



# HEREDITARY

HetERogeneous sEmantic Data integration for the guT-bRain interplaY

**Deliverable 3.6**

## **FAIRification of participating data resources**

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No GA 101137074. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.



**Funded by  
the European Union**



## EXECUTIVE SUMMARY

HEREDITARY's Work Package 3 develops a multimodal, federated analytics platform; Task 3.6 designs FAIRification workflows and 1+MG-aligned practices for consortium data, and D3.6 reports their **first implementation** between months 7 (July 2024) and 24 (December 2025). These workflows apply the FAIR Principles, which seek to make digital assets Findable, Accessible, Interoperable and Reusable, and align with the 1+MG Framework for secure, privacy-preserving, cross-border human sensitive data sharing and discovery.

A consortium-wide assessment combined surveys and bilateral meetings to identify datasets and governance constraints, leading to a **tiered FAIRification strategy**: Tier A (future FEGA-model implementation), Tier B (submission to the European Genome-phenome Archive) and Tier C (catalogue-only metadata exposure via HealthDCAT-AP and related services).

As a Tier B case study, RadboudUMC's **BaCo inflammatory bowel disease cohort** was prepared for EGA deposition by building ontology-backed data dictionaries, harmonising clinical and microbiome variables, and designing the linkage of human and non-human components across EGA, ENA and BioSamples. Legal review of the EGA Data Processing Agreement and pre-GDPR consent has now cleared the way for submission, with later re-structuring in the emerging FEGA metadata model foreseen.

For Tier C, Task 3.6 produced **HealthDCAT-AP records and supporting Zenodo packages** for the Healthy Brain Study and two ALS cohorts from UNIPD and UNITO, validated with SHACL and Metadata Quality Assessment tools and intended for publication via the European Health Information Portal and harvesting by data.europa.eu. Parallel efforts include ONTO's reusable FAIRification pipelines over **LinkedLifeData Inventory** resources, CNAG's **genomic data quality and harmonisation toolchain**, and KU Leuven's **legal analysis of AI Act dataset-completeness requirements**, which together strengthen both the technical and governance foundations for future Tier A FAIRification.



## Document information

<b>Deliverable ID</b>	D3.6
<b>Deliverable Title</b>	FAIRification of participating data resources
<b>Work Package</b>	WP3
<b>Lead Partner</b>	EMBL
<b>Due date</b>	31.12.2025
<b>Date of submission</b>	23.12.2025
<b>Type of deliverable</b>	R – Document, report
<b>Dissemination level</b>	PU – Public

## Authors

Name	Organisation
Marcos Casado Barbero <sup>1</sup>	EMBL-EBI <sup>2</sup>
Svetla Boytcheva <sup>3</sup>	ONTO <sup>4</sup>
Maria Barouh <sup>5</sup>	ONTO
Pavlin Gyurov <sup>6</sup>	ONTO

## Contributors

Name	Organisation
Elisabetta Biasin <sup>7</sup>	KUL <sup>8</sup>
Manuel Rueda <sup>9</sup>	CNAG <sup>10</sup>
Dietmar Fernandez Orth <sup>11</sup>	CNAG
Annemarie Boleij <sup>12</sup>	RadboudUMC <sup>13</sup>
Nils Kohn <sup>14</sup>	RadboudUMC

<sup>1</sup> <https://orcid.org/0000-0002-7747-6256>

<sup>2</sup> <https://ror.org/02catss52>

<sup>3</sup> <https://orcid.org/0000-0002-5542-9168>

<sup>4</sup> <https://ror.org/048zn2n32>

<sup>5</sup> <https://orcid.org/0009-0007-7059-0252>

<sup>6</sup> <https://orcid.org/0009-0006-7488-9350>

<sup>7</sup> <https://orcid.org/0000-0001-9090-3315>

<sup>8</sup> <https://ror.org/05f950310>

<sup>9</sup> <https://orcid.org/0000-0001-9280-058X>

<sup>10</sup> <https://ror.org/03mynna02>

<sup>11</sup> <https://orcid.org/0000-0002-1237-3192>

<sup>12</sup> <https://orcid.org/0000-0003-4495-5880>

<sup>13</sup> <https://ror.org/05wg1m734>

<sup>14</sup> <https://orcid.org/0000-0002-1954-2753>

### DELIVERABLE 3.6

19.12.2025, VERSION 1.0.0

Name	Organisation
Manfredo Atzori <sup>15</sup>	UNIPD <sup>16</sup>
Gianmaria Silvello <sup>17</sup>	UNIPD
Umberto Manera <sup>18</sup>	UNITO <sup>19</sup>
Carla D'Agostino <sup>20</sup>	UNITO
Alessandra Maccabeo <sup>21</sup>	UNITO
Stefano Callegaro <sup>22</sup>	UNITO
Giacomo Nebbia <sup>23</sup>	UCD <sup>24</sup>
Anna Romanovych <sup>25</sup>	UNIPD

## Revision history

Version	Date	Author	Document history/approvals
1.0.0	19.12.2025	Marcos Casado Barbero	Extension of "Milestone verification"
0.0.6	16.12.2025	Anna Romanovych	Minor format changes prior to submission
0.0.5	12.12.2025	Marcos Casado Barbero	Final review and formatting
0.0.4	11.12.2025	Gabriele Rinck <sup>26</sup>	Internal review
0.0.3	08.12.2025	Manuel Rueda Borrego, Dietmar Fernandez Orth	Project internal review
0.0.2	01.12.2025	Svetla Boytcheva, Maria Barouh, Pavlin Gyurov	Addition of ONTO's FAIRification
0.0.1	01.12.2025	Marcos Casado Barbero	First draft of D3.6

*Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.*

<sup>15</sup> <https://orcid.org/0000-0001-5397-2063>

<sup>16</sup> <https://ror.org/00240q980>

<sup>17</sup> <https://orcid.org/0000-0003-4970-4554>

<sup>18</sup> <https://orcid.org/0000-0002-9995-8133>

<sup>19</sup> <https://ror.org/048tbm396>

<sup>20</sup> <https://orcid.org/0009-0009-0643-1665>

<sup>21</sup> <https://orcid.org/0009-0005-7779-5897>

<sup>22</sup> <https://orcid.org/0009-0005-2427-2267>

<sup>23</sup> <https://orcid.org/0000-0002-4766-6278>

<sup>24</sup> <https://ror.org/02hh7en24>

<sup>25</sup> <https://orcid.org/0009-0005-2353-8166>

<sup>26</sup> <https://orcid.org/0000-0001-6428-3431>

## Contents

<b>List of Abbreviations.....</b>	<b>9</b>
<b>Glossary.....</b>	<b>12</b>
<b>1. Introduction.....</b>	<b>14</b>
<b>2. Task 3.6 objectives.....</b>	<b>16</b>
<b>3. Scope.....</b>	<b>16</b>
<b>4. 1+MG Standards.....</b>	<b>17</b>
4.1. Phenopackets.....	17
4.2. Beacon Network.....	18
4.3. DCAT.....	19
4.4. DCAT-AP.....	19
4.5. HealthDCAT-AP.....	19
4.6. Data Use Ontology.....	20
4.7. Phenotypic and Clinical Metadata Framework.....	20
4.8. Recommended ontologies.....	21
4.9. European Genome-phenome Archive.....	21
<b>5. Methodology.....</b>	<b>22</b>
<b>6. Implementation and Results.....</b>	<b>23</b>
6.1. Assessment of HEREDITARY datasets.....	25
6.2. Enhancing discoverability of consortium data resources.....	26
6.2.1. Selecting an EU-aligned cataloguing route.....	26
6.2.2. HealthDCAT-AP records for HEREDITARY datasets.....	27
6.2.3. Current status and next steps.....	30
6.3. FAIRification of RadboudUMC's BaCo Dataset.....	31
6.3.1. Dataset description.....	31
6.3.2. Standards applied.....	33
6.3.3. Preparation for submission to EGA.....	34
6.3.4. Outcome, limitations and next steps towards FEAGA.....	34
6.4. Genomic data quality protocol.....	35
6.5. ONTO's FAIRification.....	36
6.5.1. LinkedLifeData Inventory.....	36
6.5.2. FAIRification workflow (Apache Airflow).....	36
6.5.3. FAIRification of Expression Atlas.....	38
6.5.4. Datasets prioritised for HEREDITARY.....	40
6.6. Analysis of AI Act Requirement of Completeness for Training Datasets.....	41
<b>7. Challenges.....</b>	<b>42</b>
<b>8. Discussion.....</b>	<b>44</b>
<b>9. Next steps.....</b>	<b>46</b>
<b>10. Milestone verification.....</b>	<b>47</b>
<b>References.....</b>	<b>49</b>
<b>Annexes.....</b>	<b>49</b>

## List of Figures

Figure 1. Summary diagram of Tiered view of FAIRification opportunities.....	24
Figure 2. Radar Plots of Metadata Quality Assessments of the project's Tier C FAIRified datasets.....	29
Figure 3. EHDS FAIR Data Point FAIRness assessed by FAIR-Checker.....	30
Figure 4. Overview of content of the BaCo dataset submitted to EGA, following the archival model.....	32
Figure 5. Overview of BaCo Dataset generation and content and its relation to the EGA, ENA and BioSamples.....	33
Figure 6. Steps in Apache Airflow Pipeline.....	37
Figure 7. Apache Airflow DAGs.....	37
Figure 8. FAIRification principles. Source: <a href="https://www.go-fair.org/fair-principles/fairification-process">https://www.go-fair.org/fair-principles/fairification-process</a> .....	38
Figure 9. Excerpt from Expression Atlas semantic model.....	39
Figure 10. Class relationships.....	41

## List of Tables

Table 1. Source Data in CSV format (as a table) and its transformation to RDF-star....	39
--	----



## List of Abbreviations

Abbreviation	Description
<b>1+MG</b>	1+ Million Genomes initiative
<b>ALS</b>	Amyotrophic Lateral Sclerosis
<b>API</b>	Application Programming Interface
<b>B1MG</b>	Beyond 1 Million Genomes
<b>BaCo</b>	Bacteria in colitis associated cancer development
<b>CNAG</b>	National Centre for Genomic Analysis
<b>CORIS</b>	Colorado Ophthalmology Research Information System
<b>CRG</b>	Centre for Genomic Regulation
<b>CSV</b>	Comma-Separated Values
<b>DAC</b>	Data Access Committee
<b>DAG</b>	Directed Acyclic Graph
<b>dbGaP</b>	Database of Genotypes and Phenotypes
<b>DCAT</b>	Data Catalog Vocabulary
<b>DCAT-AP</b>	EU profile of DCAT for data catalogues
<b>DMP</b>	Data Management Plan
<b>DOI</b>	Digital Object Identifier
<b>DPA</b>	Data Processing Agreement
<b>DTA</b>	Data Transfer Agreement
<b>DUO</b>	Data Use Ontology (usage codes)
<b>EFO</b>	Experimental Factor Ontology
<b>EGA</b>	European Genome-phenome Archive
<b>EHDS</b>	European Health Data Space
<b>ELIXIR</b>	European Life-Science Infrastructure
<b>EMBL</b>	European Molecular Biology Laboratory
<b>EMBL-EBI</b>	European Bioinformatics Institute

