

# GENOMED4ALL

---

## D5.1

# Data homogenization requirements and specifications



# GENOMED4ALL

Genomics for Next Generation Healthcare

## D5.1

### Data homogenization requirements and specifications

Revision **v1.0**

Work package	WP5
Task	T5.3
Due date	31-12-2021
Submission date	23-12-2021
Deliverable lead	FORTH
Version	v1.0
Authors	Haridimos Kondylakis (FORTH) Valia Kalokyri (FORTH) Kostas Marias (FORTH) Victor Mateos (DEDALUS) Jose Luis Bravo (DEDALUS)
Reviewers	Tiziana Sanavia (UNITO) Gastone Castellani (UNIBO) Piero Fariselli (UNITO)



## Abstract

Considering the wide range of repositories that GENOMED4ALL will need, homogenizing the data formats from different sources and systems (e.g. EHRs, images, genomic data, etc.) is essential. This deliverable sets the requirements for data homogenization processes paving the way to a common approach for data lake, based on the FHIR standard. FHIR enables almost direct integration with most of the current information systems and automatic data homogenization and enrichment to facilitate data processing and analytics.

## Keywords

Data homogenization, linkage, requirements, specifications



## Document revision history

Version	Date	Description of change	Contributor(s)
v0.1	01-10-2021	1 <sup>st</sup> version of deliverable template	Haridimos Kondylakis (FORTH)
v0.2	01-11-2021	First draft with contributions	Haridimos Kondylakis (FORTH) Valia Kalokyri (FORTH) Kostas Marias (FORTH)
v0.3	10-12-2021	Version ready for internal review by UNITO and UNIBO	Haridimos Kondylakis (FORTH) Valia Kalokyri (FORTH) Kostas Marias (FORTH) Victor Mateos (DEDALUS)
v0.4	23-12-2021	Implementation of comments	Haridimos Kondylakis (FORTH) Valia Kalokyri (FORTH) Kostas Marias (FORTH)
V1.0	23-12-2021	Final general review	Annelore Hermann (UPM)

## Disclaimer

The information, documentation and figures available in this deliverable are provided by the GENOMED4ALL project's consortium under EC grant agreement **101017549** and do not necessarily reflect the views of the European Commission. The European Commission is not liable for any use that may be made of the information contained herein.

## Copyright notice

© GENOMED4ALL 2021-2024

## Project co-funded by the European Commission in the H2020 Programme

### Nature of the deliverable

**R**

### Dissemination level

- PU** Public, fully open. e.g., website
- CL** Classified information as referred to in Commission Decision 2001/844/EC
- CO** Confidential to GENOMED4ALL project and Commission Services



### \* Deliverable types:

- R:** document, report (excluding periodic and final reports).
- DEM:** demonstrator, pilot, prototype, plan designs.
- DEC:** websites, patent filings, press and media actions, videos, etc.
- OTHER:** software, technical diagrams, etc.



# Table of contents

<b>1</b>	<b>Executive summary</b>	<b>8</b>
<b>2</b>	<b>Introduction</b>	<b>9</b>
2.1	Relationship with other WPs	9
<b>3</b>	<b>Overview of the available data</b>	<b>11</b>
<b>4</b>	<b>State of the art on data homogenization</b>	<b>12</b>
4.1	Common Data Models, Ontologies & Terminologies	12
4.2	Data Homogenization Architectures	14
4.3	Genomic-specific data homogenization	15
4.3.1	Existing genomic data repositories	15
<b>5</b>	<b>GENOMED4ALL homogenization requirements and specifications</b>	<b>17</b>
5.1	Data protection and security	18
5.2	Common Data Model & Terminologies	21
5.3	ETL tools	21
5.4	Data Storage	22
5.5	Data Retrieval (for training AI algorithms)	23
5.6	Data quality tools	24
5.7	Data Platform	24
<b>6</b>	<b>GENOMED4ALL homogenization envisioned approach</b>	<b>27</b>
6.1	Data providers models	27
6.2	General data flow	27
6.2.1.1	Data providers to Genomed4All ecosystem	31
6.2.1.2	Data curation	31
6.2.1.3	Data API and ML/AI data formats	32
6.2.1.4	Training ML model	32
6.2.1.5	Share ML model between central server and edge nodes	33
6.3	Data homogenization process definition	33
6.3.1	Structured data	33
6.3.2	Non structured data	34
6.3.3	Images	34
6.3.3.1	Dicom images	34
6.3.3.2	No dicom images	34
6.3.4	VCF files	34
<b>7</b>	<b>Conclusions</b>	<b>38</b>
<b>8</b>	<b>References</b>	<b>39</b>



## List of figures

Figure 1. Contributions to the T5.3 from the GenoMed4All consortium. The clinical team (WP 7) set the requirements and specifications for the format of the input data for the platform. The AI team set the requirements and specifications for the format of the output data, to be used in training the AI models. The engineering and platform workpackages (WP 3,4,8) identify the needs and constraints for the platform architecture and the software implementation of the data model. ....	9
Figure 2. The research platform to be delivered in the context of GENOMED4ALL. ....	17
Figure 3. Big Data Architecture – Simplified version. ....	28
Figure 4. General data workflow, with the interactions between all the components. ....	29
Figure 5. Generic dataflow. Illustrate the component placed in each iteration. ....	30
Figure 6. Data flow from HCP to CDM. ....	31
Figure 7. Data flow for data curation. ....	31
Figure 8. Data API and ML/AI input data generation. ....	32
Figure 9. Training ML model, run AI algorithms. ....	32
Figure 10. Share ML model between Central server and edge to update local copy and edge with Central server ....	33
Figure 11. An example of a VCF file ....	35
Figure 12. Genomics Diagnostic report definition. Illustrate the relationship between a diagnostic report and the different types of (conceptual) genomics observations suggested by HL7 FHIR. ....	36
Figure 13. JSON file created by vcf2fhir tool, with one DiagnosticReport and the Observations with the genomics information ....	37



## List of tables

### Table 1. Requirements template



## Abbreviations

<b>AI</b>	Artificial Intelligence
<b>API</b>	Application Programming Interface
<b>BCO</b>	BioCaster Ontology
<b>CDM</b>	Common data model
<b>CPT</b>	Current Procedural Terminology
<b>csv</b>	Comma Separated Values
<b>DO</b>	Disease Ontology
<b>EGA</b>	European Genomics Archive
<b>EHR</b>	Electronic Health Records
<b>ENIGMA</b>	The Enhancing Neuroimaging Genetics through Meta-Analysis
<b>EU</b>	Europe
<b>ERN</b>	European Reference Network
<b>ETL</b>	Extract Transform Load
<b>FHIR</b>	Fast Healthcare Interoperability Resources
<b>FL</b>	Federated Learning
<b>GO</b>	Gene Ontology
<b>HDs</b>	Hematological diseases
<b>LOD</b>	linked open data
<b>MeSH</b>	Medical Subject Headings
<b>ML</b>	Machine Learning
<b>MM</b>	Multiple Myeloma
<b>NN</b>	Neural Network
<b>NLM</b>	National Library of Medicine
<b>MDS</b>	Myelodysplastic Syndromes
<b>OHDSI</b>	Observational Health Data Sciences and Informatics
<b>OMOP-CDM</b>	Observational Medical Outcomes Partnership Common Data Model
<b>PI</b>	Principal Investigator
<b>PCAWG</b>	The Pan-Cancer Analysis of Whole Genomes
<b>SCD</b>	Sickle Cell Disease
<b>SNOMED CT</b>	Systematized Nomenclature of Medicine Clinical Terms
<b>VCF</b>	Variant Call Format



# 1 Executive summary

The ambition of the project to pool multi-modality data is reflected in the wide variety of **heterogeneous data sources** that will contribute data for the development of AI models. As such, considering the wide range of repositories that GENOMED4ALL will need, homogenizing the data formats from different sources and systems (e.g. EHRs, images, genomic data, etc.) is essential. Consequently, we identify state of the art in health data homogenization and based on the available datasets we describe the requirements for their integration. Data relevant to the project contain not only clinical studies and hospital records but also genomic datasets and datasets available in public repositories, patient registries supported by ERNs and EU health data platforms. The different access modalities and capabilities of each data source led to different integration approaches. Different approaches included, require a consortium member to act as Principal Investigator (PI) in a request for public data, or supporting the integration of platforms like RD-Connect GPAP as federated learning edge nodes in the GenoMed4All platform. As such, besides setting the requirements we also present the basic homogenization workflow envisioned for the platform to be developed.



## 2 Introduction

The work presented in this deliverable is part of the workpackage 5 (WP5) about data engineering, anonymization algorithms, curation and privacy preservation tools. More specifically it is part of the Task 5.3 – Data homogenization which aspires to deliver a data homogenization framework as baseline and prepare all the communication interfaces based on open standards and to carry out the preparation of the platform and definition of the mechanisms to share the data gathered during the project for research purposes. The focus of this deliverable (D5.1) is to define the data homogenization processes to be included in the project and set the requirements and specifications for them.

### 2.1 Relationship with other WPs

This deliverable concludes the work on WP5 for establishing the data homogenization requirements and specification to be included in GENOMED4ALL. It presents overlap with other deliverables, notably from WP3 and WP4. Concerning the interoperability platform and data homogenization, deliverable D3.1 presents an initial plan and a preliminary summary of the general requirements, while D3.3 adds specific requirements leading to a common data model and briefly introduces the data homogenization processes. In this deliverable, we further analyze and explain these processes in more detail. In addition, concerning the federated learning architecture, while deliverables D4.1 and D4.2 present the detailed requirements list and architecture, in this deliverable we try to collect and formalize more systematically these requirements.

