/04/2024	0v1	K. Joeri van der Velde (UMCG)	First draft sent to WP8 and reviewers
28/04/2024	0v2	K. Joeri van der Velde (UMCG)	Added input and comments from WP8 members
15/05/2024	0v3	K. Joeri van der Velde (UMCG)	Updated according reviewers and sent to Coordination
15/05/2024	0V4	Mercedes Steiner (ELIXIR Hub)	Sent to the GDI Management Board for review
31/05/2024	1V0	Mercedes Steiner (ELIXIR Hub)	Final version submitted to EC Portal

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

European Genomic Data Infrastructure

Contents

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

1. Executive Summary

The Genomic Data Infrastructure (GDI) project aims to overcome or lower barriers in clinical diagnostics and treatment by facilitating access to fragmented human genomics data across Europe. The project establishes a federated, secure infrastructure, but faces interoperability challenges on all 6 layers of the refined eHealth European Interoperability Framework due to the diversity of implementations of law and regulations, various organisational setups, data sources and infrastructures.

To address this issue, a framework for solving interoperability scenarios is proposed. This framework is able to take on a variety of relevant interoperability building blocks and interoperability scenarios that are listed in this deliverable. Key framework components include shared logical models from which we derive data management systems that serve intrinsically interoperable APIs. These systems are allowed to make different model selections or use different software products while maintaining interoperability.

The main drivers for this work are Genome of Europe (GDI T7.1 and 1+MG WG12), Cancer (GDI T7.4 and 1+MG WG9), Infectious diseases (GDI T7.3 and 1+MG WG11), Rare disease (GDI T7.2 and 1+MG WG8), as well as D7.4, MS26, MS27 from WP7. The framework will support the implementation of the T8.2 semantic interoperability package, which develops and harmonizes the minimal models from the WP7 datasets, and we expect it will contribute to achieving WP3 MS7 and MS8, WP8 T8.3, and D4.3, MS11, and MS12 from WP4. Furthermore, multiple standards for data discoverability from the TEHDAS 'Recommendations to enhance interoperability' within HealthData@EU have been incorporated in the reference implementation of the framework, such as Beacon, BBMRI MIABIS and DCAT-AP. The upcoming HealthDCAT-AP standard, built on DCAT-AP and designed specifically for the HealthData@EU infrastructure by WP6 of the EHDS2 Pilot, will be implemented when released. HealthDCAT-AP will standardise the descriptions of health-related datasets in GDI to ensure interoperability with the EHDS and meet the needs of its users.

A reference implementation of this framework MOLGENIS EMX2 has resulted in a data management solution that acts as a data catalogue and as a 'local portal' towards the underlying genomics data. It features Beacon v2, FAIR Data Point, RDF, Life Science AAI, integration and synchronisation to a REMS instance for managing data access requests. The catalogue currently contains datasets like the B1MG Rare Disease Synthetic Dataset. The reference implementation is packaged in a Docker container for easy installation and testing, encouraging collaboration and feedback from all GDI partners.

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

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.00000	**Yes** (interoperability enables linkage of genomic data)
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.	**Yes** (interoperable clinical/phenotypic data can be linked and analysed)
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.	**No**
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	**No**

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

(e.g., IT and biotech companies), healthcare systems and public authorities at large.	
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** (better sustainable operation by defining what metadata needs capturing)
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**

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

3. Introduction

3.1. Background

The partners in the '1+ Million Genomes' (1+MG) initiative and its implementation project European Genomic Data Infrastructure (GDI) and Genome of Europe (GoE) aim to enable secure access to genomics and the corresponding clinical data across Europe for better research, personalised healthcare and health policy making. In addition, they participate in many other (inter)national projects that also aim to improve disease prevention, allow for more personalised treatments and support groundbreaking research. Obviously, there is a great deal of potential synergy across all of these projects. To unlock this potential, we invest in interoperability for joint discovery, queries and analyses within GDI and related projects that part of 1+MG and GDI working groups, supported and in alignment with the European Health Data Space (EHDS) and key EU-funded projects like GDI, TEHDAS, HealthData@EU and EUCAIM.

3.2. Objective

Aim of this deliverable is to propose a framework for solving interoperability scenarios within and beyond the GDI project.

Fragmented collection and analysis of human genomics including related phenotypic and clinical data represents a significant barrier for progress in clinical diagnostics, treatments and predictive medicine. The European Genomic Data Infrastructure (GDI) project is enabling access to this data across Europe by establishing a federated, sustainable and secure infrastructure to find, search and access the data. However, the data is kept in systems from different vendors, based on different interoperability building blocks, using different API specifications. Furthermore, they have been developed to be suitable for specific interoperability scenarios, use-cases and questions. The heterogeneity of the genomic data landscape is a significant challenge that requires flexible yet pragmatic solutions. This challenge is exemplified in **Figure 1** below.

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

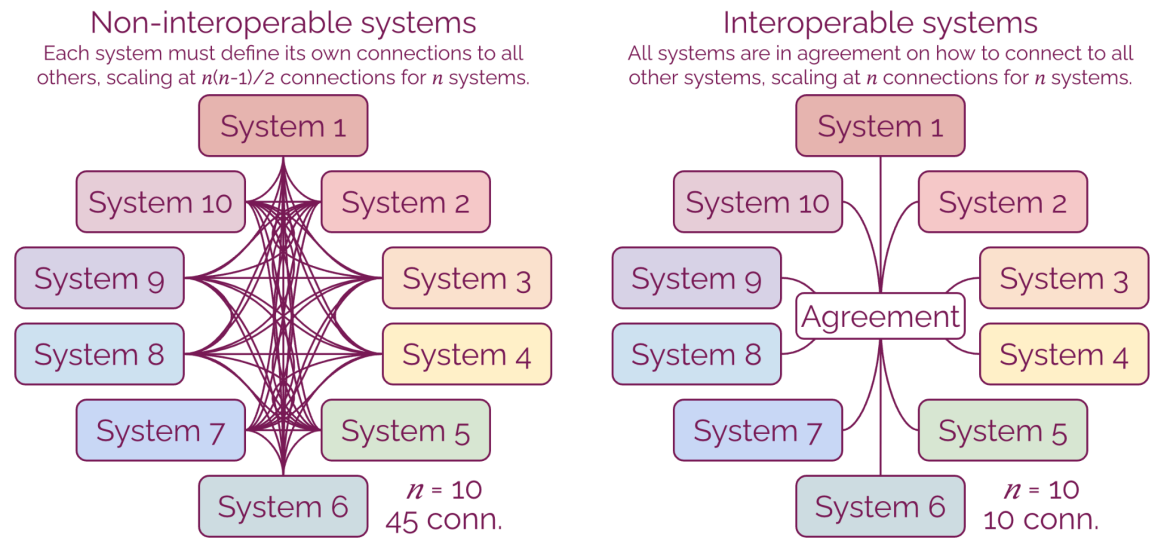

Figure 1: the connections needed for non-interoperable systems versus interoperable systems. In order for a federation of systems to be scalable, it must reach agreement on its connections.

In this deliverable, we report the development of a framework for solving interoperability scenarios within and beyond the GDI project towards establishing an interoperable European omics infrastructure. First, we will provide an overview of the framework and its components. Discussed next are the two main drivers of the framework: interoperability building blocks relevant for storing and communicating genomics data, and the interoperability scenarios that make use of these data by either asking questions with research or clinical purpose, or discovering and requesting the digital samples to do so. Finally, we will cover the current state of framework implementation as well as ongoing and future efforts. We also present tangible results produced by this framework for tackling the semantic interoperability in connection with the use cases in WP7, 1+MG/B1MG and other European projects complementary or in synergy with GDI.

3.3. Related work

This work is connected to the following work packages and deliverables as framework drivers:

- Pillar 3, WP7, GDI use cases:
  - Genome of Europe (GDI T7.1 and 1+MG WG12)
  - Rare disease (GDI T7.2 and 1+MG WG8)
  - Infectious diseases (GDI T7.3 and 1+MG WG11)
  - Cancer (GDI T7.4 and 1+MG WG9)
- Pillar 3, WP7, D7.4: Report to identify the initial set of relevant data for use cases.
- Pillar 3, WP7, MS26: Initial set of questions by the use cases.
- Pillar 3, WP7, MS27: Initial set of relevant data for use cases are identified.

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

- Pillar 3, WP8, T8.2: Semantic interoperability package, which develops and harmonizes the minimal models from the WP7 datasets.

And as development and deployment of software outputs resulting from the framework:

- Pillar 3, WP8, T8.3: Innovation in new data technologies.
- Pillar 2, WP3, MS7: Early adopter production nodes deployed with synthetic real-like data.
- Pillar 2, WP3, MS8: Incremental demonstrator
- Pillar 2, WP4, D4.3: User portal live cataloguing data within the GDI.
- Pillar 2, WP4, MS11: Development of the user portal deployed.
- Pillar 2, WP4, MS12: Production user portal deployed, with discoverability and data access functionality.

This work is also connected to the external framework drivers that are described in more detail in section 4.3 (User communities and groups).

3.4 Outline

We first engaged in desktop research and discussions with GDI partners which resulted in a proposed framework to expedite implementation of interoperability scenarios for GDI node and central services (section 4 'methods'). To evaluate the merits of this framework proposal we created a reference implementation (section 5 'description of the work accomplished). Finally, we evaluated these results, discussed how the framework can be used to develop GDI 'sunflower' model development, and provide conclusions and next steps (section 6-9). The annex provides a summary of the details of the work.

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

4. Methods & Materials

To address GDI semantic interoperability scenarios, we engaged in desktop research of previous work and a series of discussions with GDI partners to analyse interoperability requirements. Below we first summarise a framework to facilitate interoperability implementation that was proposed from these discussions (4.1). Then we provide an overview of the interoperability building blocks that have been identified to enable GDI interoperability scenarios that should feed into this framework (4.2). Subsequently we summarise the main user communities identified by GDI that should be served using these building blocks (4.3) as a basis for an analysis of the interoperability elements that will be implemented to serve interoperability needs across GDI data lifecycles. To evaluate the proposed framework we have implemented a reference implementation (section 5).

4.1. Framework for solving interoperability scenarios

In life sciences, countless interoperability building blocks have been designed such as data models, ontologies and file formats, each for a particular scope and purpose. Together with the interoperability scenarios that make use of data they capture, they are the drivers, or starting points, of the framework presented here. The interoperability building blocks and interoperability scenarios are situated at the top and top-right of **Figure 2** that provides a full framework overview.

The actual components of the framework are denoted by A-G in this figure and consist of the composite logical model (A), purposed logical model (B), physical model (C), data management system (D), request and response processing (E), Application Programming Interfaces (F), and post-system interoperability effort (G). These components are described in more detail as part of the figure and their implementation is described in later sections of this deliverable.