Abbreviation	Description
<b>ENA</b>	European Nucleotide Archive
<b>EU</b>	European Union
<b>FAIR</b>	Findable, Accessible, Interoperable, Reusable
<b>FDP</b>	FAIR Data Point
<b>FEGA</b>	Federated European Genome-phenome Archive
<b>FTP</b>	File Transfer Protocol
<b>GA4GH</b>	Global Alliance for Genomics and Health
<b>GATK</b>	Genome Analysis Toolkit
<b>GDI</b>	Genome Data Infrastructure
<b>GDPR</b>	EU General Data Protection Regulation
<b>GO</b>	Gene Ontology
<b>gVCF</b>	Genomic VCF
<b>HDAB</b>	Health Data Access Body
<b>HDN</b>	Hereditary Data Network
<b>HealthDCAT-AP</b>	Health extension of DCAT-AP
<b>HERO</b>	HEREDITARY Ontology
<b>HIP</b>	Health Information Portal
<b>HPO</b>	Human Phenotype Ontology
<b>IBD</b>	Inflammatory Bowel Disease
<b>ICD-10</b>	International Statistical Classification of Diseases and Related Health Problems, 10th Revision
<b>JSON</b>	JavaScript Object Notation (data format)
<b>JSON-LD</b>	JSON for Linked Data
<b>KUL</b>	Katholieke Universiteit Leuven
<b>LOINC</b>	Logical Observation Identifiers Names and Codes
<b>METC</b>	Medical Ethics Review Committee ( <i>Medisch-ethische toetsingscommissie</i> )

Abbreviation	Description
<b>MQA</b>	Metadata Quality Assessment
<b>NCBI</b>	National Center for Biotechnology Information
<b>NCIT</b>	National Cancer Institute Thesaurus
<b>OLS</b>	Ontology Lookup Service
<b>OMOP CDM</b>	Observational Medical Outcomes Partnership Common Data Model
<b>ORDO</b>	Orphanet Rare Disease ontology
<b>PCMF</b>	Phenotypic and Clinical Metadata Framework
<b>PXF</b>	Phenotype Exchange Format
<b>QC</b>	Quality Control
<b>RDF</b>	Resource Description Framework
<b>RDFS</b>	RDF Schema
<b>REST</b>	Representational State Transfer (web API style)
<b>SHACL</b>	Shapes Constraint Language
<b>SNOMED</b>	Systematized Nomenclature of Medicine
<b>SNOMED CT</b>	SNOMED Clinical Terms
<b>SP</b>	Submitter Portal
<b>T3.6</b>	Task 3.6 (FAIRification)
<b>UCD</b>	University of Colorado Denver
<b>UNIPD</b>	University of Padova
<b>UNITO</b>	University of Turin
<b>URI</b>	Uniform Resource Identifier
<b>VCF</b>	Variant Call Format
<b>W3C</b>	World Wide Web Consortium
<b>WP</b>	Work Package (project workstream)
<b>YAML</b>	YAML Ain't Markup Language

## Glossary

Term	Description
<b>Accession</b>	Stable identifier assigned by an archive
<b>Beaconisation</b>	Mapping local metadata to Beacon model and deploying it as an endpoint
<b>Catalogue (DCAT)</b>	A collection of dataset descriptions
<b>Controlled access</b>	Data access via approvals/policies (e.g., DAC)
<b>Controlled Vocabulary</b>	A curated and restricted set of terms that must be used consistently to describe data in a particular field or record set
<b>Dataset (DCAT)</b>	A describable, citable data resource
<b>Dataset Sample</b>	A sample distribution of the dataset as per HealthDCAT-AP standard's definition (e.g., data dictionaries, mock-up data, anonymised data, synthetic data).
<b>FAIRification</b>	Practical steps to make data FAIR
<b>FAIRify</b>	To modify and document (meta)data so that they better comply with the FAIR principles
<b>FASTQ</b>	Text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores
<b>Federated discovery</b>	Query many nodes without moving data
<b>Findability</b>	Ease of discovering datasets/metadata
<b>Harmonisation</b>	Aligning variables/metadata across sources
<b>Harvesting</b>	Pulling metadata from provider catalogues
<b>Non-identifiable metadata</b>	Descriptive information about data or samples that does not, on its own or in combination with other reasonably available information, allow an individual person to be identified
<b>Ontology</b>	Controlled concepts/relations for semantics
<b>Persistent identifier</b>	Long-lived, resolvable ID (e.g., DOI/accession)
<b>Phenopackets</b>	GA4GH phenotype data model
<b>Proprietary dataset</b>	Dataset that is generated by or controlled within the

Term	Description
	HEREDITARY project and thus restricts its access
<b>RDF-star</b>	Extension of RDF for statement-level annotations
<b>Solve-RD</b>	"Solve-RD - solving the unsolved rare diseases" is a research project funded by the European Commission (Grant No. 779257) from 01.01.2018 to 31.03.2024
<b>Triplestore</b>	Database for RDF triples/graphs

## 1. Introduction

**HEREDITARY** develops a multimodal, federated analytics platform that semantically integrates text, genomics, bioimages, and environmental data to enable privacy-preserving discovery and analysis across partner institutions.<sup>27</sup> Within its Work Package 3 (WP3),<sup>28</sup> **Task 3.6** (T3.6) "FAIRification workflows and 1+MG guidelines compliance" aims to make the participating data resources Findable, Accessible, Interoperable, and Reusable (FAIR) according to the **FAIR Principles** [1]. This task has two dedicated Deliverables, D3.6<sup>29</sup> and its revision, D3.7, to report on the design and implementation of these FAIRification workflows. This document corresponds to the former, in which we outline the first batch of concrete approaches to FAIRify HEREDITARY's data resources and their results, with a deliberate focus on discoverability.

In this document, we survey **standards** adopted by the **1+ Million Genomes** (1+MG) initiative<sup>30</sup> and the **European Genomics Data Infrastructure** (GDI) project,<sup>31</sup> especially those developed by the Global Alliance for Genomics and Health (GA4GH),<sup>32</sup> and report on their implementation by the HEREDITARY project. Alignment with the 1+MG Framework<sup>33</sup> ensures HEREDITARY's outputs remain relevant in an ever-growing federation of European discovery and access networks. GA4GH standards provide some concrete implementation pathways for that alignment. For example, these include Beacon<sup>34</sup> for discovery, Data Use Ontology (DUO) codes<sup>35</sup> for machine-readable access conditions, and Phenopackets<sup>36</sup> for phenotypic/clinical metadata representation.

Our guiding frameworks are the **FAIR Principles**, through which we emphasise machine-actionable metadata discovery and reuse across stakeholders. We also adopt these principles as evaluation lenses for measuring improvements documented here.

By **FAIRification**, we mean a set of pragmatic, incremental steps to make data FAIR, including broad approaches like: (1) enriching and harmonising metadata, (2) exposing machine-readable discovery signals while respecting data governance; and (3) enabling access to data through standardised methods. More specifically, we format data sources according to European Union (EU) **catalogue-grade schemas** (e.g., DCAT-AP 3.0)<sup>37</sup> and widely used **controlled vocabularies/ontologies** (e.g., Human Phenotype Ontology – HPO).<sup>38</sup> Additionally, where possible and in accordance with the project's goal of accelerating data sharing and access, we showcase the benefit of

<sup>27</sup> <https://hereditary-project.eu/overview>

<sup>28</sup> <https://hereditary-project.eu/work-packages>

<sup>29</sup> <https://doi.org/10.5281/zenodo.17901862>

<sup>30</sup> <https://digital-strategy.ec.europa.eu/en/policies/1-million-genomes>

<sup>31</sup> <https://gdi.onemilliongenomes.eu/about>

<sup>32</sup> <https://www.ga4gh.org/about-us>

<sup>33</sup> <https://framework.onemilliongenomes.eu/about-the-framework>

<sup>34</sup> <https://genomebeacons.org>

<sup>35</sup> <https://www.ga4gh.org/product/data-use-ontology-duo>

<sup>36</sup> <https://www.ga4gh.org/product/phenopackets>

<sup>37</sup> <https://semiceu.github.io/DCAT-AP/releases/3.0.0>

<sup>38</sup> <https://hpo.jax.org>

submitting to long-term, secure archivals like the **European Genome-phenome Archive** (EGA).<sup>39</sup> Importantly, **discoverability is not synonymous with open access**; submitting to EGA improves overall FAIRness without altering access controls and data ownership.<sup>40</sup>

Furthermore, D3.6 sits alongside WP1's **Data Management Plan** (DMP) and its associated deliverable, D1.1, translating policy into practice by documenting how standards, sharing, and curation can effectively increase the value of project data resources. These lessons will be fed back into the DMP during its revision in D1.2. Additionally, T3.6 works closely with other WP3 tasks that are directly involved in the semantic integration of project data resources. Namely, T3.1 "Multimodal semantic ontology" with the creation of the HEREDITARY Ontology,<sup>41</sup> T3.4 "Genomics data science" and its beaconisation efforts, and T3.5 "Privacy-preserving processing" with its overseeing.

WP3 spans multiple modalities (e.g., genomics, clinical tables, imaging) which require a shared semantic layer to enable federated analytics. In this first attempt at FAIRifying project datasets, we centre our efforts on data from the **project's proprietary sources**: University of Padova (UNIPD) and University of Turin (UNITO) in Italy; Radboud University Medical Center (RadboudUMC) in the Netherlands; and the University of Colorado Denver (UCD) in the United States of America. Data from these institutions are used as the spearhead for implementing FAIRification workflows in the project, spanning the following categories of data that we attempt to FAIRify in T3.6:

- **Genomics data:** Digital outputs and derivatives of measuring genomes, including raw sequence reads (e.g., FASTQ files), processed alignments (e.g., BAM files), and variant call sets (e.g., VCF files). These are used to study all genes and their interactions with each other and the environment.
- **Structured clinical/phenotypic data:** Values captured in predefined fields and codes (e.g., demographics, diagnoses, medications, labs) that support direct computable querying and interoperability.
- **Unstructured data:** They lack a fixed schema (e.g., free-text clinical notes, narrative reports, images) and typically require Natural Language Processing (NLP) or other signal-processing methods to extract computable information.

Ultimately, we emphasise **discoverability** because in a federated project like HEREDITARY, outputs can only scale if researchers can reliably find eligible datasets across the network of nodes **before requesting access to them**.

Additionally, HEREDITARY also draws on **external datasets** (i.e., non-proprietary) to which specific collaborators have been granted access. These are, for example, datasets archived and distributed by the EGA and the database of Genotypes and

---

<sup>39</sup> <https://ega-archive.org/about/ega>

<sup>40</sup> [https://rdmkit.elixir-europe.org/human\\_data#sharing-and-reusing-of-human-data](https://rdmkit.elixir-europe.org/human_data#sharing-and-reusing-of-human-data)

<sup>41</sup> <https://hereditary.dei.unipd.it/ontology>

Phenotypes (dbGaP),<sup>42</sup> which are routinely reused by CNAG. While FAIRification principles can equally be applied to these resources, this report deliberately focuses on proprietary HEREDITARY datasets, where documentation and discoverability are currently more limited, or non-existing, than for external datasets.