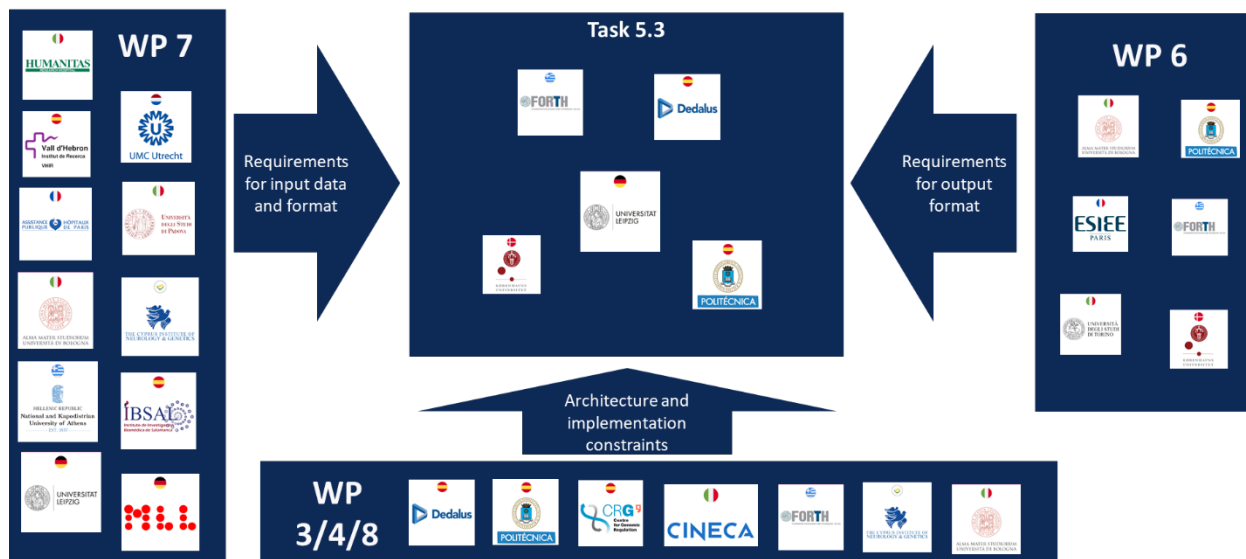


Figure 1. Contributions to the T5.3 from the GenoMed4All consortium. The clinical team (WP 7) set the requirements and specifications for the format of the input data for the platform. The AI team set the requirements and specifications for the format of the output data, to be used in training the AI models. The

engineering and platform workpackages (WP 3,4,8) identify the needs and constraints for the platform architecture and the software implementation of the data model.

Collecting the data homogenization requirements and specifications and defining the data homogenization process within WP 5 have required a collective effort and collaboration among many interested parties within the consortium. Figure 1 shows how individual partners, grouped by their respective WPs, contributed to this task. Technical experts from WP3 provided a continuous discussion on the common data model aspects for the data homogenization and weekly teleconferences were arranged with key partners from WP3 and WP5. Technical experts from WP4 provided guidance on the requirements and constraints for the implementation of the data model in a federated setting. WP6 experts in AI modeling provided the requirements from the perspective of the “end-user” of the data model, since ultimately the data managed by the GenoMed4All platform will be used for the federated training of the AI models. WP7 clinical data providers shared their requirements from the perspective of the “data sources”, by identifying the data format with the standardized terminologies, and by sharing metadata about the structure of their datasets. Finally, WP8 experts provided guidance on the modeling of -omics data, their standards and the best practices.



### 3 Overview of the available data

In this section, we present an overview of the list of available data repositories within and outside the GENOMED4ALL consortium. The types of data and their formats are analyzed in detail in deliverable D3.3, but here we present a short introduction for completeness and to enable a full understanding of the complexity and heterogeneity of data types and sources used in the project.

The most basic types of health data analyzed in GenoMed4All, which are common to all disease use cases, are demographic and clinical data. Demographic data are highly sensitive since they can be used to easily identify a person, such as birth, sex, address, socio-economic and health status, etc. Most demographic data will be completely redacted according to the aims of the project, however highly influential information such as age or gender will be included, making sure that the proper safeguards are put in place to prevent simple means of reidentification. Clinical data contain information about the disease manifestations, treatment, and the general care path followed by the patient. Only the data considered relevant for the AI modeling will be included in the GenoMed4All datasets, like for example the types of treatment, the responses, the main clinical manifestations of the disease, etc. Of course, it should be noted that the precise makeup of each dataset will be influenced by the clinical question that will be addressed through the AI analysis.

Another common data type is results from laboratory tests and assays. Such information is the most prevalent both in our current and in future datasets which will be used in GenoMed4All, since this data type is inherently well-structured and highly informative of a patient's pathological status. A subset of laboratory data particularly relevant to the analysis of disease use cases is represented by the hematological data, which include the results of all laboratory blood tests derived from each patient. Finally, complex laboratory assays such as the novel Oxygenscan technique [10] will constitute part of the available laboratory data.

A core constituent of any dataset in GenoMed4All is represented by the -omics data, which are high-dimensional data arising from the analysis of an individual's genome, or transcriptome, or other associated molecular profiles representing his/her phenotype. Here, the term phenotype is used to denote any clinical manifestation of the disease that can be reasonably attributed to an underlying genetic abnormality. Omics data may include genomics, i.e. the study of an individual's genome, transcriptomics, i.e. the study of an individual's gene expressions through RNA, proteomics, i.e. the study of an individual's expressed proteins and their interactions, metabolomics, i.e. the study of the whole set of metabolites, radiomics, i.e. the study of the imaging data related to an individual, and other fields of application. Such large amounts of high-dimensional data often require specialized techniques for storage and processing, making them one of the focuses of the whole project and a significant challenge for the development of novel techniques.

## 4 State of the art on data homogenization

### 4.1 Common Data Models, Ontologies & Terminologies

Big amounts of data are currently stored in various silos within health and social care systems (volume). Although the semantic web and the linked open data (LOD) paradigms provide the means to open, connect and integrate the available data, still a huge amount of them exists in many different formats, ranging from textual documents and web tables to well-defined relational data and APIs (variety). In addition, the data pertain to ambiguous semantics and quality standards resulting from different collection processes across sites (veracity). The large amount of data generated and collected today comes in so many different streams and forms — from carer notes, personal health records, images, sounds, videos, health conversations in social media (variability), to continuous streaming information collected from wearables and other monitoring devices (velocity).

Semantic Integration is the problem of providing unified and transparent access to a collection of data stored in multiple, autonomous and heterogeneous data sources using semantic models. During the last years, ontologies and common data models have been used to integrate structured and semi-structured data[1]. However, there is not a single correct way to model a domain and several ontologies exist. For example, the Symptom Ontology<sup>1</sup>, and the Disease Ontology<sup>2</sup> (DO) link disparate datasets through symptoms and disease concepts; the Foundational Model of Anatomy<sup>3</sup> collects the phenotypic structure of the human body, whereas Adverse Event Ontology [5] tries to model adverse events. The Experimental Factor Ontology focuses on experimental variables in Gene Expression Atlas<sup>4</sup>, the Clinical Care Classification System<sup>5</sup> tries to code health care settings and the Current Procedural Terminology (CPT)<sup>6</sup> is a medical nomenclature used to report medical procedures and services under public and private health insurance programs. UMLS<sup>7</sup>, the Unified Medical Language System, is a unifying framework that integrates different terminologies which are relevant to medicine and biomedical information technologies. The Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) is a clinical terminology, which has been promoted as a reference terminology for electronic health record (EHR) systems. SNOMED CT is used by the College of American Pathologists<sup>8</sup>; the UMLS Metathesaurus<sup>9</sup>, the

<sup>1</sup> [http://symptomontologywiki.igs.umaryland.edu/wiki/index.php/Main\\_Page](http://symptomontologywiki.igs.umaryland.edu/wiki/index.php/Main_Page)

<sup>2</sup> [http://www.obofoundry.org/cgi-bin/detail.cgi?id=disease\\_ontology](http://www.obofoundry.org/cgi-bin/detail.cgi?id=disease_ontology)

<sup>3</sup> <http://sig.biostr.washington.edu/projects/fm/AboutFM.html>

<sup>4</sup> <http://www.ebi.ac.uk/gxa/>

<sup>5</sup> [http://en.wikipedia.org/wiki/Clinical\\_Care\\_Classification\\_System](http://en.wikipedia.org/wiki/Clinical_Care_Classification_System)

<sup>6</sup> <http://www.ama-assn.org/ama/pub/physician-resources/solutions-managing-your-practice/coding-billing-insurance/cpt.page>

<sup>7</sup> <http://www.nlm.nih.gov/research/umls/>

<sup>8</sup> <http://www.cap.org/apps/cap.portal>

<sup>9</sup> <http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html>

European project epSOS<sup>10</sup> and the European project SemanticHealthNet<sup>11</sup>. The Medical Subject Headings (MeSH)<sup>12</sup> database is a medical thesaurus published and annually updated by the US National Library of Medicine (NLM). It is used for cataloguing the library holdings and for indexing the databases that are produced by the NLM (e.g. MEDLINE). ACGT MO<sup>13</sup> tries to model medical knowledge in the Cancer domain. The International Classification of Diseases<sup>14</sup> is the world's standard tool to capture mortality and morbidity data. LOINC<sup>15</sup> is a database and a universal standard for identifying medical laboratory and clinical observations, while the Medical Dictionary for Regulatory Activities<sup>16</sup> (MEDRA) is a clinically validated international medical terminology for diagnoses, symptoms, surgeries and other medical procedures. The Thesaurus of the National Cancer Institute (NCI)<sup>17</sup> covers vocabulary for clinical care, translational and basic research, public information and administrative activities. Moreover, other ontologies try to model multiscale data such as the Systems Biology Ontology<sup>18</sup> and the Gene Ontology (GO)<sup>19</sup>, which support biologically meaningful annotation of genes and their products in different databases. Besides these ontologies that refer to core medical and biological knowledge, other ontologies try to mostly cover the domain of social entities that are related to health care such as Ontology of Medically Related Social Entities<sup>20</sup> and the BioCaster Ontology [2] (BCO), which collects the terms and relations necessary to detect and assess public health events. The FHFO [8] represents the family health histories of persons related by biological and/or social family relationships (e.g. step, adoptive) who share genetic, behavioural, and/or environmental risk factors for disease.