Lastly, the component named 'Federated query or analysis' is a technical representation of the questions originating from the interoperability scenarios. They include the full range of questions including simple discovery queries such as: 'how many females with Noonan syndrome older than 20 years' all the way to an advanced federated genome-wide association study.

The main messages that **Figure 2** conveys are:

- Heterogeneity in terms of interoperability building blocks and interoperability scenarios is welcomed and can be managed effectively, even before data management systems are developed and deployed.
- Investing together in a composite logical model once leads to a family of interoperable systems, saving each deployment a significant post-system interoperability effort.
- Interoperability between systems is not a binary property, but instead a gradient that depends on the starting commonalities of these systems.
- Systems without any starting commonality can still join a family of interoperable systems by mapping their internal data model to an API-specified model and implementing the corresponding API protocol. While this is a common practice, it should be avoided, because these mappings are highly time-consuming to create and are imperfect, resulting in data loss.

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

- Facilitate the autonomy of data holders by respecting their specific data capturing choices, most likely made with good reasons, while maintaining interoperability.

Relation to the European Interoperability Framework (EIF)

The Refinement of the eHealth European Interoperability Framework[1] (ReEIF, final version of April 28, 2015) was established by the 7th Meeting of the eHealth Network (eHN). It defines six interoperability levels: Legal and regulatory, Policy, Care process, Information, Applications, and IT Infrastructure. In the framework for solving interoperability scenarios presented here, the interoperability levels Legal and regulatory, Policy, and Care process are not in scope. However, the three levels remaining levels are closely aligned: Information is about defining and coding of information. Essentially, the presented framework specifies a flow to achieve standardized information from by applying components A through C. Second, applications are about integration in healthcare systems. While not addressed directly, the central component D can be considered an integrated component situated in between standardized information and standardized communication. Lastly, IT Infrastructure is about communication protocols. Framework components E through G provide a further breakdown and specification of this concept.

[1] https://health.ec.europa.eu/document/download/7cc56460-46ea-4fd7-9064-ac1a14a2a14e_en

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

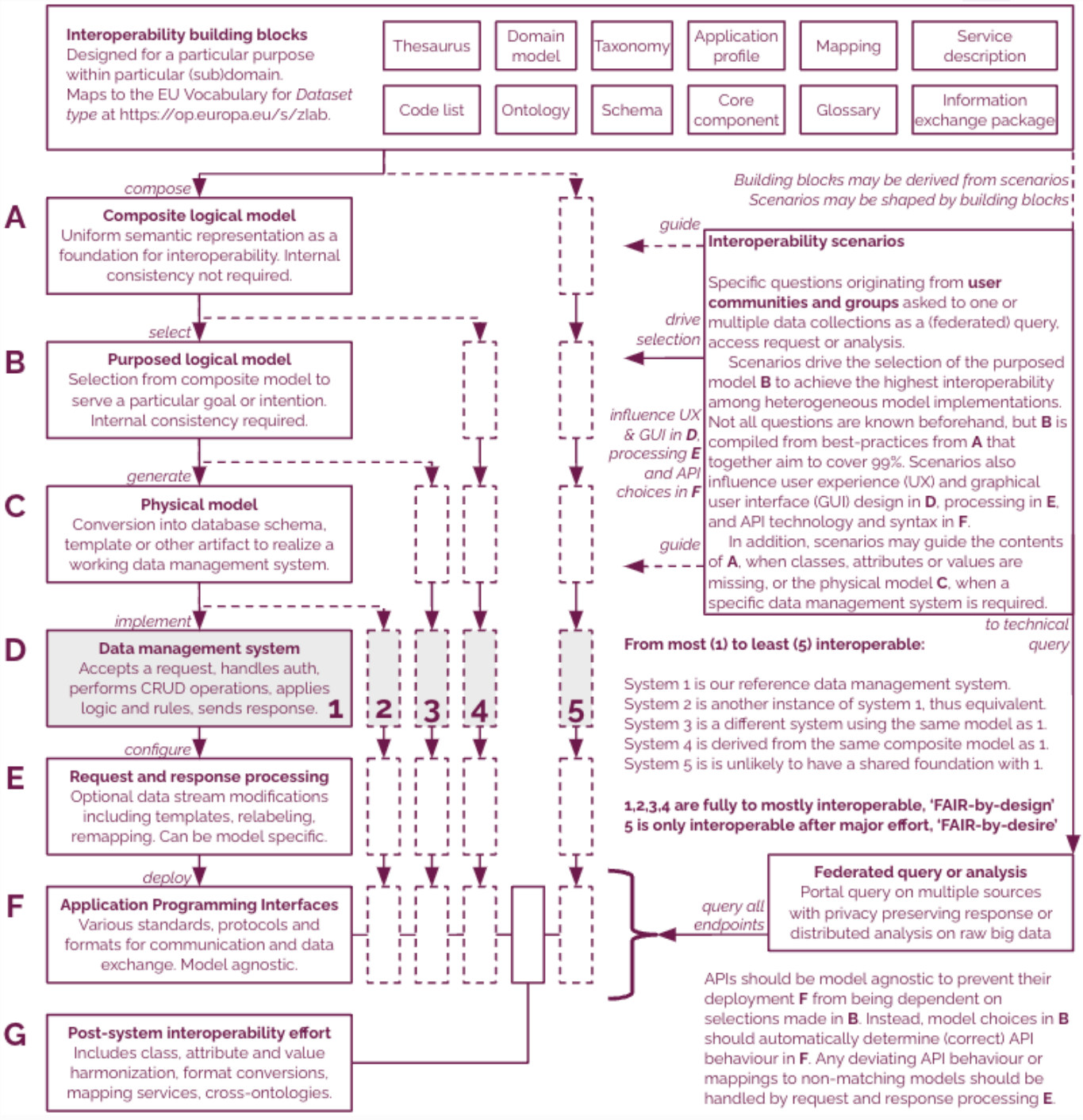

Figure 2: Proposed framework for solving interoperability scenarios.

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

4.2. Interoperability building blocks

Task 8.2, which is working on the semantic interoperability scenarios, is meeting every month to work on a mapping in the different minimal datasets that are part of GDI and 1+MG in order to find commonalities and start a cross-walk among them. The interoperability building blocks presented here are in scope of this working group to delineating different interoperability scenarios in GDI. They are essential for storing, organising and manipulating data so that it can be accessed and manipulated efficiently, as well as used in algorithms and applications. They are the driver for interoperability as shown in **Figure 2**. Here, we will briefly describe data structures that are relevant for interoperability in GDI as well as complementary projects. In this section, we will focus on 'metadata' that typically powers the data discovery journey of a user. This comprises, for instance, properties of individuals, patients, samples, analyses, and datasets as well as other layers of metadata such as terms of use and data access conditions. It also includes phenotypes (observable characteristics of an individual) and genotypes (the genetic constitution of an individual) as long as they are relevant for data discovery purposes. For instance, this may include key genomic features or aggregated genomic data such as individuals with a frameshift or stop gain mutation in a gene of interest or finding genomic variants with an allele frequency below a threshold. After data discovery and access, users can work with the complete genomics data that is typically stored in sufficiently standardized genomics data formats such as FASTQ, BAM, CRAM, VCF and gVCF.

Different types of building blocks may together contribute to interoperability. These types are explained below in **Table 1**, and will be assigned to the building blocks in this deliverable for clarity.

Building block type	Description
Application profile	In the information sciences, an application profile consists of a set of metadata elements, policies, and guidelines defined for a particular application.
Code list	A code list is a complete set of data element values of a coded simple data element.
Core component	A core component is a context-free semantic building block for creating clear and meaningful data models, vocabularies and information exchange packages.
Domain model	A domain model is a conceptual view of a system or an information exchange that identifies the entities involved and their relationships.
Glossary	A glossary is a simple list of terms and their definitions. A glossary focuses on creating a complete list of the terminology of domain-specific terms and acronyms.
Information exchange package	An information exchange package (IEP) description is a collection of artefacts that define and describe the structure and content of an IEP. An Information Exchange Package Documentation has a specific information exchange context and may refer

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

	to other semantic assets.
Mapping	A mapping is a relationship between a concept in one vocabulary and one or more concepts in another.
Ontology	An ontology is a formal, explicit specification of a shared conceptualisation.
Schema	A schema is a concrete view on a system or information exchange, describing the structure, content and semantics of data.
Service description	A service description is a set of documents that describe the interface to and semantics of a service.
Taxonomy	A taxonomy is a scheme of categories and subcategories that can be used to sort and otherwise organise items of knowledge or information.
Thesaurus	A thesaurus is a controlled and structured vocabulary in which concepts are represented by terms, organised so that relationships between concepts are made explicit, and preferred terms are accompanied by lead-in entries for synonyms or quasi-synonyms.

Table 1. Types of interoperability building blocks. These definitions map to the EU Vocabulary for Dataset type[2] which is also a recommended controlled vocabulary for the HealthDCAT-AP DataSet *type* attribute.

The 1+MG Framework Recommendations[3] for Data models, standards & ontologies include the following building blocks: Phenopackets, DUO, DCAT, DCAT-AP, and the 1+MG Minimal dataset for cancer. In addition, the B1MG prioritised standards for cross-border exchange of EHR data[4] include LOINC, SNOMED, UCUM, WHO ATC, ICD-10, ORPHAnet, EMA SMS substances, EDQM, HL7 v3/FHIR, HPO, ISO IDMP, ISCO-08, EMDN, ISO Country and language codes and FAIR genomes (a promising best practice genomics data framework within 1 + MG and B1MG). Here, we will highlight the standards that are most relevant as interoperability building blocks to support the currently selected GDI and 1+MG use cases and questions, in alphabetical order. These building blocks are further supplemented with extensions and domain standards where needed. The standard used are in alignment with TEHDAS Recommendations to enhance interoperability within HealthData@EU[5], which are *Standards for data discoverability (meta-data standards)*: Beacon, BBMRI-MIABIS, Bio-image archive, CESSDA CMM, DCAT-AP, ECRIN-CRMDR, FAIRSHARING, INSPIRE, PHIRI. *Standards that enable semantic interoperability*: CDISC SDTM, LOINC, OMOP CDM, Orphanet, SNOMED CT. *Standards for interoperable communication*: DICOM, HL7 FHIR, IDMP (SPOR), ISO 8000-110.