## 2. Task 3.6 objectives

- Boost **dataset discoverability** by exposing non-identifiable (e.g., high-level and pseudonymised data), standards-compliant metadata.
- **Harmonise metadata** with European profiles to enable federated cross-site discovery.
- Apply **FAIRification workflows** across genomics, structured, and unstructured data in participating resources.
- Align implementations with 1+MG/GDI and GA4GH standards for **interoperability**.
- Produce **documented improvements** in FAIRness and report them publicly in both D3.6 and its revision D3.7.
- Provide **reusable guidance and workflows** that pave the way for more datasets to be FAIRified.
- Improve findability without altering **access controls or ownership**.
- Inform the **DMP** (D1.1) and its revision (D1.2).

## 3. Scope

This deliverable **covers** the design and first implementations of T3.6 FAIRification activities performed from M7–M24, focusing on measurable discoverability gains in participating consortium data sources. It spans genomics, structured clinical/phenotypic, and some unstructured modalities (e.g., imaging or clinical free text); targets metadata enhancement, harmonisation to European models, and the application of 1+MG, GDI, and GA4GH discovery assets (e.g., DCAT-AP, DUO...).

The scope includes project-generated datasets and explicitly excludes changes to data-access governance or re-consent activities; the emphasis is on metadata-level improvements and their documented evidence for D3.6. Finally, it includes proposed next steps for the last year of T3.6 (M24–M36).

It **does not cover**:

- The creation of brand-new tools or standards where existing ones suffice.

---

<sup>42</sup> <https://dbgap.ncbi.nlm.nih.gov/home/>



- An in-depth privacy-preserving report (see D3.8 "Privacy-preserving analytics: first release").
- A legal and ethical study (see D7.2 "In-depth legal and ethical study").
- Active Beaconisation (i.e., setting up a Beacon network) of participating resources: T3.4 has taken the lead and T3.6 collaborates on that matter. See D3.11 "Pilot of the genomics data science ontology interconversion" for further details.

## 4. 1+MG Standards

Within the **1+MG initiative**, data and metadata are organised around a minimal cross-border metadata model and a set of both mature and emerging standards that enable discovery and access to genomic and related health data across Europe. To facilitate searching, linking and analysis of data, the 1+MG Framework has collected a **set of recommendations and guidelines**.<sup>43</sup> In this section, we will summarise the main 1+MG standards relevant to the HEREDITARY project and its use-cases.

These recommendations, originally developed in Beyond 1 Million Genomes (B1MG) and implemented through the GDI project, point participating nodes to **GA4GH standards** such as the DUO codes (for machine-readable data-use conditions) and Phenopackets (for structured phenotypic/clinical descriptions), alongside **catalogue standards like DCAT** and its European profiles **DCAT-AP** and **HealthDCAT-AP**, which support interoperability with the wider European Health Data Space (EHDS).

Together with the Phenotypic and Clinical Metadata Framework (3v0 v1, 2023),<sup>44</sup> and recommended clinical and biomedical widely used ontologies (e.g. HPO,<sup>45</sup> EFO,<sup>46</sup> SNOMED CT),<sup>47</sup> this stack defines the 1+MG reference architecture that HEREDITARY adopts in T3.6 when FAIRifying participating datasets and aligning them with European discovery and access networks. This effort is initially reported in D3.6, but will be refined throughout 2026 and finally reported back in D3.7 (see [Next steps](#)).

Furthermore, these 1+MG recommendations are complementary to the HEREDITARY-developed standards, such as the HEREDITARY Ontology (HERO).<sup>48</sup>

### 4.1. Phenopackets

The GA4GH Phenopacket schema<sup>49</sup> is a standard format for representing detailed **phenotypic and clinical information about diseases** (e.g., diagnoses), **patients** (e.g., clinical features), and **genetics** (e.g., genomic variants). It was developed to improve researchers' ability to understand, diagnose, and treat both rare and common

<sup>43</sup> <https://framework.onemilliongenomes.eu/data-models-ontologies>

<sup>44</sup> <https://doi.org/10.5281/zenodo.10058688>

<sup>45</sup> <https://hpo.jax.org/about>

<sup>46</sup> <https://www.ebi.ac.uk/efo/>

<sup>47</sup> <https://bioportal.bioontology.org/ontologies/SNOMEDCT>

<sup>48</sup> <https://hereditary.dei.unipd.it/ontology/>

<sup>49</sup> <https://www.ga4gh.org/product/phenopackets>

diseases. Phenopackets are stored as PFX (Phenotype Exchange Format) files, which can be saved in JSON or YAML formats. Each "packet" links a set of physical features to a disease and patient, providing detailed information for each (e.g., age, gender, disease onset). Furthermore, Phenopackets recommends a set of ontologies (e.g., HPO) to ensure interoperability across resources.

Within the 1+MG Framework, Phenopackets (v2, 2021)<sup>50</sup> are **recommended** as the standard to exchange phenotypic information across services. It is the required format for the rare disease use case (e.g., Solve-RD project),<sup>51</sup> with other use-cases under review.<sup>52</sup>

## 4.2. Beacon Network

The GA4GH Beacon API is an open standard for **genomic and clinical data discovery** that allows users to enquire whether datasets held by participating repositories contain records matching a set of criteria (e.g., variant or phenotype filters). The key difference with other approaches is that users can perform these queries without going through the time-consuming process of data access requests, all while preserving the privacy of data subjects by only exposing summary-level information.<sup>53</sup>

In the latest Beacon release (v2) the model and API have been extended to support **richer queries** and tighter integration with other standards such as Phenopackets and DUO. This highlights Beacon as a general "lingua franca" for federated data discovery across heterogeneous genomic resources.<sup>54</sup>

Within the 1+MG Framework, Beacon v2 is designated as a **required data discoverability standard**.<sup>55</sup> Each national node is expected to expose Beacon endpoints over summary-level descriptions of local datasets, enabling users to find relevant datasets and then follow links to the node's data access management process for full access. An example in development is GDI's allele frequency portal that is built on a network of allele frequency beacons.<sup>56</sup>

Given the highly sensitive content of the HEREDITARY datasets and its restrictive accessibility, data discovery would be greatly improved if Beacons at each participating data resource were deployed. These efforts are being undertaken by Task 3.4 of HEREDITARY (see D3.11 for further details). Furthermore, it is imperative that developments of the HEREDITARY Data Network (HDN)<sup>57,58</sup> are well coordinated with Beacon deployments and integrated in the FAIRification plans when available. See D3.2 for further details on the HDN.

---

<sup>50</sup> <https://phenopacket-schema.readthedocs.io/en/latest/schema.html#version-2-0>

<sup>51</sup> <https://solve-rd.eu/>

<sup>52</sup> <https://framework.onemilliongenomes.eu/data-models-ontologies>

<sup>53</sup> <https://www.ga4gh.org/product/beacon-api>

<sup>54</sup> <https://docs.genomebeacons.org/>

<sup>55</sup> <https://framework.onemilliongenomes.eu/data-discovery>

<sup>56</sup> <https://catalogue-test.azurewebsites.net/allele-frequency>

<sup>57</sup> <https://github.com/mircocazzaro/tdn-endpoint>

<sup>58</sup> <https://github.com/mircocazzaro/central-tdn>

### 4.3. DCAT

The **Data Catalog Vocabulary** (DCAT, v3, 2024) <sup>59</sup> is a World Wide Web Consortium (W3C) <sup>60</sup> Resource Description Framework (RDF) <sup>61</sup> vocabulary for describing **datasets and data catalogues** on the Web, enabling interoperable publication and harvesting of dataset metadata across different portals. It facilitates decentralised catalogue publishing and eases federated dataset searches across multiple catalogues, which perfectly suits the HEREDITARY network.

In 1+MG, DCAT is a **required standard** and provides the foundational model for catalogue-level metadata exposed by national nodes and the 1+MG User Portal. <sup>62,63</sup> It supports machine-actionable discovery of genomic and health datasets and facilitates interoperability with other DCAT-compatible catalogues (e.g., GOV.UK). <sup>64,65</sup>

### 4.4. DCAT-AP

The DCAT Application Profile (DCAT-AP, v3, 2024) <sup>66</sup> is the **European application profile of DCAT** that specifies shared **classes, properties and controlled vocabularies** so that public sector data portals can exchange and aggregate dataset descriptions across borders and domains. <sup>67</sup> In summary, DCAT-AP is an extension of DCAT, which further delineates requirements for dataset records.

The 1+MG Framework designates DCAT-AP as a **required profile** for data discovery. <sup>68</sup> Genomic and associated health datasets in national 1+MG and GDI nodes are expected to be exposed via DCAT-AP-compliant catalogues so they can be harvested into the 1+MG User Portal <sup>69</sup> and integrated with EU-wide dataset catalogues (e.g., data.europa.eu) <sup>70</sup> within the EHDS.

### 4.5. HealthDCAT-AP

HealthDCAT-AP (v5, 2025) <sup>71</sup> is a health-specific extension of DCAT-AP, developed in the HealthData@EU pilot. Its goal is to **standardise descriptive metadata for health datasets and catalogues**, including elements for data sensitivity, legal bases, data controllership, data use conditions and dataset quality or utility labels. Similarly to how DCAT-AP is an extension of DCAT, HealthDCAT-AP retains compatibility with DCAT-AP and, therefore, with plain DCAT.

---

<sup>59</sup> <https://www.w3.org/TR/vocab-dcat-3/>

<sup>60</sup> <https://www.w3.org/>

<sup>61</sup> <https://www.w3.org/RDF/>

<sup>62</sup> <https://framework.onemilliongenomes.eu/data-models-ontologies>

<sup>63</sup> <https://catalog-test.healthdata.nl/>

<sup>64</sup> <https://framework.onemilliongenomes.eu/data-discovery>

<sup>65</sup>

<https://www.gov.uk/government/publications/recommended-open-standards-for-government/using-metadata-to-describe-data-assets-in-a-data-catalogue>

<sup>66</sup> <https://semiceu.github.io/DCAT-AP/releases/3.0.0/>

<sup>67</sup> <https://op.europa.eu/en/web/eu-vocabularies/dcat-ap>

<sup>68</sup> <https://framework.onemilliongenomes.eu/data-discovery>

<sup>69</sup> <https://catalog-test.healthdata.nl/>

<sup>70</sup> [data.europa.eu](https://data.europa.eu)

<sup>71</sup> <https://healthdataeu.pages.code.europa.eu/healthdcat-ap/releases/release-5/>

While it is not yet listed on the 1+MG standards page, ongoing discussions within GDI and recent Digital Europe calls for tools and services explicitly require that metadata for 1+MG/GDI datasets be compatible with standards of the EHDS Dataset Catalogue (i.e., HealthDCAT-AP).<sup>72</sup> Therefore, nodes are expected to progressively adopt HealthDCAT-AP.