The exact terminologies/ontologies that will be used in the GENOMED4ALL project will soon be concluded based on the data that will gradually become available. An initial discussion is already presented in D3.4 and the interested reader is forwarded there for more information.

Besides all these ontologies and terminologies, more structured generic common health data models have been proposed such as FHIR and OMOP-CDM

**FHIR** is a universal standard for the exchange of patients' health records from EHR systems. It is composed of so-called resources, bundled in profiles that specify how a particular information element should be represented. There are around 100 resources, including diagnostics, medication, physical measurements, etc. FHIR distinguishes itself by having a description of the

<sup>10</sup> <http://www.epsos.eu/>

<sup>11</sup> <http://www.semantichealthnet.eu/>

<sup>12</sup> <http://www.ncbi.nlm.nih.gov/mesh>

<sup>13</sup> <http://bioportal.bioontology.org/ontologies/1126>

<sup>14</sup> <http://www.who.int/classifications/icd/en/>

<sup>15</sup> <http://en.wikipedia.org/wiki/LOINC>

<sup>16</sup> <http://www.meddra.org/>

<sup>17</sup> <http://ncit.nci.nih.gov/>

<sup>18</sup> <http://www.ebi.ac.uk/sbo/main/>

<sup>19</sup> <http://www.geneontology.org/GO.consortiumlist.shtml>

<sup>20</sup> <http://omrse.googlecode.com/svn/trunk/omrse/omrse.owl>

API that exposes the resources from the underlying EHR system. This allows other applications to (securely) pull data from and push data to the EHR systems. There are multiple sandboxes to explore the standards. In contrast to many other standards, FHIR is not the endpoint of the data. Rather than persisting data in this standard, it facilitates data transfer to other applications using the resource profiles and API definition. EHR systems, like openEHR, can use FHIR to expose data to other applications. These can be consumer apps that make health data insightful to patients or research platforms like OHDSI.

**OMOP-CDM** is one of the most widely used common data models for supporting analysis of observational health data, enabling the generation of reliable scientific evidence about the disease history, the effects of medical interventions and the healthcare interventions and outcomes. Besides the standard CDM, OMOP-CDM extensions are used, such as the Oncology CDM extension for representing cancer data at the levels of granularity and abstraction required to support cancer research. Although the radiological exams can be currently registered using the OMOP-CDM, the model does not enable the storage of their subsequent curation process.

## 4.2 Data Homogenization Architectures

To integrate the available information, numerous approaches have been developed, either centralized, where data are stored locally (e.g. data warehouses), or federated, where data are left at the sources and accessed on demand.

The benefit of the federated architectures is that the data sources are controlled locally by their owners and the regulations and policy restrictions can be enforced appropriately without the data leaving the local premises, which in many cases is required by hospitals and data owners. However, the federated architectures complicate data access and retrieval as the data are not available in a central point and should be queried in a distributed fashion. In addition, this requires distributed nodes to have efficient processing and storage “power” as the data should be locally stored and served. Federated nodes can also comply with a common data model to speed up query answering and to enable data homogenization.

On the other hand, centralized architectures enjoy the benefit of efficiency as all data are centrally available and are ready to be queried and analyzed. However, storing sensitive health data centrally requires compliance with certain regulations. Therefore, the extract-transform-load workflow should be appropriately managed in order to enjoy query efficiency.

Besides these two mainstreams, there are also hybrid approaches, where some data are stored centrally and some other data are stored in a federated manner enjoying both the benefits (but suffering also from the problems) of both the architectures.

## 4.3 Genomic-specific data homogenization

The majority of existing formats for describing genotype information does not include a means to share corresponding phenotypic information (e.g. observable characteristics, signs/symptoms of disease). The lack of uniformity amongst the genomic databases (with their format for representing the phenotypic information) hinders communication and limits the ability to perform further analyses across them. The Global Alliance for Genomics and Health Steering Committee approved Phenopackets<sup>21</sup> in October 2019, a standard file format for sharing phenotypic information. The Phenopackets standard aims to facilitate the communication between the research and clinical genomics communities by creating an ecosystem of interoperable tools and resources that can use phenotypic data with fewer barriers. Using Phenopackets, clinicians can search through genetic variants that produce similar phenotypes and they can determine which one best matches their patients. Overall, such matching supports better and faster diagnosis and treatment, and higher chances of remission. Phenopackets also benefit researchers by opening up opportunities to analyse more data and strengthen our understanding of human health and disease.

### 4.3.1 Existing genomic data repositories

**RD-Connect GPAP**<sup>22</sup> currently stores genomics and phenotypic information of about 12000 participants with Rare Diseases. The platform allows real-time analysis combining genetic variants and clinical information. The experiments, submitted to the platform, can have a 6 months embargo period and, after this timeframe, all the experiments will be visible to the platform users.

**European Genomics Archive (EGA)**<sup>23</sup> is a long-term repository for genomics data, consented for specific uses but not for open public distribution. The EGA follows strict protocols for information management, data storage, security and dissemination. EGA is a worldwide reference service for the management of sensitive biological data that allows the identification of citizens who donate such data (Personal Identifying Information). EGA contains more than 2 million files generated in institutions of 39 countries from all continents, for a total of more than 8 PB of data. These data are guarded and distributed to thousands of scientists from 57 countries. During 2019, the EGA has distributed more than 5 PB to research groups to carry out projects that without such data would have been impossible.

**The Pan-Cancer Analysis of Whole Genomes (PCAWG)**<sup>24</sup> study is an international collaboration to find common patterns of mutations in more than 2,600 cancer whole genomes from the International Cancer Genome Consortium. All the data are collected in a central repository but data analysis can be done either in a collaborative cloud or locally.

<sup>21</sup> <http://phenopackets.org/>

<sup>22</sup> <https://platform.rd-connect.eu>

<sup>23</sup> <https://ega-archive.org/>

<sup>24</sup> <https://www.embl.de/campaigns/pancancer/>



**The Enhancing Neuroimaging Genetics through Meta-Analysis (ENIGMA)** Consortium is a collaborative network of researchers working together on a range of large-scale studies that integrate data from more than 70 institutions worldwide. It has established more than 50 working groups (WGs), pooling worldwide data, resources and expertise to answer fundamental questions in neuroscience, psychiatry, neurology, and genetics.



## 5 GENOMED4ALL homogenization requirements and specifications

The development of the GENOMED4ALL platform will be carried out in two phases (see D3.1 and D3.2 for more details), a) a preliminary development phase where data is centralized, and b) a final federated phase that will eventually formulate the research platform shown in Figure 2

### Research Framework

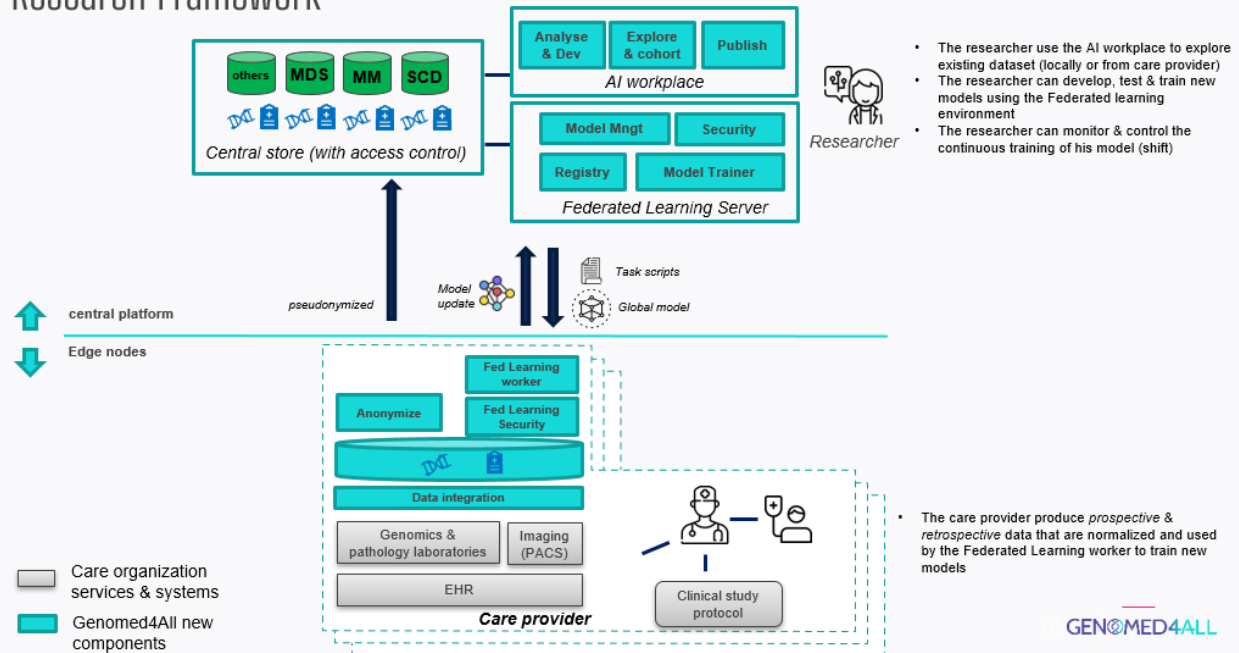


Figure 2. The research platform to be delivered in the context of GENOMED4ALL.