[2] https://op.europa.eu/s/zlab
[3] https://framework.onemilliongenomes.eu/data-models-ontologies
[4] https://zenodo.org/records/10058688
[5] https://tehdas.eu/app/uploads/2023/10/tehdas-recommendations-to-enhance-interoperability.pdf

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

1+Million Genomics Minimal Cancer Model (1+MG-MDC) - *Domain model*

To revolutionise healthcare through precision cancer medicine, seamless access to clinical and genomic data across borders is essential. Given the varied levels of digitalization globally, standardised data collection models are crucial for collaborative efforts. The European Union's 1+Million Genomes (1+MG) initiative, supported by Horizon 2020 Beyond 1 Million Genome project, aims to establish data models, best practices, and technical infrastructures for accessing sequenced genomes, including cancer genomes. The cancer-focused Working Group 9 (WP9), has created the 1+MG-Minimal Dataset for Cancer[6][7] (1+MG-MDC) – a comprehensive data model with 138 items across eight conceptual domains, facilitating the collection of cancer-related clinical information and genomics metadata. Building on existing models, the 1+MG-MDC emphasises annotation and traceability of aspects relevant to cancer's complex trajectory and treatment. Designed to be user-friendly yet comprehensive, it caters to both clinicians and researchers.

Beacon v2 - *Service description*

The Beacon v2 default (i.e. reference) model outlines the fundamental entities such as individuals or biosamples, as well as delineates the relationships between them. The key entities within the model are defined as follows: ***Collections (Datasets and Cohorts):*** These are groupings of variants or individuals that share common attributes, such as belonging to the same repository (datasets) or originating from specific study populations (cohorts). ***Genomic variations:*** These represent unique alterations within the genome, including positional information, sequence modifications, and type of variation. ***Individuals:*** This category encompasses both patients and healthy controls whose detailed information, including phenotypic and clinical data, is stored within the repository. ***Biosamples:*** These are samples collected from individuals, providing details about the procedures involved, as well as dates and times of collection. ***Analyses & Runs:*** This category includes information on the procedures used for sequencing biosamples (runs), as well as the bioinformatic processes employed to identify variants (analyses). When fully expanded across all data types including optionals and choices, the Beacon v2 model consists of over 400 variables.

DCAT - *Domain model*

The Data Catalog Vocabulary[8] (DCAT) is an RDF vocabulary crafted to enhance interoperability among data catalogues available on the web. By employing DCAT to delineate datasets within these catalogues, publishers enhance their visibility and empower applications to access metadata from various catalogues. DCAT facilitates decentralised catalogue publishing and simplifies federated dataset searches across multiple catalogues. Aggregated DCAT metadata can function as a manifest file, streamlining digital preservation efforts. The latest version of DCAT is version 3 released on 18 January 2024.

DCAT-AP - *Application profile*

[6] https://www.nature.com/articles/s41588-024-01721-x
[7] https://zenodo.org/records/8239363
[8] https://www.w3.org/TR/vocab-dcat-3

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

The DCAT Application profile for data portals in Europe (DCAT-AP) is a set of guidelines built upon W3C's Data Catalogue vocabulary (DCAT). It defines how to describe public sector datasets in Europe. Its main purpose is to facilitate searching for datasets across different data portals, improving the accessibility of public sector data across various borders and sectors. The development of DCAT-AP was a collaborative effort involving DG CONNECT, the EU Publications Office, and the Interoperable Europe Programme. A diverse Working Group, comprising representatives from 16 European Member States, various European Institutions, and the US, worked on elaborating the specification. The current version of DCAT-AP is v3.0 which was released on 12 February 2024. See: https://semiceu.github.io/DCAT-AP/releases/3.0.0/.

European Health Data Space (EHDS) - *Core component*

The European Health Data Space[9] (EHDS) is a fundamental component of the European Health Union, building upon existing regulations such as the General Data Protection Regulation (GDPR) and the NIS 2 Directive. It comprises specific guidelines, shared standards, infrastructure, and governance arrangements. These include enabling individuals to access and manage their electronic personal health data more readily, both within their national contexts and across the EU. Furthermore, it aims to cultivating a unified marketplace for electronic health record systems, relevant medical devices, and high-risk AI systems. Lastly, it establishes a reliable and effective framework for utilizing health data in research, innovation, policymaking, and regulatory endeavors (secondary data usage).

FAIR Data Point- *Application profile*

FAIR Data Point[10] (FDP) has adopted DCAT version 2. Its primary aim is to create a standardized approach for providing and accessing metadata. By introducing a few structural extensions to DCAT, it ensures that client applications have a reliable method for accessing and interacting with metadata content. The current version is 1.2, released on 20 October 2023. FDP will be updated to DCAT-AP v3.

FAIR genomes - *Domain model*

FAIR genomes[11] is a promising best practice genomics data framework within the 1 + Million Genomes (1 + MG) and Beyond 1 Million Genomes (B1MG) initiatives[12]. FAIR genomes [1] is a national (Dutch) coordination action to unite currently fragmented guidelines & tools to increase FAIR-ness of DNA data, uniting work from all types of DNA laboratories (rare disease, cancer, research, etc), patients/participants organisations, and has extensive collaborations with (inter)national initiatives, including aligned with NL and international organisations BBMRI, ELIXIR, X-omics, Solve-RD, EJP-RD, GA4GH. A key output from the project is a semantic schema 110 elements categorised into 9 modules. Built on common ontologies such as NCIT, DUO, and EDAM, it introduces new terms only when necessary. This schema is represented by a YAML file, convertible into templates for data entry

[9] https://www.european-health-data-space.com
[10] https://specs.fairdatapoint.org
[11] https://fairgenomes.org
[12] https://zenodo.org/record/5018520/files/B1MG%20D3.7%20-%20Documented%20best%20practices%20in%20sharing%20and%20linking%20phenotypic%20and%20genetic%20data%20-%201v0.pdf

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

software (Electronic Data Capture - EDC) and programmatic interfaces (JSON, RDF), streamlining genomic data sharing in both research and healthcare settings.

HealthDCAT-AP - *Application profile*

HealthDCAT-AP[13] is a fundamental specification for metadata records within the European Health Data Space (EHDS) and other European data initiatives. It is an extension of the EU DCAT Application Profile (DCAT-AP) to better suit the discovery of health data and enhance the accessibility and comprehension of electronic health records. This extension prioritizes privacy and security, ensuring responsible and efficient sharing of sensitive information. This effort is part of the broader EHDS and aligns with the objectives of the EU4Health Programme, which seeks to bolster health systems. By meeting the outlined goals in Regulation (EU) 2021/522, this project contributes significantly to building a resilient, accessible, and efficient health data infrastructure across Europe. This improved infrastructure is poised to catalyze the digital transformation of healthcare within the EU, promoting a more data-centric, streamlined, and patient-focused approach. Ultimately, these advancements will lead to substantial improvements in public health outcomes and healthcare provision throughout member states.

Joint Research Centre Common Data Elements (JRC-CDE) - *Domain Model*

The EU RD Platform has released the JRC-CDE to enhance the interoperability of RD registries. The JRC-CDE comprises 16 essential data elements[14], also available in semantic representation[15]. Each rare disease registry across Europe is tasked with registering these elements, which encompass patient personal details, diagnosis, disease history, care pathway, research-related information, and disability specifics. The development of this set was overseen by a Working Group led by the JRC and composed of experts from EU projects such as EUCERD Joint Action, EPIRARE, and RD-Connect, all dedicated to standardising data sets.

Minimal dataset for infectious diseases - *Domain model*

GDI Task 7.3 and 1+MG WG11 experts working on developing a Minimal Dataset for Infectious Diseases (MDID), going beyond COVID-19. focused on compiling ontologies covering key concept areas like data submission, host characteristics, pathogens, sequencing of host samples and pathogens, treatment, and environmental variables. These efforts are aimed at establishing standardized protocols for data collection. The MDID serves as a blueprint, offering guidance on how observational data should be collected according to specific standards.

Observational Medical Outcomes Partnership (OMOP) - *Schema*

The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) serves as an open community data standard aimed at unifying the structure and content of observational data. Its primary goal is to facilitate efficient analyses capable of generating reliable evidence. At the heart of

[13] https://healthdcat-ap.github.io
[14] https://eu-rd-platform.jrc.ec.europa.eu/set-of-common-data-elements_en
[15] https://jbiomedsem.biomedcentral.com/articles/10.1186/s13326-022-00264-6

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

the OMOP CDM lies the OHDSI standardised vocabularies. These vocabularies play a pivotal role in organising and standardising medical terminology across various clinical domains within the OMOP common data model. They empower standardised analytics by providing a knowledge base for constructing exposure and outcome phenotypes, along with other features essential for characterization, population-level effect estimation, and patient-level prediction studies.

Phenopackets - *Information exchange package*

A Phenopacket[16] connects thorough descriptions of physical traits with data on diseases, patients, and genetics. This allows clinicians, biologists, and researchers studying diseases and drugs to create more comprehensive disease models. Phenopackets are stored as PXF (Phenotype Exchange Format) files, which can be saved in JSON or YAML formats. Each packet links a set of physical abnormalities to a disease and patient, providing information such as age, gender, when symptoms started, and supporting evidence. PXF employs standard ontologies to ensure compatibility among various sources and users, streamline text analysis, and facilitate machine-based reasoning.

SNOMED-CT - *Ontology*

SNOMED-CT is a comprehensive and standardised clinical terminology used in healthcare. It provides a structured and universal language for capturing, sharing, and exchanging health information across different clinical settings and systems. SNOMED-CT plays a crucial role in enhancing interoperability and communication in the healthcare domain by offering a common vocabulary for describing clinical concepts and relationships. OMOP and SNOMED-CT are part of EOSC4Cancer SOPs for minimal clinical, image and omics data exchange. In parallel projects related to EOSC4Cancer, such as AACR GENIE, B1MG and IMPaCT-Data, similar clinical data models are present. All these projects demonstrate high interoperability with either OMOP CDM or cBioPortal, offering the potential for adaptation and harmonisation to enhance datasets within the EOSC4Cancer initiative.

4.3. User communities and groups

In this section we aim to provide examples of typical user communities that would use GDI as the basis for exchanging genomics data and depositing data in 1+MG/GDI. By depositing their data in 1+MG/GDI, who will be an authorised participant in the European Health Data Space (EHDS), projects will automatically fulfill the EHDS requirements and avoid potential penalties. Therefore, by supporting this scenario across all communities we will solve multiple challenges at once. While some communities are linked to GDI from the start, others are pre-existing and have already solved their genomics data exchange, developing best-practices on the way and offering synergistic approaches. By considering the broader communities, we can elucidate more precise requirements and methods to make future implementations of European projects within the GDI infrastructure significantly smoother. The relevant communities for Genome of Europe, Cancer, Infectious Diseases and Rare Diseases are represented in the 1+MG WGs 8, 9, 10, 11 and 12 and their corresponding

[16] http://phenopackets.org

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

requirements implemented via GDI. Below is a brief description of each community and closely related projects, each providing input for the interoperability scenarios described in section 4.4.