#### 4.6. Data Use Ontology

The GA4GH Data Use Ontology (DUO) defines a **controlled vocabulary of data use conditions** that can be attached to datasets as machine-readable tags describing permitted secondary uses. For example, restrictions by disease area (DUO:0000007),<sup>73</sup> non-commercial use (DUO:0000046)<sup>74</sup> or geographical limitations (DUO:0000022).<sup>75</sup> These attributes enable prospective researchers to filter datasets based on their data use conditions. Additionally, they simplify the access request processing performed by Data Access Committees (DAC) and Data Controllers.

In the 1+MG Framework, DUO is a **required standard**: data submitted to national nodes should encode access conditions of controlled-access datasets using DUO so that discovery services, GDI infrastructure and repositories such as the Federated EGA (FEGA)<sup>76</sup> can facilitate automatic filtering of datasets by allowed uses and support data access committees in evaluating access requests.<sup>77</sup>

#### 4.7. Phenotypic and Clinical Metadata Framework

The Phenotypic and Clinical Metadata Framework (PCMF, 3v0 v1, 2023)<sup>78</sup> was a 1+MG/B1MG deliverable (D3.6) that set out broad principles, models and recommendations for **representing phenotypic and clinical metadata and linking them to genomic data** in a cross-country vocabulary.

It is listed in the 1+MG Framework as a **required instrument**,<sup>79</sup> providing guidance on which standards, terminologies and tools national 1+MG nodes should use when designing local data models and FAIRification pipelines. In HEREDITARY we therefore treat the PCMF primarily as a normative reference, rather than as an additional data model to implement per se. Task 3.6 efforts on FAIRification adopt the standards and practices promoted in the PCMF, such as HealthDCAT-AP, DUO Codes, Beacon-based and Phenopackets-aligned HERO ontologies, and EGA/FEGA as the preferred repository for controlled-access data.

<sup>72</sup>

<https://www.euro-access.eu/en/calls/2434/Health-Data-ingestion-capacities-and-data-services-for-the-European-Genomic-Data-Infrastructure-in-the-European-Health-Data-Space-data-tools>

<sup>73</sup> [http://purl.obolibrary.org/obo/DUO\\_0000007](http://purl.obolibrary.org/obo/DUO_0000007)

<sup>74</sup> [http://purl.obolibrary.org/obo/DUO\\_0000046](http://purl.obolibrary.org/obo/DUO_0000046)

<sup>75</sup> [http://purl.obolibrary.org/obo/DUO\\_0000022](http://purl.obolibrary.org/obo/DUO_0000022)

<sup>76</sup> <https://ega-archive.org/about/projects-and-funders/federated-ega/>

<sup>77</sup> <https://framework.onemilliongenomes.eu/data-models-ontologies>

<sup>78</sup> <https://doi.org/10.5281/zenodo.10058688>

<sup>79</sup> <https://framework.onemilliongenomes.eu/data-models-ontologies>

#### 4.8. Recommended ontologies

1+MG does not define new terminologies but instead curates a set of recommended ontologies and controlled vocabularies for different domains.<sup>80</sup> For example, HPO and Orphanet Rare Disease ontology (ORDO) for rare disease phenotypes, the 10th Revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-10)<sup>81</sup> and Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT)<sup>82</sup> for diagnoses, and Logical Observation Identifiers Names and Codes (LOINC)<sup>83</sup> for laboratory measurements.

These **ontology recommendations** are intended to be used as value sets (i.e., controlled vocabularies) for variables in the datasets encompassed by the 1+MG framework, Phenopackets and other representations.

#### 4.9. European Genome-phenome Archive

The EGA is an ELIXIR core data resource, jointly operated by the European Bioinformatics Institute (EMBL-EBI)<sup>84</sup> and the Centre for Genomic Regulation (CRG),<sup>85</sup> that provides **secure long-term archival** and **controlled-access distribution of identifiable human genomic and associated phenotypic data**.<sup>86</sup> This key service eases data reusability and sustainability by filling in gaps in all stages of a dataset life cycle and adhering to FAIR principles.

Additionally, the **Federated EGA** (FEGA) responds to the challenge of datasets that have restrictions on storage outside of national borders (i.e., cannot be submitted to EMBL-EBI and CRG).<sup>87</sup> This model of national repositories deploying EGA-like services allows for sensitive data to remain under national jurisdiction while being discoverable and accessible via a shared framework. The FEGA is a crucial part of GDI and 1+MG, as it serves, in multiple scenarios, as the data-hosting backbone for the initiative.<sup>88,89</sup>

Sensitive data storage in the HEREDITARY project relies solely on the existing infrastructure of participating institutions. Furthermore, data access and distribution entail a case-by-case scenario where Data Transfer Agreements (DTAs) and bespoke distribution methods are to be prepared for each access request and institution. For this reason, to secure long-term data sustainability and ease reuse, it is highly recommended to **submit HEREDITARY datasets to the EGA/FEGA** where applicable. In the absence of an alternative, we expect this approach to be included in the future D8.6 "Final exploitation plan and final IPR strategy" led by Task 8.3.

---

<sup>80</sup> <https://doi.org/10.5281/zenodo.10058526>

<sup>81</sup> <https://icd.who.int/browse10/2019>

<sup>82</sup> <https://www.snomed.org/what-is-snomed-ct>

<sup>83</sup> <https://loinc.org>

<sup>84</sup> <https://ror.org/02catss52>

<sup>85</sup> <https://ror.org/03wyz892>

<sup>86</sup> <https://ega-archive.org/about/ega/>

<sup>87</sup> <https://ega-archive.org/about/projects-and-funders/federated-ega/>

<sup>88</sup> <https://ega-archive.org/about/projects-and-funders/federated-ega/>

<sup>89</sup> <https://ega-archive.org/about/projects-and-funders/projects/>

## 5. Methodology

- **Project coordination and document development**
  - Internal planning and drafting carried out using shared document platforms (e.g., Google Drive) <sup>90</sup> and the HEREDITARY content management system and wiki. <sup>91</sup>
  - Communication and requirements gathering conducted through email exchanges and virtual meetings (e.g. Zoom) <sup>92</sup> with project collaborators.
- **Dataset assessment and engagement**
  - Systematic identification of candidate datasets and data providers using the Description of Action (DoA), <sup>93</sup> an internal "Introduction to T3.6 benefits" document ([Annex 1](#)) and a short contributor survey. <sup>94</sup>
  - Structured one-to-one interviews with proprietary data sources (RadboudUMC, UNIPD, UNITO, UCD) to document dataset characteristics, governance constraints and feasible FAIRification tiers.
- **Standards review and FAIRification design**
  - Review of 1+MG, GDI and GA4GH documentation to define a target stack of standards and to embed them in FAIRification workflows where applicable.
  - Participation in the ELIXIR Interoperability Platform WP3 ("Standardising processes for FAIR data management") <sup>95,96</sup> 2025 workshop session on FAIR data management in Lausanne, Switzerland. <sup>97,98</sup>
- **Ontology mapping and metadata modelling**
  - Construction of dataset-level data dictionaries (e.g. BaCo) in spreadsheets, followed by manual and semi-automated mapping of variables and values to recommended ontologies using the Ontology Lookup Service (OLS, v4). <sup>99</sup>
  - Design of EGA-compatible and FEAGA-oriented metadata structures for BaCo in line with current EGA submission procedure.

---

<sup>90</sup> <https://drive.google.com/>

<sup>91</sup> <https://hereditary.dei.unipd.it/>

<sup>92</sup> <https://www.zoom.com/>

<sup>93</sup> <https://hereditary.dei.unipd.it/groupoffice/index.php?r=files/file/download&id=13>

<sup>94</sup> <https://forms.gle/7EKVHg9HQBPR6wrz7>

<sup>95</sup> <https://elixir-europe.org/platforms/interoperability>

<sup>96</sup> <https://elixir-europe.org/internal-projects/commissioned-services/2024-interoperability#wp3>

<sup>97</sup> <https://elixir-europe.org/events/fhd-hdtr-day>

<sup>98</sup> [https://docs.google.com/presentation/d/19n0E\\_XfHaw81d2SxbTM6RiYfzZ9nLJpT](https://docs.google.com/presentation/d/19n0E_XfHaw81d2SxbTM6RiYfzZ9nLJpT)

<sup>99</sup> <https://www.ebi.ac.uk/ols4>



- **Catalogue-level FAIRification (HealthDCAT-AP / DCAT-AP)**
  - Execution of a structured landscape analysis of European and international discovery portals ([Annex 2](#))<sup>100</sup> to select an EU-aligned route for HEREDITARY.
  - Creation of draft HealthDCAT-AP RDF records for selected HEREDITARY datasets (HBS, UNIPD ALS cohort, UNITO ALS baseline) using the HealthDCAT-AP editor<sup>101</sup> and sandbox catalogue<sup>102</sup> provided by the European Health Information Portal (HIP).<sup>103</sup>
  - Validation and quality checking of DCAT-AP / HealthDCAT-AP records, using the HealthDCAT-AP validator embedded in its editor, data.europa.eu's Metadata Quality Assessment (MQA),<sup>104</sup> online Shapes Constraint Language (SHACL) validator,<sup>105</sup> and the FAIR-Checker web tool [\[2\]](#).<sup>106</sup>
- **Archival preparation and submission workflows**
  - Preparation of EGA submissions by creating submitter accounts, structuring metadata entities (e.g., experiments), DAC and dataset structures, and preparing the file inventory in collaboration with data controllers.
- **Publication and outreach materials**
  - Use of Zenodo<sup>107</sup> for depositing public-facing documentation (e.g. the "HEREDITARY Findability Platforms Landscape" report)<sup>108</sup> and dataset landing pages (e.g., UNIPD's ALS Longitudinal Cohort).<sup>109</sup>
  - Production of short explanatory materials (including video content) using standard video-editing software (CapCut)<sup>110</sup> to support partner engagement and explain FAIRification workflows.