During the **preliminary development phase**, the data will be shared centrally with the WP6 contributors, allowing them to develop the AI algorithms locally before generalizing the training to a federated approach. In this phase, a centralized data storage repository will be maintained to enable a secure, trusted and GDPR-compliant sharing of data. As such the data should be **centrally-stored in a standardized, pre-defined format**, using a common data model (CDM). Data represented in this way will be easily auditable, searchable, and transformable into the format required by each AI algorithm.

During the **federated phase**, a centralized storage facility will no longer be required as each clinical data provider (acting as a federated learning edge node) will host only their private portion of the dataset locally. Only the AI models will be transmitted over the network, guaranteeing a high level of privacy and security. In the federated phase, a CDM becomes a necessity to ensure that different centres contribute to the same dataset in a standardized and interoperable way.

In agreement with the aforementioned ideas, it is essential to set the homogenization requirements for such a platform. For an effective compilation and consideration of requirements, each of them has been specified according to the following fixed format table:

Table 1. Requirements template

Requirement id:	Requirement name:
<b>Definition</b>	This field contains the specification of the requirement (description of the purpose and goals to be fulfilled), written in a preferably concise, yet clear way. At this point, one should be very specific as to which is the goal of this requirement and envisioned benefit. E.g., The gateway must support different data sources.
<b>Reference Functionality</b>	This field contains information on the GENOMED4ALL components and functionalities this requirement refers to.
<b>Success Criteria</b>	This field contains information on how to assess the fulfilment of this requirement.
<b>Requirement Dependences</b>	This field lists (the corresponding codes) of other requirements on which the specific one depends.
<b>Priority</b>	This element specifies the criticality of the requirement and it can take the values COULD for optional requirements, SHOULD for desirable, MUST for mandatory (in ascending order).

The fields are as follows:

- **Requirement ID:** This field provides a unique code to exclusively identify each requirement and easily tracking its fulfilment in the next steps of the project. This field has the following generic format: TYPE-RQT# where RQT# is a unique identifier of the requirement composed of an optional text string and a sequence of digits, while TYPE has the following values:
  - FUNC – functional requirement.
  - POL – non-functional policy requirement.
  - DAT – non-functional data requirement.
  - SP – non-functional security & privacy requirement.
  - OTH – other non-functional requirement.

In the sequel, we will try to capture and explain the requirements for developing a **scalable platform**, capable of integrating additional data sources in a seamless manner.

## 5.1 Data protection and security

All data available data should be appropriately anonymized respecting national, European and international requirements and norms:

Requirement id: SP-1	Requirement name: Data Anonymization
<b>Definition</b>	All data should be appropriately anonymized both at the edges and at the central nodes.
<b>Reference Functionality</b>	Data anonymization tools
<b>Success Criteria</b>	Data are appropriately anonymized

<b>Requirement Dependences</b>	
<b>Priority</b>	MUST

In addition, the access to the information should be regulated based on authentication/authorization mechanisms and role-based access.

<b>Requirement id: SP-2</b>	<b>Requirement name: Authentication/Authorization</b>
<b>Definition</b>	Proper authentication and authorization mechanisms should ensure regulated access to the data
<b>Reference Functionality</b>	Authentication & Authorization tools
<b>Success Criteria</b>	Access to the data is effectively regulated.
<b>Requirement Dependences</b>	
<b>Priority</b>	MUST

<b>Requirement id: SP-3</b>	<b>Requirement name: Role Based Access</b>
<b>Definition</b>	Based on the role of the user, appropriate access to relevant data should be provided
<b>Reference Functionality</b>	Authentication & Authorization tools
<b>Success Criteria</b>	Access granted based on the roles.
<b>Requirement Dependences</b>	Authentication/Authorization mechanisms
<b>Priority</b>	MUST

<b>Requirement id: SP-4</b>	<b>Requirement name: Restricting Data Access for the Federated Phase</b>
<b>Definition</b>	During the federated phase of the project, it should not be possible to access individual patient's data through the platform outside of the institution that owns the patient's data
<b>Reference Functionality</b>	Querying tools and Data Access APIs
<b>Success Criteria</b>	Data Accessible only to the institution who owns patient data
<b>Requirement Dependences</b>	Authentication/Authorization mechanisms
<b>Priority</b>	MUST

<b>Requirement id: SP-5</b>	<b>Requirement name: Restricting Data Access to Individuals</b>
---------------------------------	---

<b>Definition</b>	We never display individual patient's data to humans in a visualization. Federated queries returning aggregated statistical information are allowed. The risk of extracting individual patient samples is minimized.
<b>Reference Functionality</b>	Querying tools and Data Access APIs
<b>Success Criteria</b>	Individual patients' data are not accessible.
<b>Requirement Dependences</b>	Authentication/Authorization mechanisms
<b>Priority</b>	MUST

<b>Requirement id:</b> DAT-1	<b>Requirement name:</b> Data Logging
<b>Definition</b>	The infrastructure should keep a log file of all accesses to the data, to the errors and to the cleaning processes applied.
<b>Reference Functionality</b>	Data Cleaning, Data Access
<b>Success Criteria</b>	Successful logging of relevant information
<b>Requirement Dependences</b>	Authentication/Authorization mechanisms, Data Access tools and APIs
<b>Priority</b>	MUST

<b>Requirement id:</b> SP-6	<b>Requirement name:</b> Compliance
<b>Definition</b>	The data infrastructure should be compliant to relevant EU and national legal frameworks, with specific regard to the Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons concerning the processing of personal data and the free movement of such data (General Data Protection Regulation, 'GDPR' or 'Regulation') and relevant national legislations, including de-identification (i.e., pseudonymization) procedures, technical requirements for data security within the GENOMED4ALL infrastructure, definition of access roles, data governance, and legal roles and liabilities (i.e., data controllers, data processors).
<b>Reference Functionality</b>	Data platform
<b>Success Criteria</b>	The GENOMED4ALL research platform compliant with national and international laws and regulations.
<b>Requirement Dependences</b>	
<b>Priority</b>	MUST

## 5.2 Common Data Model & Terminologies

A common data model, unified across disease areas, protects AI developers from the combinatorial explosion of terminologies, standards and data representations from the source data providers and enables semantic homogeneity and interoperability

Requirement id: FUNC-1	Requirement name: Common Data Model
<b>Definition</b>	<p>The data from different clinical sources is highly heterogeneous. The data model must harmonize this heterogeneity.</p> <p>The data model, and especially the transformations to and from the CDM, must be designed to ensure the highest data quality for training the AI models.</p> <p>The data model shall contain at the very least all the information in the raw source data. The data model may only change the organization of the information and enrich it with additional details (after clinical validation).</p>
<b>Reference Functionality</b>	Common Data Model
<b>Success Criteria</b>	The available Common Data Model is able to model all data relevant to the GENOMED4ALL project
<b>Requirement Dependences</b>	
<b>Priority</b>	MUST

Requirement id: FUNC-2	Requirement name: Common Terminologies
<b>Definition</b>	The data model shall support the use of the same terminologies and ontologies as the raw source data.
<b>Reference Functionality</b>	Common Terminologies
<b>Success Criteria</b>	The terminologies used in the various available datasets are supported by the Common Data Model and the accompanying set of terminologies used.
<b>Requirement Dependences</b>	Common Data Model
<b>Priority</b>	MUST

## 5.3 ETL tools

Having available the raw data and the common data model, an appropriate extract-transform-load workflow should be established accompanied by the relevant tools doing the actual data transformation to the common data model.

Requirement id: FUNC-3	Requirement name: Tools for Data Mapping
---------------------------	---



<b>Definition</b>	The necessary mapping tools should be available for establishing the correspondences among the raw data, the common data model and the terminologies used. The complexity of the data mapping should be kept as close as possible to the source data
<b>Reference Functionality</b>	Data Mapping
<b>Success Criteria</b>	The available raw data mapped to the CDM and to the terminologies proposed by the GENOMED4ALL project
<b>Requirement Dependencies</b>	Common Data Model, Common Terminologies
<b>Priority</b>	MUST

<b>Requirement id:</b> <b>FUNC-4</b>	<b>Requirement name:</b> <b>Tools for ETL</b>
<b>Definition</b>	The necessary tools for extracting, transforming and loading the data to the data lake/repository should be available.
<b>Reference Functionality</b>	ETL
<b>Success Criteria</b>	The data transformed and loaded to the data lake/repository of the project
<b>Requirement Dependencies</b>	Common Data Model, Common Terminologies, Data Mappings, Data Mapping Tools
<b>Priority</b>	MUST

<b>Requirement id:</b> <b>FUNC-5</b>	<b>Requirement name:</b> <b>Support for data modifications</b>
<b>Definition</b>	Data is updated infrequently but can be updated in case of input errors as well as patient follow-up in longitudinal studies.
<b>Reference Functionality</b>	ETL
<b>Success Criteria</b>	The updated data transformed and loaded to the data lake/repository of the project
<b>Requirement Dependencies</b>	Common Data Model, Common Terminologies, Data Mappings, Data Mapping Tools
<b>Priority</b>	SHOULD

## 5.4 Data Storage

The data used through the lifetime of the GENOMED4ALL project should be stored in repositories either locally or centrally following a common data model.