4.3.1. Genome of Europe (GoE)

The GoE consortium is a part of GDI Task 7.1 and WG12 of the 1+MG initiative. It is dedicated to establishing a European network of national genomic reference cohorts, each consisting of at least 500,000 individuals. These cohorts will be chosen to accurately represent the diverse population of Europe. In pursuit of this goal, participating countries will create **population cohorts** that mirror the genetic makeup of their citizens, encompassing both healthy individuals and those affected by disease. Subsequently, these individual datasets will be integrated to form a cohesive European **reference dataset.** The sequencing of 500,000 genomes amassed through this initiative will significantly contribute to the overarching aim of sequencing one million genomes, as envisioned in the 1+MG initiative.

Projects similar to GoE that synergize with the GDI infrastructure include the Genome of the Netherlands[17] (GoNL), which was launched as a Rainbow Project by BBMRI-NL, the Dutch biobank collaboration due to its exceptional scientific potential and its promise for advancing new treatments and diagnostic methodologies. By scrutinising the DNA of 750 Dutch individuals, comprising 250 trios of two parents and their adult offspring, alongside a broad genetic profile of a significant Dutch population, this initiative promises to unveil a plethora of novel information, insights, and potential applications.

4.3.2. Cancer

GDI Task 7.4 and 1+MG WG9 comprises cancer export teams and specialists from European and other relevant cancer-related projects that together generate guidelines and methods ready for the integration in the different national cancer research activities. They contribute to feasible technical solutions in collaboration with Pillar II experts to facilitate access to genomic and phenotypic datasets provided by related projects such as EOSC4Cancer, as well as to the technology and recommendations of the parallel work of the 1+MG/B1MG cancer use-case and in alignment with the EHDS2 and the Cancer Mission. Three clinical use cases have been defined: melanoma, non-small cell lung cancer, and chronic myeloid leukaemia. This task will start by developing solutions primarily for colorectal cancer, making best use of the existing wealth of experience and systems in this area. However, it will also prioritize the adaptation of these solutions across other cancer types.

Furthermore, GDI Pillar III activities are connected to relevant European developments on biomedical research. This includes the European Open Science Cloud (EOSC) for Cancer (EOSC4Cancer), which is considered one of the projects critical for the success of the 1+MG infrastructure. The architectural requirements evolved in GDI will support the Pillar III use cases and ensure compatibility with EHDS and the EOSC.

[17] https://www.nlgenome.nl

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

EOSC4Cancer aims to enable access to a wide array of cancer data, encompassing genomics, imaging, medical records, clinical data, environmental factors, and socio-economic indicators. Building on federated and interoperable systems, it ensures secure identification, sharing, processing, and reuse of FAIR data across borders. Through community-driven analysis environments, EOSC4Cancer facilitates the exploration and utilisation of these datasets including advanced techniques like machine learning and artificial intelligence. Across its five distinct use cases, EOSC4Cancer spans the entire patient journey from cancer prevention through diagnosis to treatment. By establishing robust data trajectories and workflows, it lays the groundwork for future European Cancer Mission projects.

Other cancer projects that would synergize with the GDI infrastructure include the FORCE, EUCAIM and UNCAN projects. FORCE aims to improve the detection and treatment of rare cancers. To achieve this, FORCE collects information (medical data and bodily materials) from patients with rare cancers, so that it can be used to answer important questions and thus improve care in the future. FORCE makes the best possible use of medical data that is already recorded, for example in the hospital record or in the Dutch Cancer Registry (NKR). In addition, FORCE also stores imaging materials, such as CT scans or MRI scans, to closely monitor the growth of the cancer. The Cancer Image Europe (EUCAIM) project offers a reliable platform where researchers, clinicians, and innovators can access a variety of cancer images. This enables them to benchmark, test, and pilot AI-driven technologies. By linking top-notch cancer image data with AI expertise, Cancer Image Europe fosters collaboration and speeds up the advancement of state-of-the-art solutions for cancer diagnosis and treatment. The European Initiative to UNderstand CANcer (UNCAN.eu) proposes to set up a European Federated Cancer Research data hub and generate a series of use cases, addressing major challenges in cancer research.

4.3.3. Infectious diseases (ID)

GDI Task 7.3 and 1+MG WG11 will organize, classify, and disseminate information, links, and software related to accessing genomic data repositories (both human and pathogen) and other resources developed during the COVID-19 pandemic. This includes advancements in the 1+MG/B1MG use-case and the BY-COVID project connected to the European COVID-19 data portal, along with various national and international initiatives. Its objective is to develop sustainable solutions that complement and strengthen national projects, ensuring their longevity beyond the pandemic and applicability to other infectious diseases. Currently, this task is working on the Minimal Dataset Infectious Diseases (MDID). It has links to ELIXIR CONVERGE WP7: Federated European Genome-phenome Archives for transnational access of COVID-19 host data. This work package will work together with the ELIXIR expert data managers network (WP1) to collate the national COVID-19 DMPs and metadata requirements. Furthermore, this task will make major technological contributions to the infectious diseases use-case by bringing in the existing experiences on ViralBeacon.

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

4.3.4. Rare diseases (RD)

GDI Task 7.2 and 1+MG WG8/WG10 ensure that the insights gained concerning the use cases Rare Diseases (WG8) and Common Complex Diseases (WG10) from the 1+MG/B1MG initiative are integrated into GDI. It will incorporate the lessons learned and advancements made in the 1+MG/B1MG Proof of Concept (PoC) demonstrator to implement other use cases within Pillar III. This involves setting up infrastructure for querying and extracting genomic information. Additionally, it will collaborate closely with the 1+MG/B1MG to provide recommendations on data sources, standards, analysis, visualization systems, and resources. These recommendations will guide the implementation of various national projects, such as the Dutch FAIR Genomes initiative.

The European Rare Diseases Research Alliance (ERDERA) has the ambition to improve the health and well-being of the 30 million people living with a rare disease in Europe, by making Europe a world leader in RD research and innovation, to support concrete health benefits to rare disease patients, through better prevention, diagnosis and treatment. Its goals of sharing data and enabling research are perfectly aligned with GDI and therefore offers a critical opportunity to achieve synergy between these projects. ERDERA is considered the spiritual successor of both Solve-RD and EJP-RD.

The Solve-RD project[18] provided molecular diagnoses to many rare disease cases and increased the understanding of rare diseases for better diagnostic capabilities and outcomes through advanced omics approaches. Integrated with European Reference Networks (ERNs) for rare diseases, particularly ERN-RND, EURO-NMD, ERN-ITHACA, and ERN-GENTURIS, Solve-RD collaborated with patient cohorts across all 24 ERNs and undiagnosed disease programs in Spain and Italy. The consortium comprised clinicians, geneticists, researchers, patient organisations, and experts in omics technologies and bioinformatics. The RD3 database was made available to members of the project. Users with permission can interact with the data in the browser to find patients that meet specific criteria or use the Discovery Nexus tool to build groups of similar patients. For example, users would like to find patients and samples by a specific analysis (e.g., SR-RNAseq). Using the browser, users can navigate to the samples table and create a new filter for on analysis types.

In the European Joint Programme on Rare Diseases (EJP-RD), the Virtual Platform (VP) was developed. The VP streamlines access to resources crucial for rare disease research. Serving as a unified gateway, it enables seamless discovery, querying, and eventual access to patient registries of the 24 ERNs, biobanks, genomics and multi-omics repositories, knowledge bases, and specialised resources like animal models and cell line libraries. Through a federated ecosystem approach, resources are optimised to align with rare disease research needs and made FAIR within this context. Data remains securely stored at its source but can be queried remotely from an EJP RD query point, enabling federated discovery, query, and analysis while safeguarding patient privacy and respecting each resource's access conditions.

[18] https://solve-rd.eu

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

4.4. Analysis of functional interoperability scenarios

We base the approach in this report on over a decade of experience in a diverse array of research projects and groups. However, to make much of this implicit knowledge we have inventoried different dimensions of interoperability scenarios. We first collected a set of interoperability scenarios with the aim to provide understanding of functional aspects of interoperability of GDI infrastructure. Subsequently, we took a technical approach aiming to tease apart technical challenges within these scenarios that would need to be addressed to successfully deliver functionality serving the functional scenarios. Finally, we analysed technical scenarios that GDI could implement to address these functional and technical scenarios.

GDI nodes exist as part of a complex data lifecycle bridging between data collection, capture, and integration phase of genomics data, which is upstream from each GDI node, and the data pooling/analysis/study phase, which is downstream from each GDI node. These concepts are shown in **Figure 3**. Each of the 1+MG WG use cases, as well as the WP7 defined use cases, have or will soon have a 1+MG approved/confirmed minimal dataset. The GDI data inclusion will be based on these minimal datasets to answer the use case questions. Possible sources of these data include national genomics plans for personalized medicine, health-care providers (HCPs) and research groups/teams such as genome diagnostics departments and population biobanks. We are in particular mindful of the challenges that will be faced by GDI nodes, as UMCG might be one of them. Based on our experience running a precursor of GDI functionality in Groningen one of the biggest challenges is how to collect sufficiently rich metadata so that the GDI datasets will be useful. This data is locked up in existing systems that often don't have reuse by third parties as their priority. In addition, as partners in many EC consortia, we are mindful of the diversity of post-processing software used by research consortia. We believe that GDI would have greatly increased acceptance if both data collectors and data users would be facilitated (within reason). In **Table 2** we outline the interoperability challenges that come with the complete data lifecycle, including collection, discovery and analysis.