## 6. Implementation and Results

In HEREDITARY, Task 3.6 operationalises the 1+MG and GDI vision of federated, standards-based genomic data sharing. To do so, we constantly work with project partners to assess their datasets and thus identify candidates for FAIRification. In this

<sup>100</sup> <https://doi.org/10.5281/zenodo.17901819>

<sup>101</sup> <https://ehds.healthdataportal.eu/editor2>

<sup>102</sup> <https://ehds.healthdataportal.eu/index.py>

<sup>103</sup> <https://www.healthinformationportal.eu/services/find-data>

<sup>104</sup> <https://data.europa.eu/mqa/methodology>

<sup>105</sup> <https://data.europa.eu/mqa/shacl-validator-ui/data-provision>

<sup>106</sup> <https://fair-checker.france-bioinformatique.fr/>

<sup>107</sup> <https://zenodo.org/>

<sup>108</sup> <https://doi.org/10.5281/zenodo.17901819>

<sup>109</sup> <https://doi.org/10.5281/zenodo.17671189>

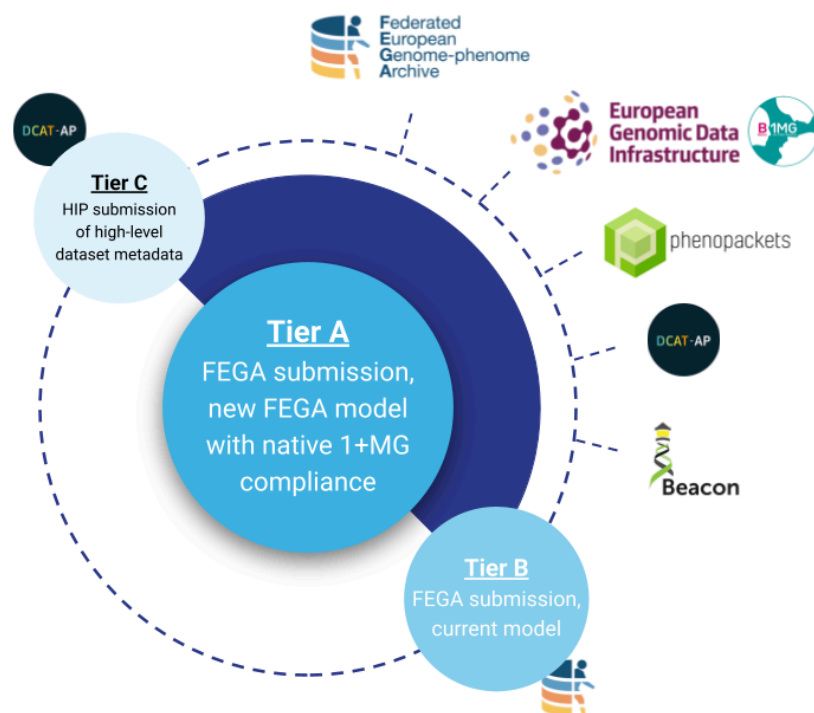
<sup>110</sup> <https://www.capcut.com/>

section, we detail how the [1+MG Standards](#) are implemented for HEREDITARY by T3.6.

Within this context, we first surveyed HEREDITARY partners and onboarded candidate data sources. This assessment yielded a **tiered view** of FAIRification opportunities ([Figure 1](#)):

- **Tier A.** Alignment with the full 1+MG stack by submitting datasets to a FEGA node adhering to the emerging FEGA Metadata Model. This approach maximises FAIRness, including long-term sustainability, findability, ease of reusability and interoperability with metadata models of Beacon, Phenopackets and HealthDCAT-AP.
- **Tier B.** Submission of suitable datasets to the EGA using its current technology stack and metadata model. Although it does not fully integrate all 1+MG recommendations, a submission to a long-term archive like EGA represents the most cost-effective way to greatly improve FAIRness of the project datasets.
- **Tier C.** Minimal but standards-based metadata exposure through indexed catalogues to ensure basic findability. This approach is designed for datasets that cannot be presently moved (i.e., submitted) to field archival but can be made discoverable in EU-wide catalogues.

*Figure 1. Summary diagram of Tiered view of FAIRification opportunities.*



The implementation effort in the reporting period up to M24 focused on the Tiers B and C, while Tier A is postulated as an approach to work on during the next period. We first identified a case study for Tier B FAIRification in RadboudUMC's "BaCo dataset",



selected because it combines rich clinical and genomic content with a realistic pathway to long-term archival in the EGA and, ultimately, to compliance with the 1+MG framework through FEAGA. Furthermore, we also picked out four datasets from project proprietary data sources to trial Tier C FAIRification.

### 6.1. Assessment of HEREDITARY datasets

The assessment of HEREDITARY datasets for Task 3.6 combined a general **call for contributions** with targeted **bilateral one-to-one discussions**. We started this process in early 2025 by contacting all 17 project institutions, enquiring about their possible datasets subject to FAIRification. To aid with the onboarding and training process, we circulated an internal "Introduction to T3.6 benefits" document ([Annex 1](#)) and a short survey,<sup>111</sup> outlining in accessible terms how FAIRification could support long-term preservation, reuse and integration of partner datasets. Although the form elicited only four responses, it helped organise the following round of interviews with data sources. Based on the diverse responses and the scope of T3.6, we initially focused further efforts on the four institutions explicitly listed in section "1.2.4 Project Data" of the Description of Action (DoA)<sup>112</sup> as proprietary data sources: RadboudUMC, UNIPD, UNITO and UCD.

Given the heterogeneous governance, sensitivity and technical readiness of these datasets, we then conducted a series of structured 1-to-1 meetings with each of the key partners to clarify what was realistically achievable in a first FAIRification round and within the span of the HEREDITARY project:

- For **RadboudUMC**, discussions covered both the "Bacteria in colitis associated cancer development" (BaCo) and the "Healthy Brain Study" (HBS)<sup>113</sup> datasets, including existing ENA submissions,<sup>114</sup> data types (e.g., genomic, clinical CSVs, imaging), current governance arrangements and the potential for EGA/FEAGA-based archival and HealthDCAT-AP catalogue records.
- At **UNIPD**, we reviewed ALS, MS and Parkinson and gut microbiome datasets described with the HERO phenoclinical ontology. Given strict governance and workload constraints, the most realistic short-term contribution was to share data dictionaries and high-level descriptions rather than relocating the underlying data.
- At **UNITO**, we explored how existing BRAINTEASER-derived resources<sup>115</sup> and registry cohorts (e.g., ALS demographics) could benefit from public-facing documentation, Zenodo-based minimal packaging and subsequent HealthDCAT-AP records.
- For **UCD**, we assessed EHR- and imaging-based Parkinson datasets held under strict institutional constraints following EPIC's guidelines. We identified

---

<sup>111</sup> <https://forms.gle/7EKVHg9HQBRP6wrz7>

<sup>112</sup> <https://hereditary.dei.unipd.it/groupoffice/index.php?r=files/file/download&id=13>

<sup>113</sup> <https://www.healthybrainstudy.nl/en/home>

<sup>114</sup> <https://www.ebi.ac.uk/ena/browser/view/PRJEB87602>

<sup>115</sup> <https://zenodo.org/records/8083181>

aggregate-level descriptions as the most realistic short-term FAIRification targets.

These dialogues with project partners elucidated that, as of the last quarter of 2025, only one dataset, BaCo, could be FAIRified following the Tier B, while the rest could belong to the Tier C.

In multiple cases the data were intertwined across research groups and institutions, and there was not a clear separation between "datasets". Thus, the decision on where to split these into formal datasets (e.g., based on phenotype and access policies) was entirely up to the data sources.

## 6.2. Enhancing discoverability of consortium data resources

In this section, we detail the work done to make HEREDITARY datasets findable, even when their sensitivity requires that **data do not leave their corresponding institutions** (i.e., Tier C FAIRification). The importance of this fundamental FAIR principle lies in the fact that datasets are often siloed, scattered across institutional black-boxes and barely documented, limiting their discoverability. Considering the great effort it takes to create them and the missed opportunity to maximise their scientific output with the help of the research community, it is crucial to improve the documentation and findability in order to augment the return on investment, especially from a funding perspective.

### 6.2.1. Selecting an EU-aligned cataloguing route

For datasets that cannot currently be moved to external archives, Task 3.6 focused on **improving discovery through catalogue-level metadata** rather than data transfer. To choose an appropriate route, we conducted a systematic landscape analysis of dataset portals and platforms, comparing scope, standards, access models and development overhead. This landscape included 13 services (e.g., EGA, BioStudies,<sup>116</sup> Maelstrom,<sup>117</sup> Zenodo,<sup>118</sup> [data.europa.eu](https://data.europa.eu))<sup>119</sup> and was documented in a comprehensive report titled "HEREDITARY Findability Platforms Landscape" (see [Annex 2](#)).<sup>120</sup>

The landscape review and project constraints pointed to a pragmatic **mix of services** rather than a single catalogue (see further details in section 5 of [Annex 2](#)). For datasets where data transfer is possible (see [Section 6.3](#)), the preferred route is to submit data and metadata to EGA or a FEAGA node, which provides controlled-access archival, persistent identifiers and portal/API-based discovery. For all HEREDITARY datasets, regardless of whether data can be moved or not, non-sensitive metadata should be described following **HealthDCAT-AP** via the **European Health Information Portal** (HIP) [3], which offers an official HealthDCAT-AP editor<sup>121</sup> and sandbox catalogue<sup>122,123</sup>

<sup>116</sup> <https://www.ebi.ac.uk/biostudies>

<sup>117</sup> <https://www.maelstrom-research.org>

<sup>118</sup> <https://zenodo.org>

<sup>119</sup> [data.europa.eu](https://data.europa.eu)

<sup>120</sup> <https://doi.org/10.5281/zenodo.17901819>

<sup>121</sup> <https://ehds.healthdataportal.eu/editor2/>

<sup>122</sup> <https://www.healthinformationportal.eu/services/find-data>

<sup>123</sup> <https://ehds.healthdataportal.eu/index.py>

for health data holders. The resulting catalogue should then be registered for **harvesting by data.europa.eu** so that records appear in the EU open-data infrastructure. On top of this, **Zenodo** and **BioStudies** can provide a citable, dataset- and study-level landing page that links to other resources (e.g., EGA accessions and HIP/data.europa.eu entries), while **Beacon v2** offers federated discovery across genomic datasets without centralising data.

For **Tier C FAIRification**, where raw data cannot be moved from the host institutions, we implement the **metadata-only** part of this mix of services: (1) describing health datasets using HealthDCAT-AP through HIP's "EU Datasets" sandbox catalogue<sup>124</sup> in preparation for the later addition to HIP's production catalogue;<sup>125</sup> (2) exposing these records for EU-wide discovery via harvesting by data.europa.eu; (3) creating per-dataset Zenodo records where applicable (i.e., if there is no further documentation to serve as landing page elsewhere); and (4) aggregating study-level metadata, including multiple datasets, as BioStudies records.

This approach gives HEREDITARY datasets standards-based, machine-actionable metadata that are compatible with emerging EHDS infrastructure, without requiring any bespoke portal development.