<b>Requirement id:</b> <b>FUNC-5</b>	<b>Requirement name:</b> <b>Data Lake/Data Repository</b>
<b>Definition</b>	The transformed data should be available in a data repository ready to be queried.

Reference Functionality	Data storage
Success Criteria	The transformed data, following the common data model stored in either the federated or the central repository of the project.
Requirement Dependences	Common Data Model, Common Terminologies, Data Mappings, Data Mapping Tools, ETL tools
Priority	MUST

Requirement id: FUNC-6	Requirement name: Data Cleaning
Definition	Appropriate processes for data cleaning, based on the properties of each dataset, must be defined and implemented. These processes should be configurable, in order to be applied properly and efficiently for the various data sources (e.g. specific sensors may provide indicators on the validity/quality of the stored measurements).
Reference Functionality	Data Cleaning tools
Success Criteria	Proper adaptation of the data cleaning process for each dataset ensuring that the invalid data are not considered.
Requirement Dependences	Data repository
Priority	MUST

## 5.5 Data Retrieval (for training AI algorithms)

The AI models are going to process the data as numerical values in a tabular format. This implies that non-numerical data such as coded terms, categorical variables, genomics, and free text need to be converted into zeroes and ones and to be inserted into a table, prior to being provided as input to the training algorithm. **WP6 has expressed the requirement that the format of the source data should be completely transparent to the AI model**, which will be trained on clean and wrangled data in the numerical, tabular format expressed above.

Requirement id: FUNC-5	Requirement name: Data Querying
Definition	The infrastructure should provide an appropriate API for performing analytical queries based on the available data. Queries should be based on the available common data model and the terminologies selected and used. Specific attention will be shown in querying and transforming non-numerical data features.
Reference Functionality	Data Querying Tools & APIs
Success Criteria	Timely and correct access to the data is ensured
Requirement Dependences	Data Repository
Priority	MUST

Requirement id: FUNC-6	Requirement name: Enabling Federated Learning
Definition	GENOMED4ALL will innovate compared to existing large repositories because it will be based on a federated approach, in this way it will overcome existing barriers of a fully centralized repository, such as legal barriers to transfer genomics data and clinical data out of the hospitals/region/country.
Reference Functionality	Federated Learning Infrastructure
Success Criteria	Federated Learning algorithm have access to the available data
Requirement Dependences	Data repository, Data querying mechanisms
Priority	MUST

Requirement id: FUNC-7	Requirement name: Metadata
Definition	Metadata should be always available for exploration to platform subscribers.
Reference Functionality	Data Querying Tools & APIs
Success Criteria	Metadata provided to the interested parties
Requirement Dependences	Data Repository
Priority	MUST

## 5.6 Data quality tools

Further, the GenoMed4All architecture should allow automatic ways to assess and monitor the quality of data

Requirement id: FUNC-8	Requirement name: Data quality tools
Definition	Tools should be available to enable the automatic assessment and monitoring of the quality of the data
Reference Functionality	Data Querying Tools & APIs
Success Criteria	Data quality descriptions on the offered data are available.
Requirement Dependences	Data Repository
Priority	SHOULD

## 5.7 Data Platform

Finally, there are some additional requirements for the data platform as a whole.



Requirement id: FUNC-9	Requirement name: Multiple data architectures
Definition	GENOMED4ALL should support centralized, hybrid and federated data architecture
Reference Functionality	N/A
Success Criteria	Various instances of the platform are available supporting both centralized, hybrid and federated nodes.
Requirement Dependencies	Data Repository, Data Access Tools
Priority	MUST

Requirement id: FUNC-10	Requirement name: Local Deployment
Definition	Local independent deployment of the platform must be supported for the care providers.
Reference Functionality	N/A
Success Criteria	Independent functional deployments of the platform can be created through a specified process.
Requirement Dependencies	Data Repository, Data Access Tools
Priority	MUST

The available data sources should be interoperable and linked with additional external resources such as EU data platforms, registries, and public repositories, in order to enrich data with further information and enhance the potentialities of federated learning. Among others, GENOMED4ALL should be connected to existing repositories, such as RDConnect GPAP and EGA, in this way it will connect thousands of genomic and clinical data from EurobloodNet to other rare diseases ERN datasets (NMD - Neuromuscular diseases, RND - Rare Neurological diseases, Genturis - Rare genetic tumor risk syndromes). It will be also linked to the European Joint Programme on Rare Diseases, in this way it will be connected to European rare diseases infrastructure that it is currently built.

Requirement id: FUNC-11	Requirement name: External Data Sources
Definition	In order to ensure greater availability of information sources, connectivity with external clouds and services for data acquisition should be pursued. The infrastructure should be able to pull data from external data sources through REST APIs.
Reference Functionality	N/A
Success Criteria	Data pulling from external sources is supported.

Requirement Dependences	ETL and mapping tools
Priority	MUST

Requirement id: FUNC-12	Requirement name: FAIR
Definition	In order to ensure Findability, Accessibility, Interoperability and Reuse (FAIR) the consortium should ensure that the appropriate <b>metadata</b> are available to make the sources easily findable, <b>common data models</b> to enhance interoperability
Reference Functionality	N/A
Success Criteria	Independent functional deployments of the platform can be created through a specified process.
Requirement Dependences	N/A
Priority	MUST



## 6 GENOMED4ALL homogenization envisioned approach

The Homogenization process will consist in the extraction, transformation, and load of the data from the different data providers to a common data storage model. This chapter presents the strategy to perform this process.

### 6.1 Data providers models

Due to the diversity of data providers, we will find a heterogeneous data ecosystem, with several data formats (images, reports, csv files, relational database, etc..), data types and data models. In most of the cases, these are unstructured data, which will complicate both data access and data transformation processes.

Another important challenge is the security and legacy in data access. It won't be possible, in some of the cases, to have full access to the data in real time. This will require different approaches to the data that will depend on each specific data provider. In some cases, data will be provided in an asynchronous mode and decoupled from the healthcare data system, to avoid direct access to the data.

In the deliverable 3.1 there is an introduction to data type and data repository, that will be covered in depth in deliverable 3.3.

### 6.2 General data flow

The OHC Digital Health Platform **includes a set of components** – some open source, some vendor, some DXC developed IP – that are orchestrated together. Components that will be used for the GENOMED4ALL project are:

1. **Integrated Record on ELASTIC:** Our storage layer for the integrated, longitudinal care record consolidating data from a broad range of systems of record. Standardized to HL7's FHIR resource specifications and based on the FHIR Data Model.
2. **API Director:** It helps organizations to create new secure channels for accessing patient data, exposing information in a controlled and managed way. It publishes a catalogue of interface end-points.
3. **AI Ready adapter:** It allows data from the integrated records to be extracted and optimized for consumption by AI and BI platforms.

## Big Data Architecture - Simplified version

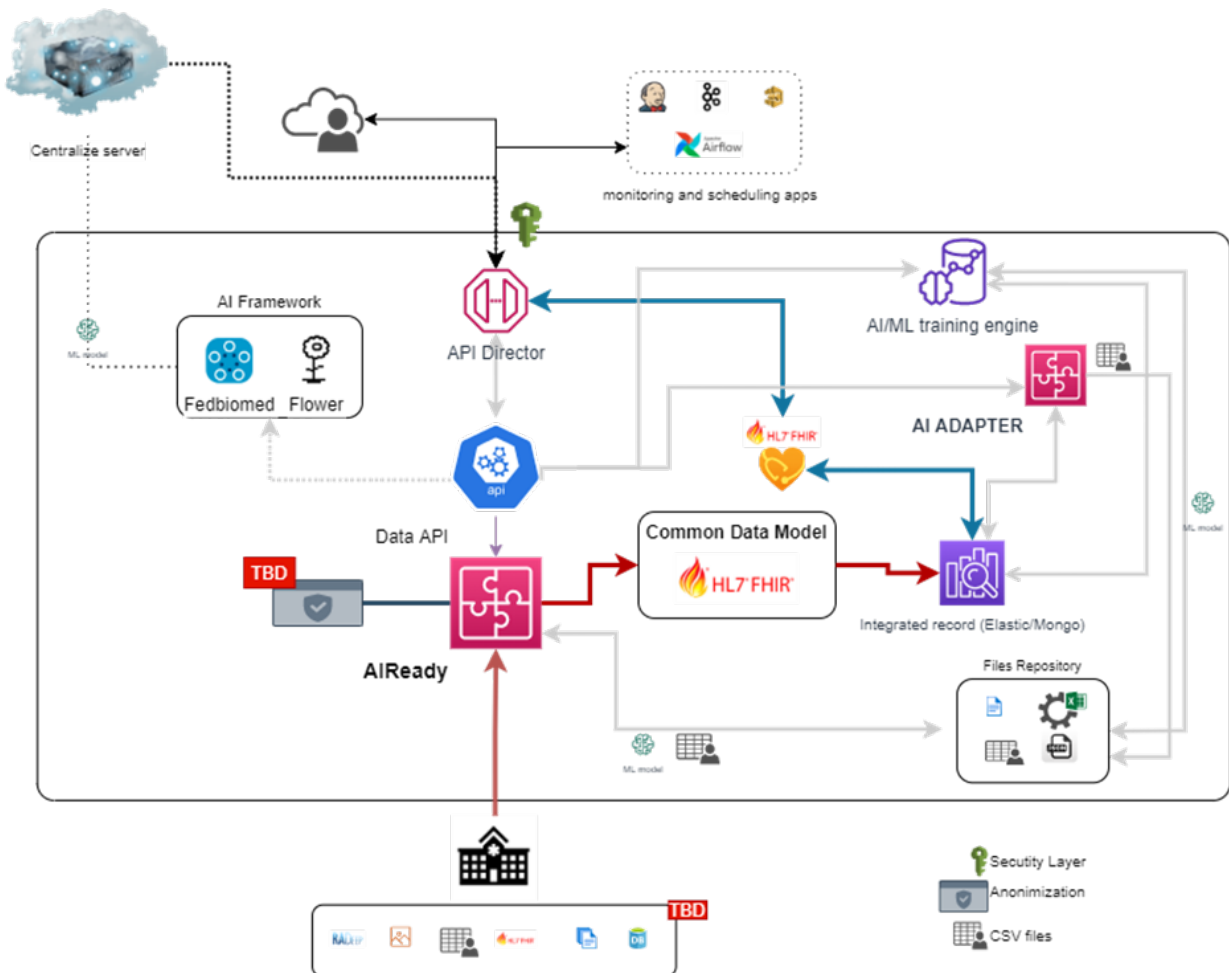


Figure 3. Big Data Architecture – Simplified version.