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

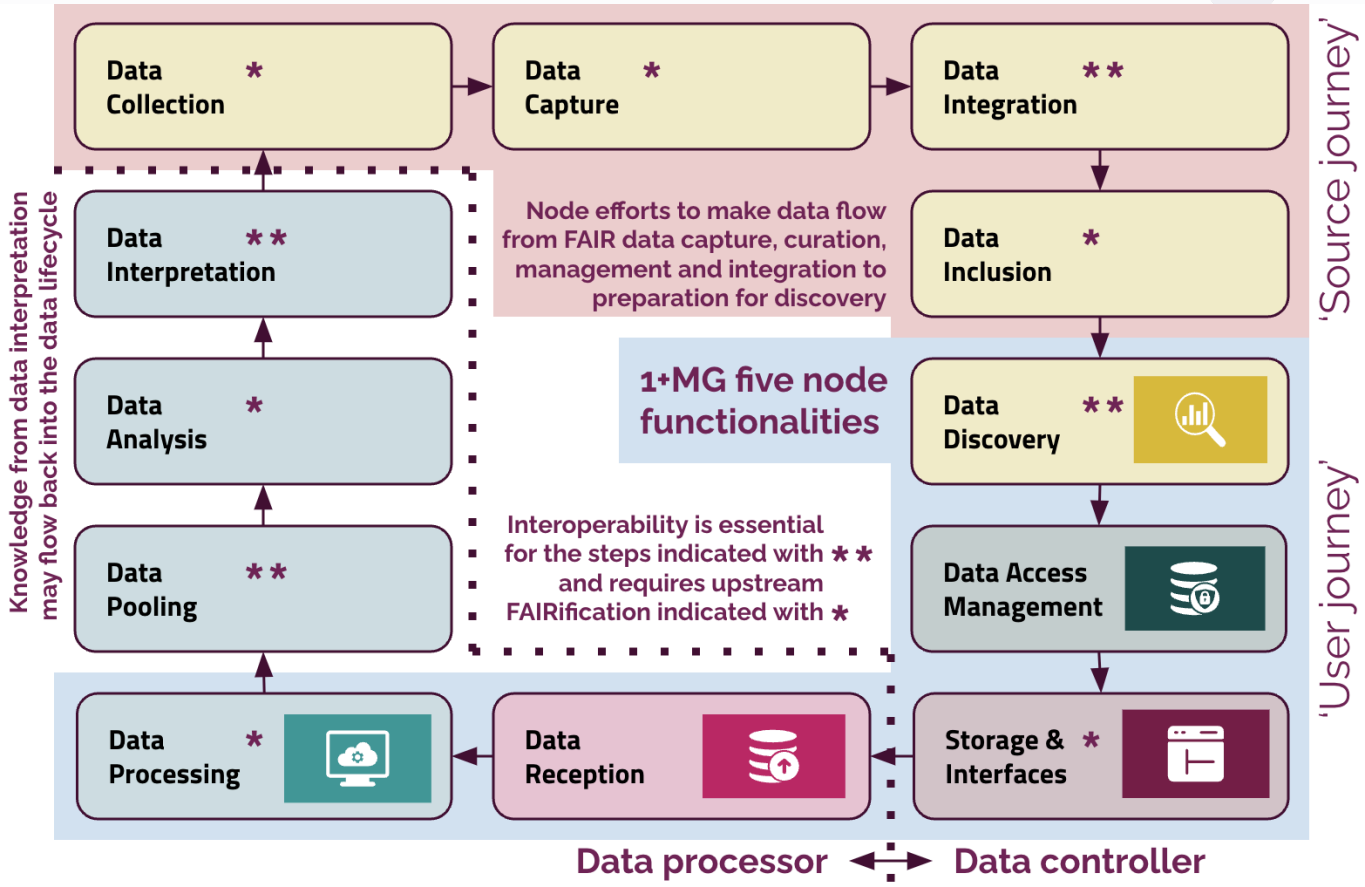

Figure 3: GDI node data lifecycle. For certain steps, interoperability is essential for successful completion. However, their upstream steps are responsible for delivering the necessary FAIR data.

Table 2: Examples of interability scenarios. This list is not exhaustive but should provide sufficient examples to validate interoperability requirements.

Functional scenario	Interoperability elements
Scenarios where data is ingested* into GDI	
As a HCP I want to register new patients and samples. The IDs should already be generated and I would like the system to suggest an ID or allow me to select a new one. It should also be possible to link a patient to one or more sample IDs.	Domain model (a sunflower union of all models applicable to patients and samples) and Application profiles (HealthDCAT-AP) and Data Management Systems capable of generating and linking identifiers.

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

As a HCP I want to share diagnostic genomes data. We would like to add new data and manage existing records, as well as mark records that are complete/verified.	EHRs/LIMS/systems specific to pathology/genetics capable of exporting data into GDI nodes fulfilling 1+MG Minimal metadata model via sunflower union and corresponding data and quality requirements.
As a HCP I want to see how my site is performing. I want to see how much data I've submitted and how my site is progressing in terms of the project's objective.	Data Management Systems capable of tracking submissions and providing visual dashboards.
As a biobank organisation I want to share sample information for the genomics data.	Domain model (a sunflower union of all models applicable to samples) and BIMS capable of export into GDI node.
As a cohort study I want to clarify the collection structure of the shared genetics data.	Domain model (to specify subpopulations and timelines of data collection).
As an ERN I have a patient registry of patients whose genetics data could be shared in GDI.	Domain model (Domain models (a sunflower union of all models applicable to patients, in particular JRC-CDE) and Application profile (HealthDCAT-AP).
As a project coordinator I want to understand how the project is progressing in terms of recruitment/registration/submission by a number of factors/outcomes: e.g., overall vs center, diagnosis, genes, etc.	Data Management Systems capable of tracking processes such as recruitment, registration, and submission, all displayed in faceted visual dashboards.
Scenarios where GDI data is discovered and requested	
(GoE/Cancer/ID/RD) As a rare disease researcher I want to count available samples carrying a variant.	Service description (e.g. Beacon v2) that have the appropriate domain model on board.
(RD) As a researcher, I want to find patients and samples by a specific analysis (e.g., SR-RNAseq), phenotype, known mutations, ERN, status (unsolved/solved).	Domain model (a sunflower union of all models applicable to patients, samples, omics data, inclusion status, etc) and Application profiles (HealthDCAT-AP).
(RD) As a rare disease researcher I want phenotypes associated to the genetics data.	Information exchange package (e.g. Phenopackets).
Scenarios where GDI data is prepared for reuse	

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

(GoE/Cancer/ID/RD) As a bioinformatician I need to create a sample sheet as input for centralized and federated interoperability scenarios, depending on the data use conditions.	Info on file locations, sample characteristics, patient ids, identifiers; requirements on input files to be processed.
(GoE/Cancer/ID/RD) As a researcher I want to create a database for a multicenter project where each center has a dedicated schema to import data into and linked to a central schema. Each center level schema should be flexible enough that they reference a central schema without having to hardcode schema names.	Data Management System capable of federative databasing and supporting flexible data models) and Domain model (serving as central schema and specific extensions).
(GoE) Lookup of individual genetic variants (of specific ancestries) for allele frequencies for e.g. diagnostics.	Service description (e.g. Beacon v2) and Domain model (sunflower union of all models supporting variants and allele frequencies).
(Cancer) As a researcher I want to request genetics data available in cBioPortal.	Information exchange package (import format of cBioPortal) or Data Management System capable of delivering the result of the data request into cBioPortal).
(RD) As a rare disease researcher I want phenotypes associated to the genetics data.	Information exchange package (e.g. Phenopackets)
Make cross-national GWAS over full genomes possible by creating interoperability between older and newer sequencing technologies that the member states have access to.	Domain model (including ring or petal with rich sequencing metadata including technology and various quality control metrics).
Dynamic capture of phenotypic parameters because science focus will change over time.	Domain model (flexible enough to add phenotypes at runtime) or Core component (EAV-model for phenotypic data).
Scenarios where GDI data is used for specific research questions/use cases. These require (specific) domain-specific models in combination with appropriate standardized file formats.	
(GoE) Recalibration of polygenic risk scores	
(GoE) Performing ancestry-specific imputation	
(Cancer) Can the tumour relapse be explained by the NGS molecular profile at recurrence?	
(Cancer) What are risk factors in cancer prevention?	

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

(Cancer) How can we detect cancer earlier?
(Cancer) How can we improve the risk/benefit ratio of screening?
(Cancer) Who to treat with a given (neo)adjuvant anti-cancer therapy?
(Cancer) Which patient to assign to what trial?
(Cancer) Can we match tumour mutations against a database with known variants responsible for tumour regrowth or therapy resistance?
(Cancer) Is there a second line of treatment against this second mutational hit?
(Cancer) Determining MicroSatellite Instability (MSI) score in ColoRectal Cancer for prognostic and therapeutic choices as well as studying defective DNA repair mechanisms.
(ID) GWAS (validation): risk variants of severe COVID-19 (research).
(ID) Identify variants that may guide prognosis and/or treatment (healthcare).
(RD) Why do people with certain disease-specific genes not develop the disease?
(RD) Why do some gene variants cause adverse side effects for medications?

* For now, we assume that data ingestion will be routed via the national nodes. This means that individual institutes or HCPs will not submit to the GDI infrastructure directly, but only via their national nodes or contacts. This is the case even when a country does not operate their own national node, but has transferred this role to a central European-wide node or a selected country node. Broader advice and vision on this topic is required given the implications on various scenarios with accompanying interoperability challenges.

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

5. Description of work accomplished

To collect evidence that allows GDI members to evaluate the proposed framework, we have developed a **reference implementation of the interoperability framework intended to support the integration and development of GDI interoperability building blocks and implementation of GDI interoperability scenarios in future work**. This reference implementation is a work in progress for demonstration purposes, and together with our partners we want to find out how this approach can best benefit us all as a basis for diverse implementations by GDI node and central services. To enable rapid prototyping with option for longer term sustainability, we used existing MOLGENIS software, however, the concepts in this implementation aim to be reusable without preference for a specific software. Below, we summarise implementation of all elements (A-G) of the framework based on the concepts described in section 4.1 and Figure 2.

5.1 Composite logical model (A) implementation: 'metadata registry'

We have used a simple format to almagate of the results of all GDI WGs united by integrated interoperability building blocks into one composite logical model, This serves as a 'metadata registry' listing all data elements and using 'tags' to enable selection of elements from a particular building block. The metadata registry will become available via an interactive app at the GDI local portal demonstrator.

This 'metadata registry' aims to provide a starting point that different software providers in GDI could adopt to implement interoperability scenarios. Here, we demonstrate the viability of this concept using a reference implementation that can select items and create smaller, purposed models from a larger composite model. This fits the strategy as adopted by 1+MG to form a sunflower model for metadata. Core tables and columns include DCAT-AP. Attached to this core, HealthDCAT-AP is added as the health domain specific extension. HealthDCAT-AP is then further specified using additional petals per example use case, omics datatype, disease, and so on. This sunflower model is also applied to datasets that will be defined using a minimal core in combination with extended rings and petals that fully capture all details of a dataset while keeping commonalities with metadata of other datasets intact and cross-usable for discovery purposes.

The current reference implementation of the composite model (at April 19th 2024) consists of 64 tables and 1238 columns. It is expressed in CSV files reusing MOLGENIS emx2 format with the following columns: tableName, tableExtends, columnName, columnType, key, required, refSchema, refTable, refLink, refBack, validation, semantics, description, profiles. The exact files in which the table and column definitions are located is arbitrary, meaning that they can be easily grouped by subject per file, e.g. "Individuals", "Cohorts", "Studies", etc. Part of this model are so-called tags for 'profiles' that allow selecting groups of tables and columns that belong together, aligned to the sunflower model. There are currently 13 used tags: SharedStaging, RD3, Beacon v2 EMX2 add-on,

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

Beacon v2 Vivify add-on, GDI T8.2 union, FAIR Genomes metabolomics add-on, DCAT, Beacon v2, JRC-CDE, DataCatalogue, CohortStaging, EMA, FAIR Genomes, DCAT files add-on. An overview of the tables and column counts can be found in the Annex. The current model is available at the MOLGENIS EMX2 GitHub[19] with a versioned snapshot available at Zenodo [todo]..