### 6.2.2. HealthDCAT-AP records for HEREDITARY datasets

Using the HealthDCAT-AP editor and sandbox catalogue provided by HIP, we created HealthDCAT-AP records for three HEREDITARY datasets:

- **Healthy Brain Study** (HBS, RadboudUMC).<sup>126,127</sup> Leveraging the extensive public documentation of HBS,<sup>128</sup> we compiled a HealthDCAT-AP record ([Annex 3](#)) that was later enriched with internal reviews. This effort turns an already well-documented study into a machine-readable, EHDS-compatible dataset description that can later be harvested by data.europa.eu.
- **Longitudinal ALS cohort** (UNIPD).<sup>129,130</sup> For the UNIPD Amyotrophic Lateral Sclerosis (ALS) cohort, which had no public documentation prior to this work, we drafted the HealthDCAT-AP dataset record ([Annex 4](#)) along a Zenodo submission<sup>131</sup> containing a brief dataset description, Data Access Policy ([Annex 5](#)) and Data Dictionary ([Annex 6](#)). This record acts as the first public,

<sup>124</sup>

<https://ehds.healthdataportal.eu/datasetcollection.py?cat=https://ehdsfdp.healthdataportal.eu:443/catalog/cdb4d6ad-5a31-4428-ac03-2c6e17686597>

<sup>125</sup> <https://www.healthinformationportal.eu/services/find-data>

<sup>126</sup>

<https://ehds.healthdataportal.eu/metadata.py?meta=https://ehdsfdp.healthdataportal.eu:443/dataset/8be0668e-537c-4235-85bd-b091665403d3>

<sup>127</sup> <https://ehdsfdp.healthdataportal.eu/dataset/8be0668e-537c-4235-85bd-b091665403d3>

<sup>128</sup> <https://www.healthybrainstudy.nl/en/home>

<sup>129</sup>

<https://ehds.healthdataportal.eu/metadata.py?meta=https://ehdsfdp.healthdataportal.eu:443/dataset/f3eb1cb8-5b06-4c64-8a45-286ad3edfee7>

<sup>130</sup> <https://ehdsfdp.healthdataportal.eu:443/dataset/f3eb1cb8-5b06-4c64-8a45-286ad3edfee7>

<sup>131</sup> <https://doi.org/10.5281/zenodo.17671190>

citable description of the cohort, greatly improving its findability while leaving the underlying data under institutional control.

- **ALS demographics and baseline data (UNITO).** <sup>132,133</sup> Similarly, this dataset did not have any public documentation, and thus we created not only a HealthDCAT-AP record ([Annex 7](#)), but also a complementary Zenodo submission, <sup>134</sup> including its dataset description, Data Access Policy ([Annex 8](#)) and Data Dictionary ([Annex 9](#)). Like the UNIPD cohort, this approach covers both machine- and human-readable records as discoverable entry points for future reuse.

In addition to these three mature dataset records, similar work has been started for UCD's Colorado Ophthalmology Research Information System (CORIS) dataset. Nevertheless, due to time constraints, this draft was not included in this report, but will instead be covered in the next FAIRification Deliverable Report D3.7.

All three released records follow the HealthDCAT-AP specification as per its editor. It is thus expected, but also easy to check through data.europa.eu's SHACL validator tool, <sup>135</sup> that the three dataset files ([Annex 3](#), [Annex 4](#) and [Annex 7](#)) comply with DCAT-AP (v3.0.0).

Using the Metadata Quality Assessment (MQA) <sup>136</sup> service provided by data.europa.eu for DCAT-AP / HealthDCAT-AP metadata, we can quickly evaluate the quality and utility of each dataset's metadata based on FAIR-guided dimensions (see [Figure 2](#)). This serves as an objective measure of FAIRness for both project members to compare new updates with and requesters to know the quality of the dataset before going through the process of gaining access. Eventually, these metrics for each dataset will be part of the overall data.europa.eu's MQA dashboard <sup>137</sup> and will be available as accessible metrics for HEREDITARY's datasets.

---

<sup>132</sup>

<https://ehds.healthdataportal.eu/metadata.py?meta=https://ehdsfdp.healthdataportal.eu:443/dataset/caf55cc7-0205-4956-a1ab-d5e8008e5a97>

<sup>133</sup> <https://ehdsfdp.healthdataportal.eu:443/dataset/caf55cc7-0205-4956-a1ab-d5e8008e5a97>

<sup>134</sup> <https://doi.org/10.5281/zenodo.17779232>

<sup>135</sup> <https://data.europa.eu/mqa/shacl-validator-ui/data-provision>

<sup>136</sup> <https://data.europa.eu/mqa/methodology>

<sup>137</sup> <https://data.europa.eu/mqa/>

**Figure 2.** Radar Plots of Metadata Quality Assessments of the project's Tier C FAIRified datasets.



In parallel to the MQA analysis, the HIP Sandbox Catalogue also exposes dataset records through the **EHDS FAIR Data Point (FDP)**,<sup>138</sup> a metadata service for publishing machine-actionable metadata in line with the FAIR principles.

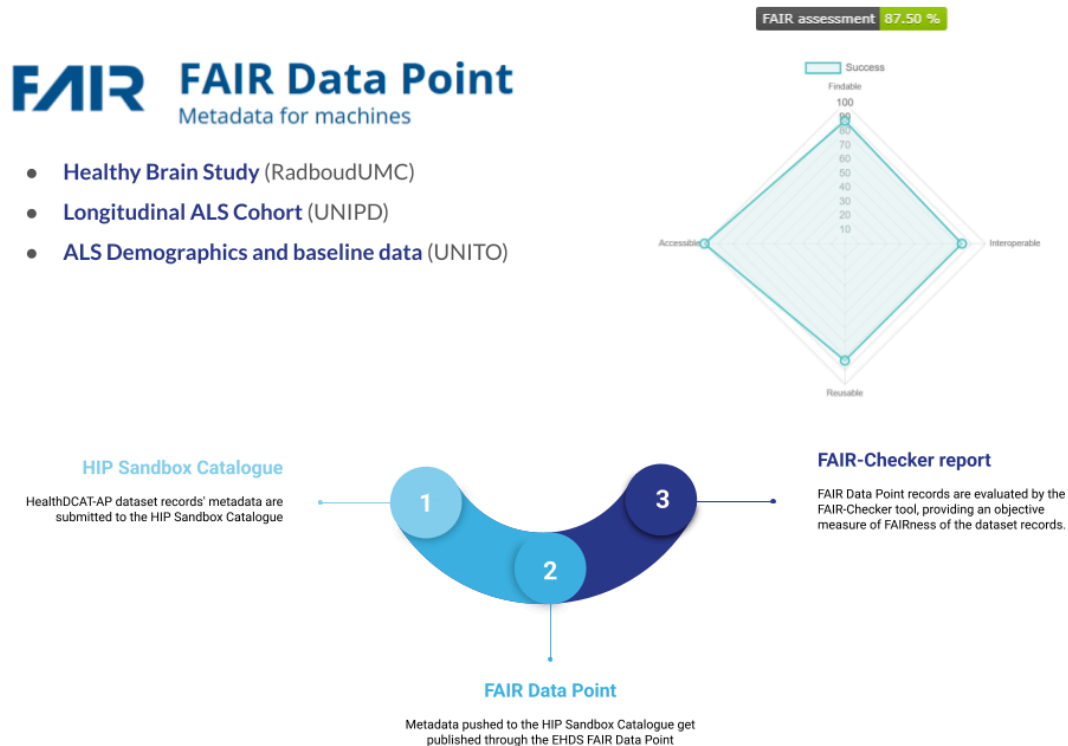
As a complementary objective measure of FAIRness of the EHDS FDP, we submitted the three FDP dataset record Uniform Resource Identifiers (URI) to the **FAIR-Checker** web tool<sup>139</sup> using its "Check" feature.<sup>140</sup> FAIR-Checker harvests embedded semantic annotations from the web pages and evaluates a set of FAIR metrics to test compliance with selected FAIR principles. The resulting FAIR-Checker reports for the three HEREDITARY records (see [Figure 3](#)) were effectively identical, each yielding an **overall FAIRness score of 87.50%**. This suggests that, in our current configuration, FAIR-Checker primarily reflects properties of the shared FDP environment, while confirming that the HEREDITARY records are exposed on the web with a consistent FAIRness profile.

<sup>138</sup> <https://ehdsfdp.healthdataportal.eu/catalog/cdb4d6ad-5a31-4428-ac03-2c6e17686597>

<sup>139</sup> <https://fair-checker.france-bioinformatique.fr/>

<sup>140</sup> <https://fair-checker.france-bioinformatique.fr/check>

Figure 3. EHDS FAIR Data Point FAIRness assessed by FAIR-Checker.



As introduced above, these three records do not cover the full extent of HEREDITARY project datasets, but were deemed as the **best starting point for the implementation** of the FAIRification workflow. Furthermore, the dataset records are not to be considered fully finished products, but rather an **evolving documentation**: details (e.g., per-dataset DUO codes) will be added, amended or removed as the HEREDITARY project and EU standards develop. Consequently, records can be edited after submission to the HIP. This allows project members and T3.6 to further increase the fidelity with which the project datasets are exposed to the public.

We emphasise **training collaborators** to enable their autonomy in independent FAIRification. In line with the aforementioned materials, we also prepared a brief video tutorial ([Annex 10](#)) to quickly get project colleagues up to speed on how to create their own HealthDCAT-AP dataset records through the editor.

### 6.2.3. Current status and next steps

By the end of the M24 reporting period (December 2025), the three HealthDCAT-AP dataset records were: (1) **generated and validated through the editor**, and then **publicly exposed** in the HIP sandbox catalogue and its associated EHDS FDP.

Next steps are to: (1) push dataset records to the production catalogue of HIP; (2) assert HIP records are harvested by data.europa.eu; (3) maintain and review the existing records in scheduled rounds of reviews with project collaborators; and (4) add more datasets to the project's dataset repertoire. These last two steps will include



adding new datasets from already onboarded institutions (e.g., UNITO), completing the documentation of their dataset collection, and finalising dataset records of other institutions (e.g., UCD's CORIS dataset). Depending on the amount of subscribed collaborators and available resources, we may be able to include non-sensitive datasets (e.g., linguistic datasets from Task 3.3) before the next reporting period (M36).

Additionally, as per the above introduced landscape results, we will start compiling all dataset records into **structured studies** in BioStudies, following the preference of each data contributor.

In this way, Tier C FAIRification provides a **lightweight but coherent path for HEREDITARY datasets to enter the wider EHDS and EU open-data ecosystem**, even when the underlying data cannot yet be transferred to community archives.

## 6.3. FAIRification of RadboudUMC's BaCo Dataset

### 6.3.1. Dataset description

The "Bacteria in colitis associated cancer development" (BaCo) dataset is a **multimodal inflammatory bowel disease (IBD) cohort** collected at RadboudUMC. It revolves around ulcerative colitis patients and matched controls. It includes faecal samples and multiple colonic biopsies per participant, together with detailed clinical information (diagnosis, demographics, medication and treatment, oncological traits, biofilm status, etc.) and several microbiome data types (16S rRNA, shotgun metagenomics, assemblies and BIOM summary tables).

A subset of the non-human reads is already publicly archived at the European Nucleotide Archive (ENA) under accession **PRJEB87602**,<sup>141</sup> while **human-derived data** and rich clinical metadata are planned for controlled-access **deposition in the EGA** (see [Figure 4](#)). In HEREDITARY, BaCo therefore represents a realistic Tier B FAIRification case: a dataset that can be moved to a long-term archive, while keeping non-human components in open repositories and using shared BioSample<sup>142</sup> sample identifiers (e.g., SAMEA118253018)<sup>143</sup> to cross-link records across archives. See a summary of the dataset in these archives in [Figure 5](#).