The data flows will define, at a high level, how the data will be moved between several components implied in the data homogenization process.

As a general overview, healthcare providers will provide data in some kind of format, extracted by themselves. This approach will guarantee privacy and security access to the data, avoiding direct access to the healthcare data. Each data provider will provide the data definition to make possible the data homogenization process.

From the provider data, an ETL tool will read the data, run a transformation pipeline that will transform source data into the common data model definition (FHIR resources), and load the resources in a data integrated record (Elasticsearch or mongo).

As FHIR [6] is going to be the common data model, a FHIR server could be available to make possible the access to these data.

Once data is homogenized and it's available in the integrated record, next processes for data curation, AI/ML data pre-processing, AI/ML input file generation (csv) and so on, can be launched.

*This data flow and the data formats is susceptible to change in the future, once the other WPs involved in the project release each deliverable.*

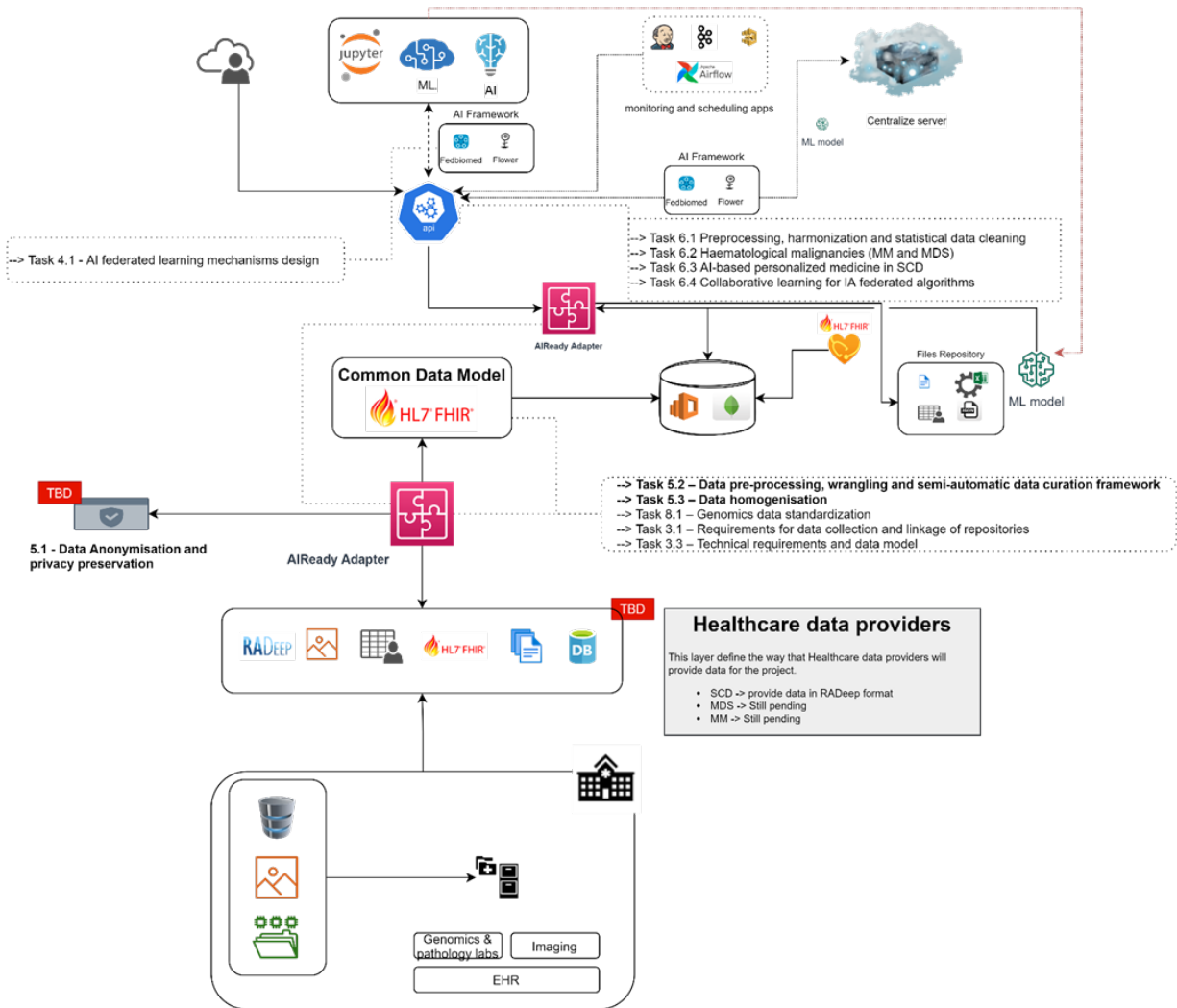
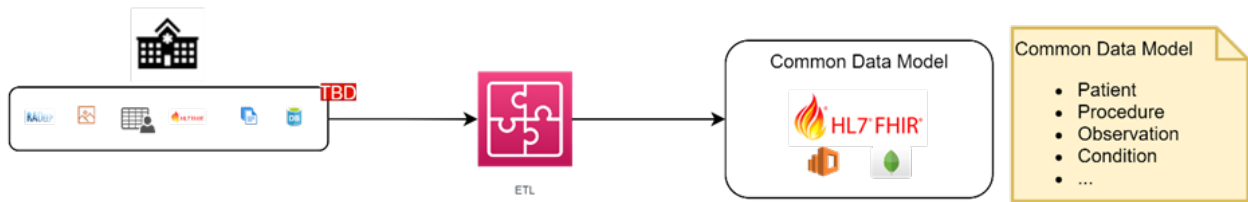
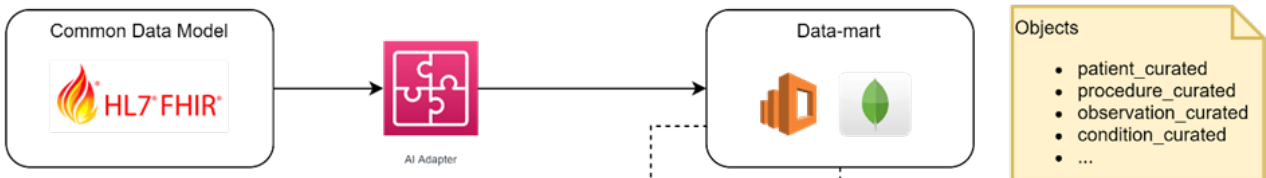


Figure 4. General data workflow, with the interactions between all the components.

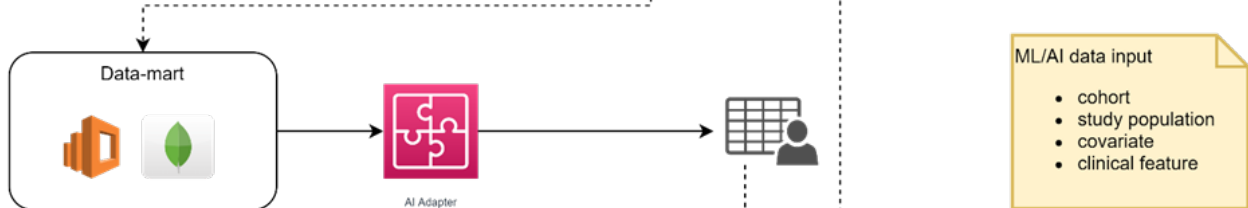
## 1) Data provider to Common Data Model - Data Homogenization



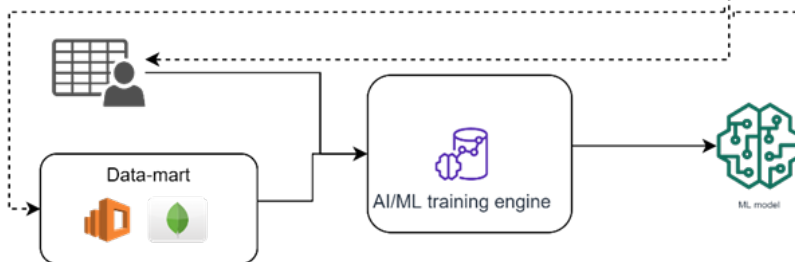
## 2) Preprocessing Common data model - Data Curation



## 3) Data API will define a common standard for ML/AI algorithms an Data Analyst to create training data input format



## 4) Model training, run AI algorithms, Data analyst



## 4) Upload/download model to/from Central server

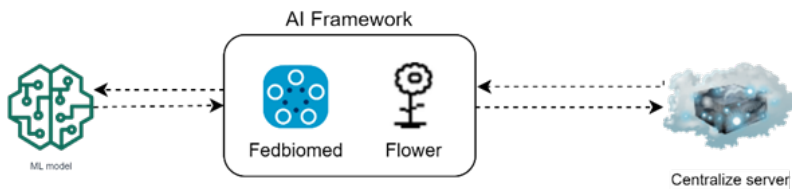


Figure 5. Generic dataflow. Illustrate the component placed in each iteration.