5.2 Purposed logical model (B) configuration: 'application profiles'

Profiles offer a convenient method to define purposed data models by selecting tables and columns from the composite logical model, i.e. the big shared data model. For this, we have tagged all data elements in aforementioned metadata registry to ease the selection of elements that a part of a particular building block, e.g. 'dcat' or 'beacon v2'. Then in an application profile one can configure quickly what parts of the metadata registry should be included for a particular application.

These profiles do more than selecting data items. They can be used to define applications that consist of various resources bundled with EMX2, including ontologies, example data, application data, and predefined system settings. Advanced options include adding user permissions and nesting of profiles. In the reference implementation, these profiles are represented as simple YAML files, which are also available at the MOLGENIS EMX2 GitHub. For more details, see the EMX manual[20].

Profiles are automatically applied to the shared data models located in the models directory. Tags from your profile are compared against those present in these models. When selected columns reference ontologies, they are automatically sourced from the ontologies folder.

There are currently 9 application profiles drawing from this model by using 2299 profile tags on the tables and columns. An overview of the profiles, tables and column counts can be found in the Annex. The 'FAIR Data Hub' profile contains many core features needed in GDI, but will most likely be extended to fit particular needs.

5.3 Physical model (C): database generation

To actually make the application profile 'work', the application profile must be implemented into working software implementations. Many data management systems (DMS) offer the flexibility of defining part of their data model at runtime to accommodate diverse data capture needs. This is typically achieved by importing a data model template, or blueprint, into the DMS. Once the model is validated and configured, the system can accept data conforming to this blueprint, which can be inputted through programmatic interfaces or human-friendly forms, all built on the same model.

[19] https://github.com/molgenis/molgenis-emx2
[20] https://molgenis.github.io/molgenis-emx2

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

Therefore, to streamline the recording data directly from sources with different DMS, we've developed software capable of parsing logical models and generating these templates/blueprints for setting up DMSs in the FAIR genomes project[1]. Currently, this provides direct support for several widely used DMSs in Dutch academic hospitals, including MOLGENIS EMX1/EMX2, Castor, REDCap, and OpenClinica. Additionally, models can be implemented in HL7/FHIR-based standards using the ART-DECOR framework hosted by Nictiz and use the iCRF Generator as a tool to facilitate re-use of hosted and approved (community maintained) standards (e.g. ART-DECOR, CDISC, MDM, openEHR) in prospective studies that use an Electronic Data Capture tool (e.g. MOLGENIS, Castor, REDCap or OpenClinica). By utilizing these blueprints, DMSs gain interoperability and cross-query capabilities, reducing or eliminating the need for data extraction, transformation, and loading into a centralized database for data retrieval and reuse. For the purpose of this reference implementation we only generated towards MOLGENIS for step 'D'.

5.4 A running data management system (D): application requirements

A data management system (DMS) is a software solution or framework designed to efficiently store, retrieve, manipulate, and analyze data. It typically encompasses tools and processes for organizing, securing, and optimizing data to meet the needs of users and applications within an organization.

Requirements for an effective DMS in the life science & health domain include:

- Flexibility and interoperability: Changing data models of live systems to fit new requirements while facilitating seamless integration with other systems via APIs such as Beacon v2.
- Data Quality: Maintaining high-quality data through well-defined models and data types in addition to input validation using specific rules (e.g. Age > 0 & < 150) to ensure accuracy, completeness, and consistency.
- Scalability: The system should be able to handle growing volumes of data. This is typically handled by only storing the metadata while linking to the underlying raw genomic data.
- Ease of use: Offering intuitive user interfaces, simple import/export data formats and programmatic interfaces such as a Python client to simplify data management tasks.
- Data security: Ensuring data confidentiality and privacy, which may include authentication such as Life Science AAI, access control such as REMS, and encryption of underlying genomic data.
- Open source: to promote collaboration, transparency, and innovation by allowing anyone to access, modify, and distribute software freely under permissive licences such as LGLPv3.
- Professionally hosted and supported, adhering to service-level agreement (SLA) that still have to be defined and agreed upon by the GDI project.

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

5.5 Request and response processing (E): payload mappings

Next to internal data management, a key element to serve interoperability scenarios is the ability to provide interfaces to represent data in different serialisation formats and structures (aka 'payloads') adhering to different standards. Therefore, processing of requests and responses is necessary. This challenge might be exacerbated when an API must deviate from standard behaviour, or when an API must be implemented in project-specific ways by design, all without changing the underlying data models or source code. For instance, the European Joint Programme on Rare Diseases has developed an enhanced version of the Beacon v2 query specification[21]. To enable such 'payload mappings' at scale, efficient mechanisms must be available to enable mapping between internal and standard representations.

In this reference implementation, to explore how to efficiently support such additional data types and capabilities, including variations such as extending but without breaking the original Beacon v2 specification, we have investigated the use of JSON query and transformation language (JSLT) templates. These templates are able to translate both requests and responses to match internal data structures when there are differences. In this way, these JSLT templates can support project-specific API flavours such as Beacon v2 extensions and modelling choices, or DCAT extensions by mechanisms such as relabeling, remapping, merging, splitting or any other complex function to achieve intended interoperability.

5.6 Application Programming Interfaces (F): protocol mappings

In addition to mapping of 'payloads' also different service interfaces will be needed, i.e. API's. An Application Programming Interface (API) is a set of rules and protocols that allows different software applications to communicate and interact with each other, enabling data exchange and functionality sharing. Typically, we consider the following types of APIs to be the most useful:

- Straightforward data upload (import) and download (export): here we choose formats that are simple to edit and handle, such as CSV files, zipped CSV bundles, and Excel files.
- Advanced data selection and manipulation: GraphQL, a query language that enables clients to request or edit precisely the data they need for high efficiency and flexibility.
- Linking data: Resource Description Framework (RDF) and its formats such as Turtle (TTL) are a powerful way to express detailed semantic relations and enables linking to similar sources.
- Generic discoverability: datamodel mappings to the DCAT model to enable FAIR Data Point
- Advanced discoverability: domain-specific, potentially sensitive queries using Beacon v2.

[21] https://github.com/ejp-rd-vp/vp-api-specs

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

In the reference implementation we mainly focussed on standard API's (RDF); the Beacon v2 interface still required some manual programming however the payload from section 5.5 did much of the heavy lifting.

5.7 Post-system interoperability effort (G):

As described in section 3.1, in case all above can not be implemented then an intermediate system is needed to still enable integration with an interoperable data federation. While these efforts are outside the analysis scope for this reference implementation, discussions in GDI yielded some pointers which we summarize here. These efforts may include mappings on class, attribute or value level and may be exact or inexact, depending on its purpose. Several examples are listed below. While these are sometimes necessary and worthwhile, they must be implemented on a per-system basis which scales poorly, as depicted in **Figure 1**.

- Attribute matching [2]
- Ontology recoding [3]
- Ontology cross-referencing[22]
- Non-exact mapping using SKOS [4]
- Value recoding [5]
- Query expansion [6]

Furthermore, additional technical challenges include:

- Format mapping: translating or converting data from one file format to another while preserving its structure, content, and meaning.
- Structure mapping: aligning and transforming data from one structure to another while ensuring that the information is accurately represented and preserved.
- Identifier mapping: linking or associating identifiers from one dataset to corresponding identifiers in another dataset.

[22] https://www.ebi.ac.uk/spot/oxo

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

6. Results

This deliverable has resulted in a framework proposal for solving interoperability scenarios containing steps to integrate interoperability building blocks into interoperable systems serving interoperability scenarios of GDI data providers and users. This framework has the potential to guide IT development in GDI complex heterogeneous data landscapes with a variety of stakeholders. We have demonstrated how relevant interoperability building blocks, such as HealthDCAT-AP and Beacon, as well as relevant user communities and groups, such as cancer and infectious diseases, can be converted into working solutions using the framework in context of the semantic interoperability package developed by Task 8.2. Therefore, we have developed a reference implementation of the framework to find out how well it covers interoperability requirements in practice. The goal of this reference implementation is not to be normative, but instead to be used as a free and open source example where ideas can be tried and tested with all partners before investing further in the development of GDI node and central solutions that are deployed across the larger infrastructure.

The reference implementation is a demonstrator based on the long-living free and open source MOLGENIS EMX2 data platform[23], and we simulated it to act as a 'local portal' for the GDI project. This means it had to represent the data management and catalogue in between data discovery and the underlying raw genomics data. These components are integrally connected to handle all local capture, management and discovery of patient and sample metadata in one place. It features a user-friendly graphical interface for (meta)data management, bulk data import and export using Excel/CSV files, RDF extraction, DCAT and DCAT-AP data representation via FAIR Data Point for connecting to the (inter)national portal, Beacon v2 API for genomics data discovery, and a live synchronisation with a REMS instance for GDI to start data access requests, using the Life Science AAI.

A demonstrator is available online[24] and is deployed along with components and specifications from the GDI Starter Kit. It currently contains the B1MG Rare Disease Synthetic Dataset[25] in alignment with GDI Task 7.2 and 1+MG WG8. It consists of 18 Illumina HiSeq 2000 sequencing samples. The catalog is linked to the actual raw sequencing data stored on a high-performance storage and compute environment. We have packaged the demonstrator in a Docker container so others can easily install it for testing and collaboration purposes. This includes EMX2, REMS, KeyCloak for LS-AAI mockup, and a PostgreSQL database. This container is also open source and available on GitHub[26]. We welcome feedback and contributions from all partners.

[23] https://github.com/molgenis/molgenis-emx2
[24] https://portal-gdi-nl.molgeniscloud.org
[25] https://ega-archive.org/studies/EGAS00001005702
[26] https://github.com/molgenis/gdi-localportal

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

7. Discussion

In this deliverable we have showcased a framework to solve interoperability issues based on various interoperability building blocks, communities and interoperability scenarios relevant within the context of the GDI project. A reference implementation demonstrates the potential of using the framework in developing the semantic interoperability package by Task 8.2.

Concretely, the domain-specific data models originating from 1+MG WG3 as well as GDI WP 6 and 8 can be expressed as independent entities, yet managed in an integrated way, to define and prove the interoperability between all of these models. This is an implementation of the sunflower philosophy, where any number of stakeholders, including GDI WP7 T7.1-7.4, can deliver overlapping but slightly different data model specifications without raising overall complexity or causing interference. At the heart of the complete sunflower is the 1+MG metadata model, representing the union of all needs from each of the GDI communities and scenarios. The framework may then help guide implementations of the 1+MG metadata model across different software products to reach a federation of interoperable local and country nodes. For instance, by mapping data from local systems to this model as system-integrated processing steps before chosen API technologies enable query and analytical capabilities.