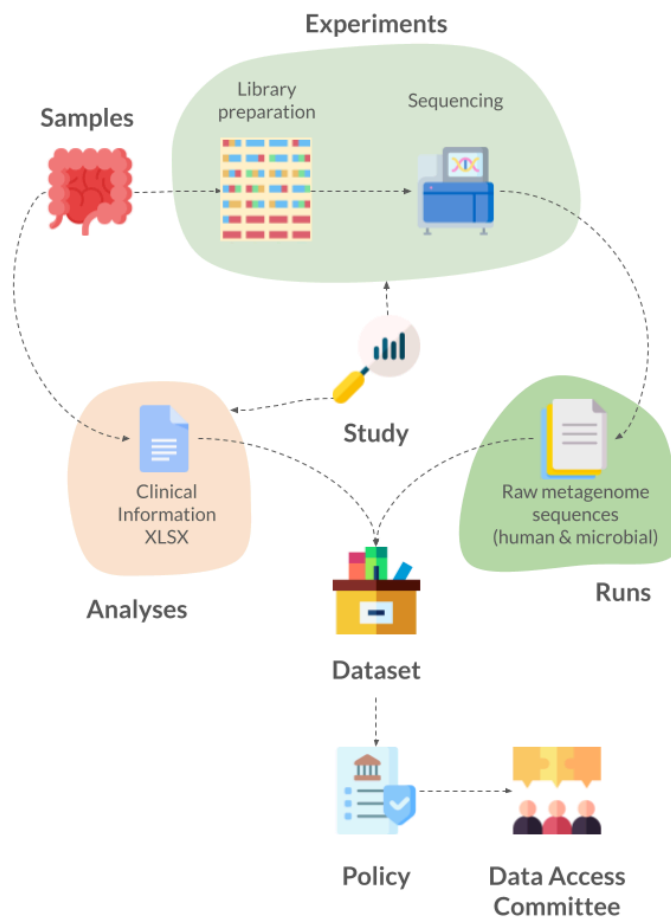
---

<sup>141</sup> <https://www.ebi.ac.uk/ena/browser/view/PRJEB87602>

<sup>142</sup> <https://www.ebi.ac.uk/biosamples/>

<sup>143</sup> <https://www.ebi.ac.uk/ena/browser/view/SAMEA118253018>

**Figure 4.** Overview of content of the BaCo dataset submitted to EGA, following the archival model.



*Icons made by Freepik, obtained from Flaticon.*



The flowchart illustrates the study design for the European Genome-Phenome Archive (EGPA). It shows the process from patient samples to data processing and storage. The process starts with a patient providing fecal samples and colon biopsies (2-3 per patient). These samples are used for library preparation and sequencing. The resulting raw metagenome sequences (human & microbial) are processed into processed metagenome (human & microbial). The processed metagenome is then used to generate a summary table BIOM, assembly, and microbial FASTQ files. The summary table BIOM is linked to the patient's clinical information (XLSX). The processed metagenome is also linked to the European Genome-Phenome Archive (EGPA) and the European Nucleotide Archive (ENA) (PRJEB87602).

### 6.3.2. Standards applied

Some free-text elements (e.g., medication names) were normalised into controlled lists and grouped by therapeutic class, and outcome variables were aligned with standard disease and procedure codes. The resulting dictionary transforms the original free-text metadata into an **ontology-backed** structure that can be reused both in the EGA submission templates and in future FEGA- and Phenopackets-based representations.

In parallel, we defined a **subject-sample linkage model** (see [Figure 5](#)) in which each biopsy and faecal sample (record in ENA) is tied to a stable subject identifier (record in EGA). Both archives register corresponding BioSamples accessions, enabling a linkage between the two through the data subject record in BioSamples. For example, semantically tagging that the "*microbial sample*" record (in ENA) and the "*colon biopsy*" record (in EGA) derive from the same "*study subject*" record (in EGA). This will allow

the EGA dataset record, when released, to be cross-referenced with the existing ENA study, so that users can discover both human and non-human components of BaCo across archives. The metadata of the dataset have been structured against the current EGA metadata schema (e.g., Samples, Experiments), while a richer FEGA-model-compliant record of BaCo will be developed in the following year (2026), further FAIRifying the dataset up to Tier A.

### 6.3.3. Preparation for submission to EGA

On the operational side, an **EGA submitter** account was created for the RadboudUMC lead (Annemarie Boleij), who was then guided through the EGA submission workflow in several one-to-one sessions.

EGA requires submitters to sign a standard **Data Processing Agreement** (DPA) that sets out the roles of the Data Controller and EGA as Data Processor and describes how personal data are handled under EU General Data Protection Regulation (GDPR).

In preparation for the submission, we put together the **file inventory** (raw FASTQ, decontaminated reads and clinical spreadsheets) that would be hosted at EGA. This involved mapping terms, complying with filename conventions consistent with subject and sample identifiers, and outlining the encryption and upload steps.

The **planned submission pathway** follows the standard EGA sequence,<sup>144</sup> including: (1) DPA (v1.5, 2025)<sup>145</sup> signature, providing the legal framework for sensitive data processing by EGA, thereby clearing the way for submission; (2) DAC registration;<sup>146</sup> (3) local encryption of files through Crypt4GH;<sup>147</sup> (4) secure upload to the EGA inbox/FTP;<sup>148</sup> (5) metadata registration for all entities through the Submitter Portal (SP);<sup>149</sup> and (6) final EGA curation leading to the public release of the dataset.

Throughout, we designed the **governance model** so that a RadboudUMC-based DAC can evaluate access requests, consistent with EGA's requirements. This request and evaluation system of EGA is one of the key benefits for collaborating data controllers: not only are long-term data storage and distribution resolved, but also accepting or rejecting data access requests becomes more efficient and transparent, enabling quick reuse of released datasets.

### 6.3.4. Outcome, limitations and next steps towards FEGA

By M24 (December 2025), the **technical preparation** for submission was largely complete: the data dictionary was ontology-mapped, subject-sample linkages and file inventories were in place, and EGA-compatible metadata had been drafted.

<sup>144</sup> <https://ega-archive.org/submission/quickguide/>

<sup>145</sup> [https://ega-archive.org/assets/files/EGA\\_Data\\_Processing\\_Agreement\\_for\\_information.pdf](https://ega-archive.org/assets/files/EGA_Data_Processing_Agreement_for_information.pdf)

<sup>146</sup> <https://ega-archive.org/access/data-access-committee/what-is-dac/>

<sup>147</sup> <https://ega-archive.org/submission/data/file-preparation/crypt4gh/>

<sup>148</sup> <https://ega-archive.org/submission/data/uploading-files/inbox/>

<sup>149</sup> <https://ega-archive.org/submission/metadata/submission/sequencing-phenotype/submitter-portal/>

Nevertheless, the actual submission to EGA was delayed by **legal and ethical blockers**: DPA signature and Informed Consent review.

The diligent **review and sign-off of EGA's DPA** by RadboudUMC's legal office took a total of 6 months, and was recently completed in November 2025.

Additionally, there were concerns that the original **BaCo informed consent**, drafted before GDPR, might not explicitly cover international archival and controlled-access sharing. RadboudUMC's Medical Ethics Review Committee, Medisch-ethische toetsingscommissie (METC), <sup>150</sup> dutifully reviewed the dataset's consent form and concluded that the wording is adequate under GDPR Article 9(2)(a), <sup>151</sup> which allows processing of special-category data, including health and genetic data, where the data subject has given explicit consent. This approval by the METC was provided by the end of November 2025.

Following these two decisions, RadboudUMC's EGA submitter account for the BaCo dataset was activated, allowing for the submission to proceed in line with EGA's data processing legal framework.

The BaCo submission preparation has nevertheless highlighted **limitations of the current central EGA metadata model** when applied to complex multimodal human studies. For example: *individuals* are only implicitly represented via samples; detailed treatment trajectories must be flattened into generic attributes; and there is no native way to express DCAT-AP or HealthDCAT-AP catalogue views, or GDI/Beacon/Phenopackets-ready representations within a single coherent record. Ongoing FEGA metadata work addresses precisely these gaps by introducing a new standardised, JSON-LD-based model with explicit hooks to 1+MG Framework standards (e.g., DCAT-AP).

Once the BaCo EGA deposition is complete, our plan is to re-express the cohort in this unified **FEGA model**, using the BaCo dataset as a spearhead example of FAIRification in the FEGA/1+MG Network. This will turn BaCo from a Tier B EGA case into an early Tier A demonstrator, showing how a real-world multidisciplinary IBD cohort dataset can be FAIRified consistently across ENA, EGA/FEGA and European discovery services. We expect these efforts to be reported by M36 in D3.7.

#### 6.4. Genomic data quality protocol

The National Center for Genomic Analysis (CNAG) is codifying a **genomic data-quality protocol** and assembling internal and external datasets for **harmonisation** and subsequent **analysis**, building on established CNAG pipelines for QC and variant calling. Concretely, CNAG maintains three open-source toolchains used in this protocol:

---

<sup>150</sup>

<https://www.radboudumc.nl/patientenzorg/uw-afspraak/meer-informatie/patient-in-een-umc/mee-doen-aan-onderzoek/alles-over-meedoen-aan-wetenschappelijk-onderzoek/medisch-ethische-toetsingscommissie>

<sup>151</sup> <https://gdpr-info.eu/art-9-gdpr>

1. **CBICall**, a framework that takes paired-end FASTQ and produces variant calls (VCF/gVCF) within a reproducible variant-calling pipeline based on GATK best practices.<sup>152</sup>
2. **Beacon2-cbi-tools**, which validates/annotates inputs and converts VCF/microarray data into Beacon-Friendly Format and loads them for Beacon v2 implementations.<sup>153</sup>
3. **Convert-Pheno**, which harmonises clinical/phenotypic data and interconverts Beacon v2, Phenopackets v2, and OMOP CDM data exchange formats [4].<sup>154</sup>

Collectively, these activities emphasise Interoperability and Reusability, improving data quality and easing cross-node harmonisation within HEREDITARY and leading to a better integration in European initiatives aligned with the 1+MG framework. Additionally, they help with the active beaconisation of project data resources undertaken by T3.4.

Work is ongoing. Consortium-reported operational challenges (timely dataset acquisition from project institutions and coordinating across an international network of nodes) have slowed the path to completion. Nevertheless, next steps are clear: finalise and submit the data-quality protocol for publication, and put the QC and harmonisation pipeline to the test based on the above repositories.

## 6.5. ONTO's FAIRification

### 6.5.1. LinkedLifeData Inventory

Ontotext's (ONTO) **LinkedLifeData Inventory** (LLDI)<sup>155</sup> provides unified, RDF-based access to a large catalogue of life-science and healthcare datasets and ontologies, using shared vocabularies, persistent identifiers and FAIR-aligned metadata to support cross-dataset linking and querying.

The LLDI framework offers a consistent, standards-based environment for integrating diverse biomedical datasets, from molecular to clinical. This unification ensures semantic alignment, simplifies cross-dataset linking, and supports efficient querying and reuse. Consequently, LLDI strengthens HEREDITARY's ontology construction, ensures knowledge maintainability, and facilitates downstream analytical workflows for healthcare objectives.