### 6.2.1.1 Data providers to Genomed4All ecosystem

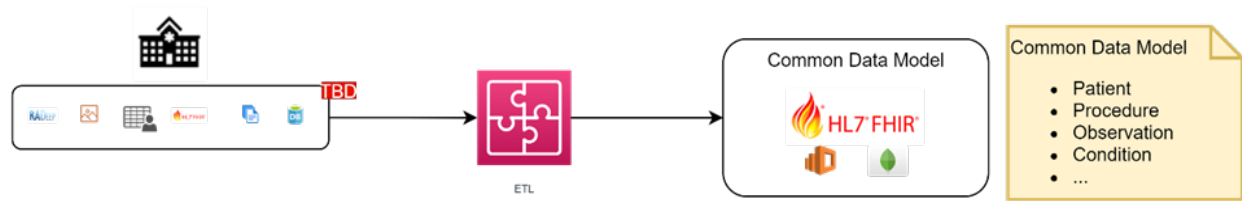


Figure 6. Data flow from HCP to CDM.

Data will be read from data healthcare providers, and data will be transformed into a common data model (defined in D3.3) using ETL techniques. The transformation will include the anonymization/pseudo-anonymization of the data to preserve patient identities.

As the data formats and types will be heterogeneous, it will require a customized analysis in each data provider, to understand each ecosystem and create specific ETL processes. For SCD, data input model will follow RADep definition [9]. MM and MDS data input model/format are still pending to be defined.

The data-mart will store data in a homogenized data model, common and reusable in all the use cases, that will facilitate data processing and analytics. This homogenization process will follow the guidelines defined in Task 8.1 – Genomics data standardization, for the standardization of Genomics data.

### 6.2.1.2 Data curation

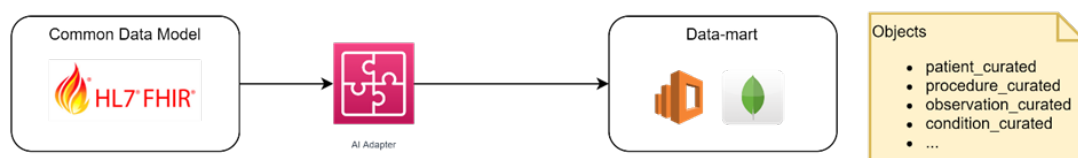


Figure 7. Data flow for data curation.

Once data has been stored in the data-mart, data curation process will be run. The main objective is analyzing data, curating the information, and generating structured meaningful data. The goal is that the generated information will be enough to create the data set and data model for the ML/AI algorithms, and provide enough information to the Data Analyst to process the information.

### 6.2.1.3 Data API and ML/AI data formats

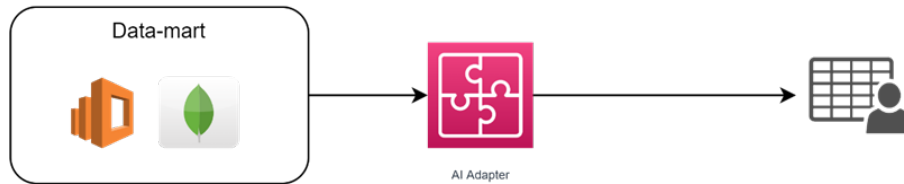


Figure 8. Data API and ML/AI input data generation.

A common Data API will be defined as a unique, scalable, and standard way to define the data required for each use case. This Data API must be as generic as possible to be reusable, not only in the three use cases covered by the Genomed4ALL project (SCD, MM and MDS), but also by any new potential use case in the future.

Data API will define the services to extract the data in the formats, types and models defined by the use cases. The expectation is to define as generic as possible the data models and data type that are going to be used by each use case.

Final output will be a csv file, with all the features that are needed to train the ML model.

### 6.2.1.4 Training ML model

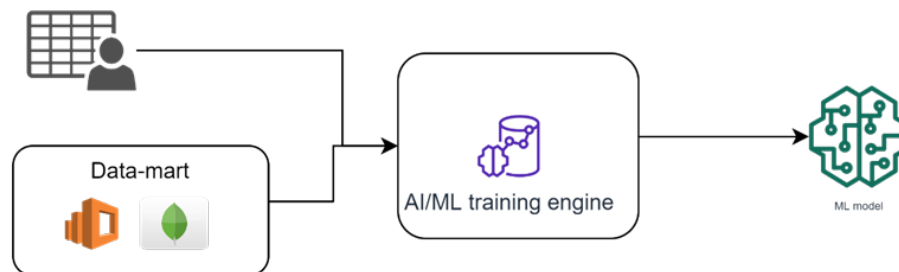


Figure 9. Training ML model, run AI algorithms.

CSV files generated in the previous step will be the entry point for the algorithms defined in WP6 for training the ML model and to run the AI algorithms. If the algorithms required some extra information, all the data will be accessible in the CDM

### 6.2.1.5 Share ML model between central server and edge nodes

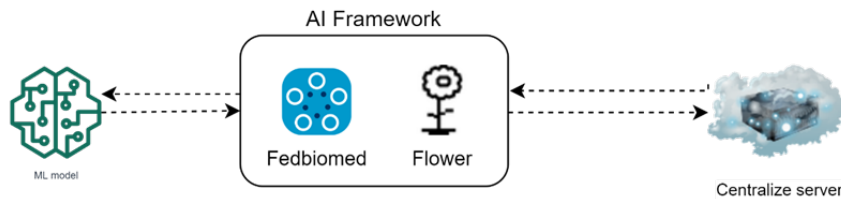


Figure 10. Share ML model between Central server and edge to update local copy and edge with Central server

The ML model will be shared with the Central server to merge it, following a hybrid federate learning approach, with the model created by all the edge nodes. The central server will make the merge of all the models received and will send back the new ML model to the edges to update it locally. This will improve the capacity of the consortium to train the models, increasing the potential capacity to generate more robust and exhaustive models for prediction.

To share this model, WP4 (D4.2 Data sharing platform architecture and components) will define the AI Framework. In each edge, Data API will use the AI Framework to coordinate with the central storage to share, in both directions, the ML model.

## 6.3 Data homogenization process definition

To homogenize data, an ETL tool will be used. This tool allows the easy connection to a specific data source, and a target, and requires only the definition and the subsequent execution of the mappings in order to transform the input data format into the output data format.

Once the CDM is defined by D3.3, it will be required a deep analysis of each use case to define how the data must be modeled into FHIR. Enabling a common understanding of the available data through the CDM makes possible the next data curation process, and data preprocessing.

To describe the data homogenization process, we have defined some generic guidelines for some of the most common data formats that will be in Genomed4ALL. At this moment, and with the current information, it is not possible to define a deep homogenization process. In next steps, it will be required to define by use case what information it will be needed, how this information will be modeled into FHIR Server.

### 6.3.1 Structured data

In case of structured data, such as relational database data or csv files, a mapping table will be required to link the source data with the target format.

This mapping table pretends to be a guide for the developers to create the mapping class that will generate the FHIR resources.

Source	Target
<b>mds.gender</b>	patient.gender
<b>mds.Hemoglobin (g/L)</b>	observation.quantity
<b>mds.genome.result</b>	Create one observation for genome with next three fields
<b>mds.genome.n mutations</b>	Observation.component
<b>mds.genome.load</b>	Observation.component

### 6.3.2 Non structured data

For Non structured data, a strong collaboration with the healthcare provider will be required to define the way to extract the significant information from the source and to define the mappings tables to create and load the information in the proper FHIR resources.

### 6.3.3 Images

#### 6.3.3.1 Dicom images

Integrate DICOM images [1] will be done in two different ways: image extraction and patient information.

Image extraction means that the DICOM image can be exported into a jpg file, even transform it into a FHIR resource, like Media or Imaging Study, in case that those images should be accessible and read.

Patient information contained in the DICOM metadata should be analyzed by the Data Analyst, that will define what information is susceptible to be exported and used in Genomed4ALL.

#### 6.3.3.2 No dicom images

No Dicom images will be loaded in FHIR only in case they are going to be accessible through the FHIR Server. In that case, images will be converted and loaded as FHIR Media resources.

Information related to those images must be contributed in other ways (structured or not) and there must be a way to match the information and the images.

### 6.3.4 VCF files

VCF (Variant Call Format) [4] files for genomic variants contain meta-information lines, a header line, and then data lines reporting the information about a position in the genome. The format also has the ability to contain genotype information on samples for each position.

```

##fileformat=VCFv4.2
##fileDate=20090805
##source=MyImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA000001 NA000002 NA000003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0/0:48:1:51,51 1/0:48:8:51,51 1/1:43:5:...
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0/0:49:3:58,50 0/1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1/2:21:6:23,27 2/1:2:0:18,2 2/2:35:4
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0/0:54:7:56,60 0/0:48:4:51,51 0/0:61:2
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0/1:35:4 0/2:17:2 1/1:40:3

```

Figure 11. An example of a VCF file

These files will be converted into FHIR formats, following the approach suggested by HL7 [7], based on DiagnosticReport and Observations FHIR resources.