We are aware that other frameworks, technologies, and software exist that help to separate metadata modelling from creating entry forms for interoperable data. For instance, iCRF Generator [7] is an interactive builder of case report forms by selecting items from online codebooks, LinkML [8] is a powerful language for expressing and sharing data structures with close links to scripting and semantic web, FAIR Data Station [9] can be used to deriving Excel-based data entry forms suitable for offline field work that are later re-uploaded into the larger semantic database, and CEDAR workbench [10] creates semantic metadata templates for data entry using BioPortal terms. While our proposed framework has a larger scope since it ranges from building blocks and scenarios to deploying running federated systems, any existing solutions within this spectrum are highly welcome as complementary, or even synergistic components in solving the interoperability challenge that we face. In fact, the worth of interoperability can only be proven by putting to the test multiple alternative implementations, each with different underlying requirements yet compliant to an agreed upon specification, combined into an effective federation of autonomous, loosely coupled nodes.

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

8. Conclusions & Impact

The inherent heterogeneity of genomic data, stemming from disparate systems, structures, and API specifications, poses a significant challenge, requiring a flexible and pragmatic approach, as outlined in the framework presented. By focusing on interoperability scenarios within and bordering the GDI project, the framework aims to establish a cohesive European (gen)omics infrastructure. Central to this framework are the diverse interoperability building blocks and interoperability scenarios that drive its implementation. This framework may guide and provide requirements for the selection of tools and models that are part of the Starter Kit. It may also act as a template to gather and validate requirements for comparable systems in a constructive and standardised way.

The framework's components, including logical and physical models, data management systems, APIs, and interoperability efforts, form the infrastructure's backbone. Through concerted efforts and investment in a composite logical model, interoperability among systems can be achieved, mitigating the need for post-system interoperability adjustments. Importantly, interoperability is not a binary state but rather a continuum, dependent on the commonalities among systems. While challenges persist, such as the mapping of internal data models to standardised APIs, collaborative efforts can streamline this process and minimise data loss.

Ultimately, the goal extends beyond the framework itself; it lies in achieving interoperability to advance various facets of healthcare, from research to clinical practice. By embracing heterogeneity and adopting innovative solutions, the framework lays the groundwork for a unified ecosystem that empowers researchers and clinicians alike in their pursuit of precision medicine across Europe and beyond.

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

9. Next steps

Task 8.2 will deliver an integrated semantic interoperability package consisting of information, methods and recommendations that will be provided to all initiatives for harmonisation and standardisation of their data to maximise reuse across countries, including through spaces such as the EHDS.

The next project iteration towards this goal, and making use of the semantic interoperability package, will be the demonstration of data access and analysis based on a Pillar 3 use case in Milestone 8. By applying this package, we can scale up the interoperability between available data sets and potential analyses, as shown in **Figure 1**. A key ingredient of the package are the minimal data models. Task 8.2 is developing and harmonising these minimal models from the WP7 datasets towards semantic interoperability of genomic and medical information in close collaboration with WP3 of B1MG and WG3 1+MG.

The task 8.2 harmonisation work currently consists of expressing the confirmed minimal datasets per 1+MG use case working Group in online spreadsheets. These models are then minimized, by determining what is critical and what may be left out, and unionized by combining them together, resulting in a sunflower model for datasets where the core contains the common minimal data items, the rings contain partly specific and partly overlapping data items, and the petals contain data items specific to a particular domain. The complete sunflower represents the maximum set across all use cases. Within their petal, ring or core, data items are assigned 'optionality' which may be mandatory, recommended or optional. This is a low-barrier way of working, allowing all partners to interact freely and create models together. However, these spreadsheets lack structure, validation and versioning, making it cumbersome to manage their content and translate the proposed models into IT infrastructure components such as data management systems and analysis tools. The framework for solving interoperability scenarios as outlined in this deliverable may help guide the translation from proposed models to interoperable IT components.

A concrete next step would be to take up the published minimal dataset on cancer in the sketched interoperability scenarios. This model can be unionized with the other (smaller) models and represents a number of well defined use cases that would prove the principles outlined here and pave the way for further harmonization work. In addition, we may need to find the most effective solution to develop and manage the proposed logical models. The framework piloted here as well as work done by many others (e.g. iCRF Generator, FAIR Data Station, LinkML) offer important ways to shorten the route from data modelling to implementation and test-driving. A good, perhaps combined solution would enable low-barrier contributions from any partner while maintaining a structured and consistent representation that can flow into the framework for downstream implementations.

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

10. References

1. van der Velde KJ, Singh G, Kaliyaperumal R, Liao X, de Ridder S, Rebers S, Kerstens HHD, de Andrade F, van Reeuwijk J, De Gruyter FE, Hiltemann S, Ligtvoet M, Weiss MM, van Deutekom HWM, Jansen AML, Stubbs AP, Vissers LELM, Laros JFJ, van Enckevort E, Stemkens D, 't Hoen PAC, Beliën JAM, van Gijn ME, Swertz MA. FAIR Genomes metadata schema promoting Next Generation Sequencing data reuse in Dutch healthcare and research. Sci Data. 2022 Apr 13;9(1):169. doi: 10.1038/s41597-022-01265-x. PMID: 35418585; PMCID: PMC9008059.
2. Pang C, Kelpin F, van Enckevort D, Eklund N, Silander K, Hendriksen D, de Haan M, Jetten J, de Boer T, Charbon B, Holub P, Hillege H, Swertz MA. BiobankUniverse: automatic matchmaking between datasets for biobank data discovery and integration. Bioinformatics. 2017 Nov 15;33(22):3627-3634. doi: 10.1093/bioinformatics/btx478. PMID: 29036577; PMCID: PMC5870622.
3. Pang C, Sollie A, Sijtsma A, Hendriksen D, Charbon B, de Haan M, de Boer T, Kelpin F, Jetten J, van der Velde JK, Smidt N, Sijmons R, Hillege H, Swertz MA. SORTA: a system for ontology-based re-coding and technical annotation of biomedical phenotype data. Database (Oxford). 2015 Sep 18;2015:bav089. doi: 10.1093/database/bav089. PMID: 26385205; PMCID: PMC4574036.
4. Thomas Baker, Sean Bechhofer, Antoine Isaac, Alistair Miles, Guus Schreiber, Ed Summers. Key choices in the design of Simple Knowledge Organization System (SKOS). Journal of Web Semantics. Volume 20, 2013, Pages 35-49, ISSN 1570-8268, https://doi.org/10.1016/j.websem.2013.05.001.
5. Swertz M, van Enckevort E, Oliveira JL, Fortier I, Bergeron J, Thurin NH, Hyde E, Kellmann A, Pahoueshnja R, Sturkenboom M, Cunnington M, Nybo Andersen AM, Marcon Y, Gonçalves G, Gini R. Towards an Interoperable Ecosystem of Research Cohort and Real-world Data Catalogues Enabling Multi-center Studies. Yearb Med Inform. 2022 Aug;31(1):262-272. doi: 10.1055/s-0042-1742522. Epub 2022 Dec 4. PMID: 36463884; PMCID: PMC9719789.
6. Khader A, Ensan F. Learning to rank query expansion terms for COVID-19 scholarly search. J Biomed Inform. 2023 Jun;142:104386. doi: 10.1016/j.jbi.2023.104386. Epub 2023 May 12. PMID: 37178780; PMCID: PMC10174726.
7. de Ridder S, Beliën JAM. The iCRF Generator: Generating interoperable electronic case report forms using online codebooks. F1000Res. 2020 Feb 4;9:81. doi: 10.12688/f1000research.21576.2. PMID: 32566137; PMCID: PMC7291075.
8. Moxon, S.T., Solbrig, H.R., Unni, D.R., Jiao, D., Bruskiewich, R.M., Balhoff, J.P., Vaidya, G., Duncan, W.D., Hegde, H.B., Miller, M., Brush, M.H., Harris, N.L., Haendel, M.A., & Mungall, C.J. (2021). The Linked Data Modeling Language (LinkML): A General-Purpose Data Modeling Framework Grounded in Machine-Readable Semantics. International Conference on Biomedical Ontology.
9. Bart Nijsse, Peter J Schaap, Jasper J Koehorst, FAIR data station for lightweight metadata management and validation of omics studies, GigaScience, Volume 12, 2023, giad014, https://doi.org/10.1093/gigascience/giad014
10. Gonçalves RS, O'Connor MJ, Martínez-Romero M, Egyedi AL, Willrett D, Graybeal J, Musen MA. The CEDAR Workbench: An Ontology-Assisted Environment for Authoring Metadata that

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

Describe Scientific Experiments. Semant Web ISWC. 2017 Oct;10588:103-110. doi: 10.1007/978-3-319-68204-4_10. Epub 2017 Oct 4. PMID: 32219223; PMCID: PMC7098808.

11. Annex

The full metadata registry, including all profiles, tables and columns, will become available via an interactive browser app at the GDI local portal demonstrator[27].

11.1. Application profiles

Name	Description	Profile tags
European Genomic Data Infrastructure (GDI)	The Genomic Data Infrastructure (GDI) project is enabling access to genomic and related phenotypic and clinical data across Europe.	GDI T8.2 union, DCAT, RD3, JRC-CDE, FAIR Genomes, Beacon v2
FAIR Data Hub	The FAIR life science and omics model	DCAT, DCAT files add-on, FAIR Genomes, FAIR Genomes metabolomics add-on, Beacon v2, Beacon v2 EMX2 add-on, Beacon v2 Vivify add-on
Set of Common Data Elements for Rare Diseases Registration defined by the European Commission's Joint Research Centre	The "Set of common data elements for Rare Diseases Registration" is the first practical instrument released by the EU RD Platform aiming at increasing interoperability of RD registries. It contains 16 data elements to be registered by each rare disease registry across Europe, which are considered to be essential for further research. They refer to patient's personal data, diagnosis, disease history and care pathway, information for research purposes and about disability. The "Set of common data elements for Rare Diseases Registration" was produced by a Working Group coordinated by the JRC and composed of experts from EU projects which worked on common data sets: EUCERD Joint Action, EPIRARE and RD-Connect. For more information, see: https://eu-rd-platform.jrc.ec.europa.eu/set-of-common-data-elements_en	JRC-CDE
RD3	The Rare Disease Data about Data (RD3) database for Solve-RD	RD3
W3C's Data Catalog Vocabulary (DCAT) - Version 2	DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web. DCAT enables a publisher to describe datasets and data services in a catalog using a standard model and vocabulary that facilitates the consumption and aggregation of metadata from multiple catalogs. This can increase the discoverability of datasets and data services. It also makes it possible to have a decentralized approach to publishing data catalogs and makes federated search for datasets across catalogs in multiple sites possible using the same query mechanism and structure. Aggregated DCAT metadata	DCAT, DCAT files add-on

[27] https://portal-gdi-nl.molgeniscloud.org

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

	can serve as a manifest file as part of the digital preservation process. See: https://www.w3.org/TR/vocab-dcat-2/	
European Networks Health Data and Cohort Catalogue	This catalogue contains metadata on cohorts/data sources, the variables they collect, and/or harmonization efforts to enable integrated reuse of their rich valuable data. The contents is grouped by 'networks', such as harmonization projects, EU projects or by navigating all cohorts directly below.	DataCatalogue
CohortStaging	This schema is used by cohorts to enter their data.	CohortStaging
FAIR Genomes metadata schema	The FAIR Genomes semantic metadata schema to power reuse of NGS data in research and healthcare.	FAIR Genomes, FAIR Genomes metabolomics add-on
BeaconV2	The Beacon Project is developed under a Global Alliance for Genomics and Health (GA4GH) Initiative for the federated discovery of genomic data in biomedical research and clinical applications. The version 2 (v2) of the Beacon protocol has been accepted as GA4GH standard in Spring 2022. MOLGENIS EMX2 features a partial implementation of Beacon v2 with a number of extensions. For more information on Beacons, see https://beacon-project.io/	Beacon v2, Beacon v2 EMX2 add-on, Beacon v2 Vivify add-on
SharedStaging	This schema contains communal tables that can be altered by all users.	SharedStaging

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

11.2. Tables and column count

Name	Description	Semantics	Profile tags	Nr. of cols.
Catalog	DCAT v2 catalog entries.	http://www.w3.org/ns/dcat#Catalog	DCAT	9
GenomicVariationsClinInterpr	Beacon v2 GenomicVariations Clinical Interpretations	http://purl.obolibrary.org/obo/NCIT_C125009	Beacon v2	5
IndividualsDiseases	Beacon v2 Individuals Diseases	http://purl.obolibrary.org/obo/NCIT_C15607	Beacon v2	10
IndividualsPhenotypicFeatures	Beacon v2 Individuals PhenotypicFeatures	http://purl.obolibrary.org/obo/NCIT_C116555	Beacon v2	5
Publications	Publications following bibtex format		DataCatalogue, CohortStaging	12
Resources	Generic listing of all resources. Should not be used directly, instead use specific types such as Databanks and Studies	dcat:Resource	DataCatalogue, EMA, SharedStaging, CohortStaging	8
Study	A detailed examination, analysis, or critical inspection of one or multiple subjects designed to discover facts.	http://purl.obolibrary.org/obo/NCIT_C63536	FAIR Genomes	9
datasets	container of files belonging together (different datasets can contain overlapping files, except for the initial datasets containing startfiles (before analysis, startfiles= BAM, gVCF files (one per chromosome) and phenopacket file))	http://purl.obolibrary.org/obo/NCIT_C47824	RD3	6
library	sequencing library information		RD3	6
IndividualsMeasures	Beacon v2 Individuals Measures	http://purl.obolibrary.org/obo/NCIT_C25209	Beacon v2	8
Organisations	Research departments and research groups	org:Organization, foaf:Agent	DataCatalogue, EMA, SharedStaging	21
Cohorts	A group of individuals, identified by a common characteristic.	http://purl.obolibrary.org/obo/NCIT_C61512	Beacon v2, RD3	75
ContactPersons	A person is an object that has certain capacities or attributes constituting personhood. In RD3: researcher or contact person involved in the study and/or affiliated with organisation.	http://semanticscience.org/resource/SIO_000498, http://purl.obolibrary.org/obo/NCIT_C25461	RD3, DCAT	8
DAPs	Data access provider relationship where an institution can provide access to (parts of) a resource		DataCatalogue	6
Dataset	DCAT v2 dataset entries.	http://www.w3.org/ns/dcat#Dataset	DCAT	28

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

Extended resources	Generic listing of all extended resources. Should not be used directly, instead use specific types such as Databanks and Cohorts	dcat:Resource	DataCatalogue, EMA, CohortStaging	19
LeafletAndConsentForm	A document explaining all the relevant information to assist an individual in understanding the expectations and risks in making a decision about a procedure. This document is presented to and signed by the individual or guardian.	http://purl.obolibrary.org/obo/NCIT_C16468	FAIR Genomes	9
Models	Data models	dcat:Dataset	DataCatalogue, EMA	20
Quantitative information	Quantitative information on the resource		DataCatalogue, EMA	9
Subcohorts	Subcohorts defined for this resource	dcat:Dataset	DataCatalogue, CohortStaging	14
datareleases	The act of making data or other structured information accessible to the public or to the user group of a database.	http://purl.obolibrary.org/obo/NCIT_C172217	RD3	7
studies	A detailed examination, analysis, or critical inspection of one or multiple subjects designed to discover facts.	http://purl.obolibrary.org/obo/NCIT_C63536	RD3	14
Aggregates	Resource counts for e.g. samples, clinical data		DataCatalogue, CohortStaging	7
Collection events	Definition of a data collection event for a resource	dcterms:PeriodOfTime	DataCatalogue, CohortStaging	17
Data resources	Resources for data	dcat:Dataset	DataCatalogue, EMA, CohortStaging	40
Datasets	Definition of a dataset within a (common) data model	dcat:Dataset	DataCatalogue, CohortStaging	11
Documentation	Documentation attached to a resource		DataCatalogue, CohortStaging	6
Individuals	Data, facts or figures about an individual; the set of relevant items would depend on the use case. In FAIR Genomes: Personal. In RD3: Subjects.	http://purl.obolibrary.org/obo/NCIT_C90492, http://purl.obolibrary.org/obo/ExO_0000127	Beacon v2, FAIR Genomes, RD3, JRC-CDE, GDI T8.2 union	50
Mappings	Mapping from collected datasets to standard/harmonised datasets, optionally including ETL syntaxes		DataCatalogue, EMA	9
RWE resources	Real world data collections	dcat:Dataset	DataCatalogue, EMA	107
Subcohort counts	Number of participants per subcohort age group, optionally divided per sex		DataCatalogue, CohortStaging	5

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

samples	Samples used as input for analyses in RD3	http://purl.obolibrary.org/obo/SCDO_0002829	RD3	23
All variables	Generic listing of all source variables. Should not be used directly, please use SourceVariables or RepeatedSourceVariables instead		DataCatalogue, CohortStaging	8
Clinical	Findings and circumstances relating to the examination and treatment of a patient.	http://purl.obolibrary.org/obo/NCIT_C25398	FAIR Genomes, JRC-CDE	19
Data sources	Collections of multiple data banks covering the same population	dcat:Dataset	DataCatalogue, EMA	107
Dataset mappings	Mappings from collected or source datasets to harmonised or target datasets		DataCatalogue, CohortStaging	7
External identifiers	External identifier(s) for this resource		DataCatalogue, EMA, CohortStaging	4
GenomicVariationsCaseLevel	Beacon v2 GenomicVariations Case Level Data	http://purl.obolibrary.org/obo/GSSO_000660	Beacon v2	3
Linked resources	Links between datasource and databank		DataCatalogue, EMA	8
Submissions	Documentation of data submissions related to a resource		DataCatalogue, EMA	11
Variables	Definition of a non-repeated variable, or of the first variable from a repeated range		DataCatalogue, CohortStaging	20
Biosamples	A natural substance derived from living organisms such as cells, tissues, proteins, and DNA. In FAIR Genomes: Material.	http://purl.obolibrary.org/obo/NCIT_C43376	Beacon v2, FAIR Genomes	18
Contacts	Listing of contact persons per resource	vcard:Individual	DataCatalogue, EMA, CohortStaging	14
GenomicVariations	Beacon v2 GenomicVariation	http://purl.obolibrary.org/obo/NCIT_C17248	Beacon v2	14
MetabolomicMaterialProcessing	A metabolomics material processing is a protocol application including material enrollments and biomaterial transformations.	http://purl.obolibrary.org/obo/OBI_0000073	FAIR Genomes metabolomics add-on	8
Repeated variables	Definition of a repeated sourceVariable. Refers to another variable for its definition		DataCatalogue, CohortStaging	9
SamplePreparation	A sample preparation for a nucleic acids sequencing assay.	http://purl.obolibrary.org/obo/OBI_0001902	FAIR Genomes	9
Variable mappings	Mappings from collected variables to standard/harmonised variables, optionally including ETL syntax		DataCatalogue, CohortStaging	12
labinfo	Information or processes in the lab that are linked to samples		RD3	20

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

Databanks	Data collection from real world databases such as health records, registries	dcat:Dataset	DataCatalogue	128
IndividualConsent	Consent given by a patient to a surgical or medical procedure or participation in a study, examination or analysis after achieving an understanding of the relevant medical facts and the risks involved.	http://purl.obolibrary.org/obo/NCIT_C16735	FAIR Genomes, JRC-CDE	12
MetabolomicMassSpecAssays	A qualitative or quantitative analysis performed to determine the amount of a particular constituent in a sample or the biological or pharmacological properties of a drug.	http://purl.obolibrary.org/obo/NCIT_C60819	FAIR Genomes metabolomics add-on	9
Networks	Collaborations of multiple institutions	dcat:Catalog	DataCatalogue, EMA	29
Studies	Collaborations of multiple institutions, addressing research questions using data sources and/or data banks	dcat:Dataset	DataCatalogue, EMA	78
overview	Overview on RD3 Subjects, samples, and experiments	http://purl.obolibrary.org/obo/NCIT_C80271	RD3	20
Catalogues	A collection of resources within a network or consortium or about a common topic	dcat:Catalog	DataCatalogue	3
MetabolomicAnalyses	Basic information, annotation or documentation concerning a metabolomics analysis workflow (but not the workflow itself).	http://edamontology.org/data_0949	FAIR Genomes metabolomics add-on	12
Runs	The determination of complete (typically nucleotide) sequences, including those of genomes (full genome sequencing, de novo sequencing and resequencing), amplicons and transcriptomes. In FAIR Genomes: Sequencing. Ideally: SequencingRuns.	http://edamontology.org/topic_3168	Beacon v2, FAIR Genomes	17
Analyses	An analysis applies analytical (often computational) methods to existing data of a specific type to produce some desired output. In FAIR Genomes: Analysis.	http://edamontology.org/operation_2945	Beacon v2, FAIR Genomes	19
Files	An electronic file is an information content entity which conforms to a specification or format and which is meant to hold data and information in digital form, accessible to software agents. In RD3: files about Individuals are stored in the sandbox environments. Files are linked to subjects, samples, and experiments, as well as with EGA accession number.	http://purl.obolibrary.org/obo/STATO_0000002, http://purl.obolibrary.org/obo/NCIT_C42883	RD3, DCAT files add-on	21

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

Variable values	Listing of categorical value+label definition in case of a categorical variable		DataCatalogue, CohortStaging	9
Distribution	DCAT v2 distribution entries.	http://www.w3.org/ns/dcat#Distribution	DCAT	5
Network variables	Listing of the variables used in a network		DataCatalogue	2

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