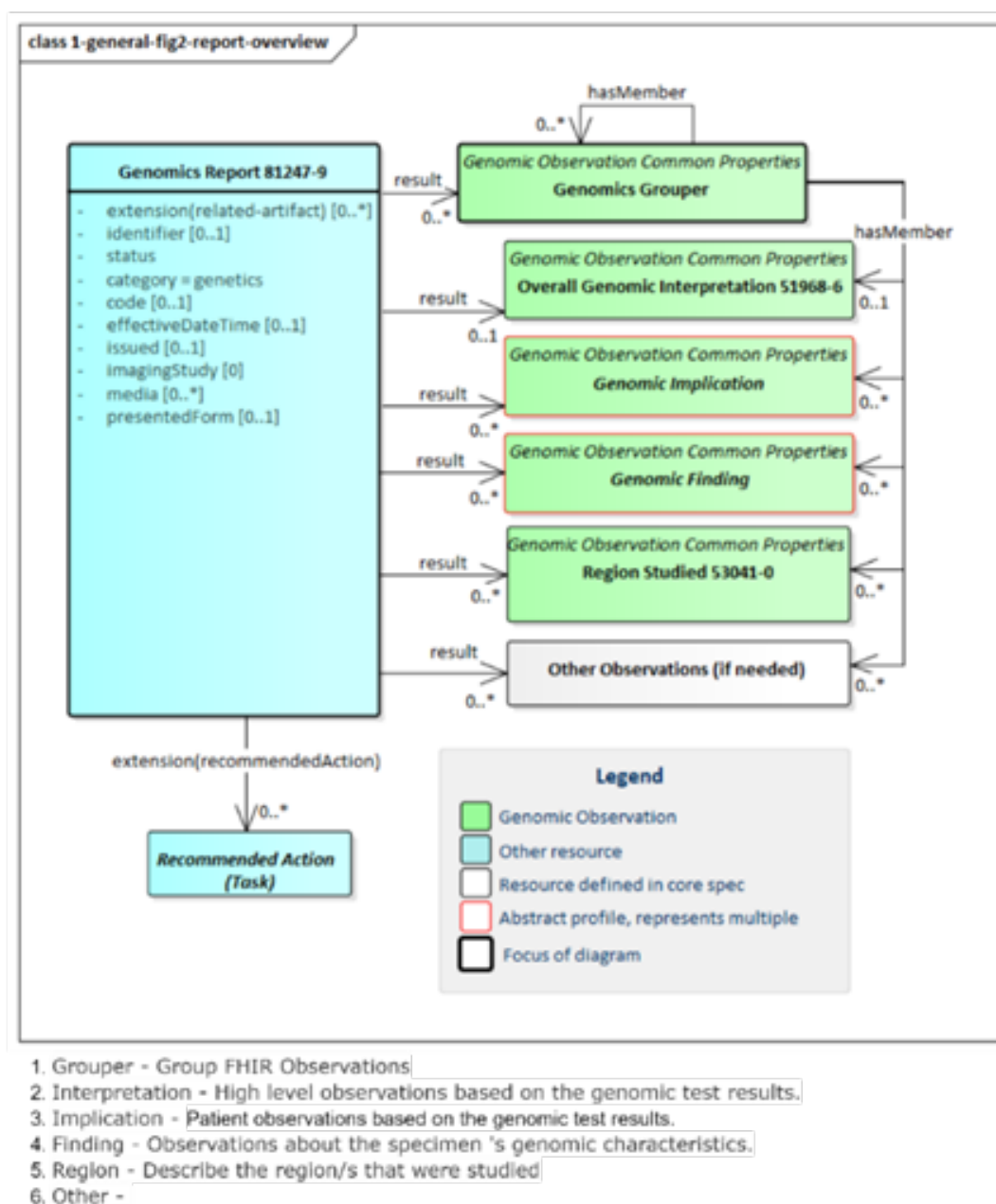


Figure 12. Genomics Diagnostic report definition. Illustrate the relationship between a diagnostic report and the different types of (conceptual) genomics observations suggested by HL7 FHIR.

For this purpose, an external tool [11], will be used. This tool is able to read the VCF file and to generate in FHIR format as many resources as needed to model the information contained in the file.

The image shows a JSON file viewer with a tree view on the left and the JSON content on the right. The tree view shows a root node with a 'DiagnosticReport' resource. The 'DiagnosticReport' resource has a 'meta' field, a 'profile' field, a 'status' field, a 'category' field, a 'coding' field, a 'subject' field, and a 'result' field. The 'result' field contains an array of 'Observation' resources. The 'Observation' resource has a 'meta' field, a 'profile' field, a 'status' field, a 'category' field, a 'coding' field, a 'subject' field, and a 'component' field. The 'component' field contains an array of 'code' objects, each with a 'coding' field and a 'valueCodeableConcept' field. The 'coding' field contains a 'system' field, a 'code' field, and a 'display' field. The 'valueCodeableConcept' field contains a 'coding' field with a 'system' field, a 'code' field, and a 'display' field.

```

1 {
2   "resourceType": "DiagnosticReport",
3   "id": "",
4   "meta": {
5     "profile": [
6       "http://hl7.org/fhir/uv/genomics-reporting/StructureDefinition/genomics-report"
7     ],
8     "status": "final",
9     "category": [
10      {
11        "coding": [
12          {
13            "system": "http://terminology.hl7.org/CodeSystem/v2-0074",
14            "code": "GDE"
15          }
16        ],
17        "display": "Master HL7 genetic variant reporting panel"
18      }
19    ],
20    "subject": {
21      "reference": "Patient/NA12878"
22    },
23    "issued": "2021-11-22T15:30:52+00:00"
24  },
25  "result": [
26    {
27      "resourceType": "Observation",
28      "id": "",
29      "meta": {
30        "profile": [
31          "http://hl7.org/fhir/uv/genomics-reporting/StructureDefinition/region-studied"
32        ],
33        "status": "final",
34        "category": [
35          {
36            "coding": [
37              {
38                "system": "http://terminology.hl7.org/CodeSystem/observation-category",
39                "code": "laboratory"
40              }
41            ],
42            "code": {
43              "coding": [
44                {
45                  "system": "http://loinc.org",
46                  "code": "53041-0",
47                  "display": "DNA region of interest panel"
48                }
49              ]
50            },
51            "subject": {
52              "reference": "Patient/RG00628"
53            },
54            "component": [
55              {
56                "code": {
57                  "coding": [
58                    {
59                      "system": "http://loinc.org",
60                      "code": "92822-6",
61                      "display": "Genomic coord system"
62                    }
63                  ],
64                  "valueCodeableConcept": {
65                    "coding": [
66                      {
67                        "system": "http://loinc.org",
68                        "code": "LA30102-0",
69                        "display": "1-based character counting"
70                      }
71                    ]
72                  }
73                }
74              ]
75            }
76          }
77        ]
78      }
79    ]
80  }
81 }

```

Figure 13. JSON file created by vcf2fhir tool, with one DiagnosticReport and the Observations with the genomics information

This information, will be loaded into an integrated record to be accessible by a FHIR Server, or consumed by the next data curation and preprocessing pipelines.

## 7 Conclusions

This deliverable briefly presents and discusses the state of the art on data homogenization, illustrating the requirements to develop the related framework within GENOMED4ALL and defining the data homogenization process to be followed through the lifetime of the project. The development of the homogenization process is expected to run iteratively, as more data sources become available and as the project evolves from a centralized homogenization architecture to a distributed one. The developed solutions along with the expected updates will be reported in the subsequent deliverables of WP5, which in line with the tasks from other workpackages will enable the homogeneous access of the integrated information.



## 8 References

- [1] Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Poggi, A., Rodriguez-Muro, M., Rosati, R. Ontologies and Databases: The DL-Lite Approach, Reasoning Web, 2009, 255-356.
- [2] Collier, N., et al. An ontology-driven system for detecting global health events, Int. Conf. on Computational Linguistics (COLING), 2010, 215-222.
- [3] DICOM. (2021). DICOM. Retrieved from <https://www.dicomstandard.org/>
- [4] Global Alliance for Genomics & Health (GA4GH). (2021, 07 27). The Variant Call Format (VCF) Version 4.2 Specification. <https://samtools.github.io/hts-specs/VCFv4.2.pdf>. Retrieved from <https://samtools.github.io/hts-specs/VCFv4.2.pdf>
- [5] He, Y., Xiang, Z., Sarntivijai, S., Toldo, L., Ceusters W. AEO: A Realism-Based Biomedical Ontology for the Representation of Adverse Events, Int. Conf. on Biomedical Ontology, Representing Adverse Events Workshop, July 26, 2011
- [6] HL7. (2021). HL7 FHIR Release 4. Retrieved from <http://hl7.org/fhir/>
- [7] HL7 FHIR. (2021). General Genomic Reporting. Retrieved from General Genomic Reporting: <http://hl7.org/fhir/uv/genomics-reporting/general.html>
- [8] Peace, J, Brennan, P.F. Ontological representation of family and family history, at AMIA Annu Symp Proc. 2007.
- [9] RADeep. (2021). Retrieved from <https://www.radeepnetwork.eu/>
- [10] Rab, Minke AE, et al. "Rapid and reproducible characterization of sickling during automated deoxygenation in sickle cell disease patients." American journal of hematology 94.5 (2019): 575-584. <https://doi.org/10.1002/ajh.25443>
- [11] vcf2fhir. (n.d.). Retrieved from vcf2fhir: a utility to convert VCF files into HL7 FHIR format for genomics-EHR integration: <https://github.com/elimuinformatics/vcf2fhir>



# GENOMED 4ALL

[genomed4all.eu](https://genomed4all.eu)



@genomed4all



/genomed4all



GENOMED4ALL receives funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement No. 101017549