### 6.5.2. FAIRification workflow (Apache Airflow)

For each LLDI-processed dataset, ONTO orchestrates a reproducible pipeline in **Apache Airflow**,<sup>156</sup> which uses Directed Acyclic Graphs (DAG),<sup>157</sup> with the following steps:

<sup>152</sup> <https://github.com/CNAG-Biomedical-Informatics/cbicall>

<sup>153</sup> <https://github.com/CNAG-Biomedical-Informatics/beacon2-cbi-tools>

<sup>154</sup> <https://github.com/cnag-biomedical-informatics/convert-pheno>

<sup>155</sup> <https://www.ontotext.com/solutions/healthcare-and-life-sciences/linked-life-data-inventory>

<sup>156</sup> <https://airflow.apache.org/>

<sup>157</sup> <https://airflow.apache.org/docs/apache-airflow/2.5.2/core-concepts/dags.html>

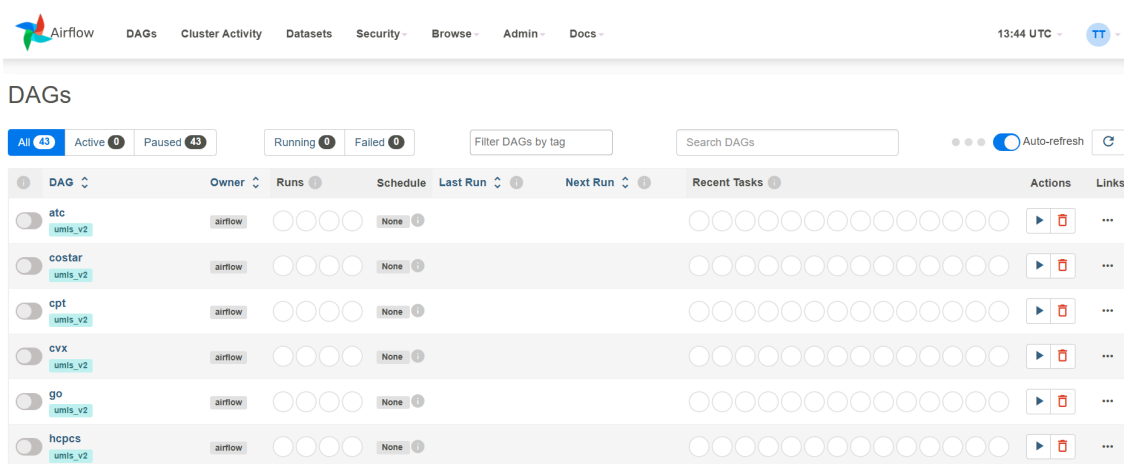
- **Step 1:** The data pipeline begins with data download, where the data is retrieved from its source and stored in a dedicated secure location.
- **Step 2:** The data is assigned a new version (new\_data).
- **Step 3:** If necessary, it undergoes transformation to RDF format. This transformation step is omitted for data that is already in RDF format.
- **Step 4:** Integration tests are conducted to verify the correctness of the final data format, preventing erroneous data entry.
- **Step 5:** Following successful integration tests, metadata is generated and verified to ensure compliance with relevant standards.
- **Step 6:** Deployment of the data to another secure storage location.

Airflow DAGs provide scheduling, monitoring and reuse across datasets (see [Figure 6](#) for the DAG step list and [Figure 7](#) for example DAGs in the UI).

*Figure 6. Steps in Apache Airflow Pipeline.*

data_download
branching
new_data
changed_dag
no_processing
integ_tests
rdf_validate
create_void
add_dcat
validate_dcat
deploy

*Figure 7. Apache Airflow DAGs.*



DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	Links
atc umis_v2	airflow	0	None				▶ 🗑	...
costar umis_v2	airflow	0	None				▶ 🗑	...
cpt umis_v2	airflow	0	None				▶ 🗑	...
cvx umis_v2	airflow	0	None				▶ 🗑	...
go umis_v2	airflow	0	None				▶ 🗑	...
hpcps umis_v2	airflow	0	None				▶ 🗑	...

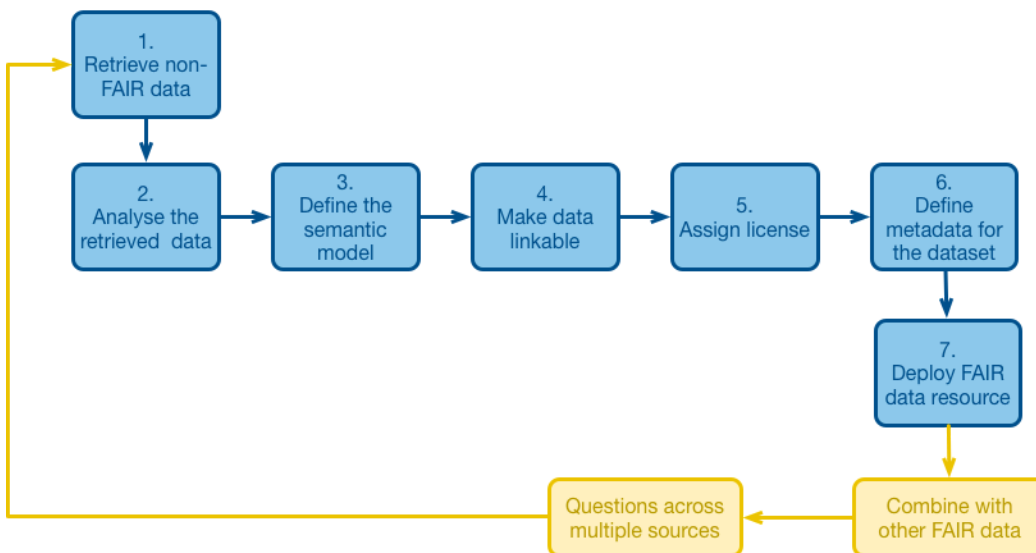
Tabular sources (e.g., CSV) are mapped to RDF using Tarql ("SPARQL for Tables"),<sup>158</sup> enabling declarative, testable transformations via SPARQL 1.1.

### 6.5.3. FAIRification of Expression Atlas

The Gene Expression Atlas,<sup>159</sup> developed by EMBL-EBI, is a resource that centralises and curates data from high-throughput transcriptomics experiments, like RNA-seq and microarrays. It offers a standardised, queryable interface to view gene expression patterns, including baseline and differential expression results, across various biological conditions, tissues, cell types, and species.

Following the FAIRification process (see [Figure 8](#)), ONTO retrieves Expression Atlas in tabular form, analyses entities and relations, and defines a semantic model (see [Figure 9](#)).

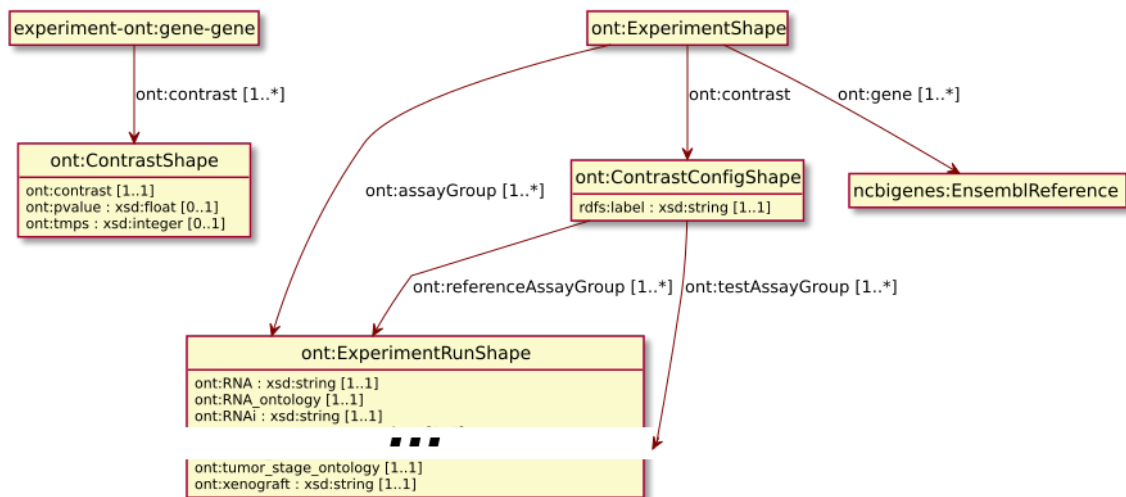
**Figure 8.** FAIRification principles. Source: <https://www.go-fair.org/fair-principles/fairification-process>.



<sup>158</sup> <https://tarql.github.io/>

<sup>159</sup> <https://www.ebi.ac.uk/gxa/home>

Figure 9. Excerpt from Expression Atlas semantic model.



Then data is transformed to RDF-star using Tarql tool <sup>160</sup> to capture statements about other statements (e.g., contrasts and statistics, see [Table 1](#)).

Table 1. Source Data in CSV format (as a table) and its transformation to RDF-star.

Gene ID	Gene Name	g2_g1.p-value	g2_g1.log2foldchange
ENSG000000000003	TSPAN6	0.489158292454121	0.4

```

experiment:E-CURD-45 ont:gene ensembl_id:ENSG000000000003 .

<<experiment:E-CURD-45 ont:gene ensembl_id:ENSG000000000003>>
  ont:contrast
<https://linkedlifedata.com/resource/experiment/E-CURD-45/ENSG000000000003/g2_g1> .

<https://linkedlifedata.com/resource/experiment/E-CURD-45/ENSG000000000003/g2_g1>
  ont:pvalue      "0.489158292454121"^^xsd:float ;
  ont:log2foldchange "0.4"^^xsd:float ;
  ont:experimentGroup
<https://linkedlifedata.com/resource/experiment/E-CURD-45/g2_g1> .

```

<sup>160</sup> <https://tarql.github.io/>



Once in RDF-star format, metadata is produced using DCAT (v2,<sup>161</sup> for catalogue-level description) and VoID (for dataset structure and interlinks).<sup>162</sup> Finally, metadata is released.

#### 6.5.4. Datasets prioritised for HEREDITARY

- **Expression Atlas.** Curated baseline and differential gene/protein expression across species and biological conditions, with a public, queryable interface and regular releases.
- **NCBI Gene.**<sup>163</sup> Integrates gene-centred information across species (nomenclature, RefSeqs, pathways, variation, phenotype) and links to genome/phenotype resources worldwide.
- **ClinVar.**<sup>164</sup> Public archive of human variations classified for diseases and drug responses, including supporting evidence. It improves access to and communication about the asserted relationships between human variation and observed conditions, along with the history of those assertions.
- **Unified Medical Language System (UMLS).**<sup>165</sup> A set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems.

[Figure 10](#) illustrates the connectivity within the NCBI Gene dataset. Each segment represents an RDF class, imported from sources such as UMLS (semnet:T-codes) or Clinvar (clinvar:Variant). The ribbons connecting the segments represent the number of RDF statements linking instances between two classes. The thickness of each ribbon corresponds to the volume of those relations.

<sup>161</sup> <https://www.w3.org/TR/vocab-dcat-2/>

<sup>162</sup> <https://www.w3.org/TR/void/>

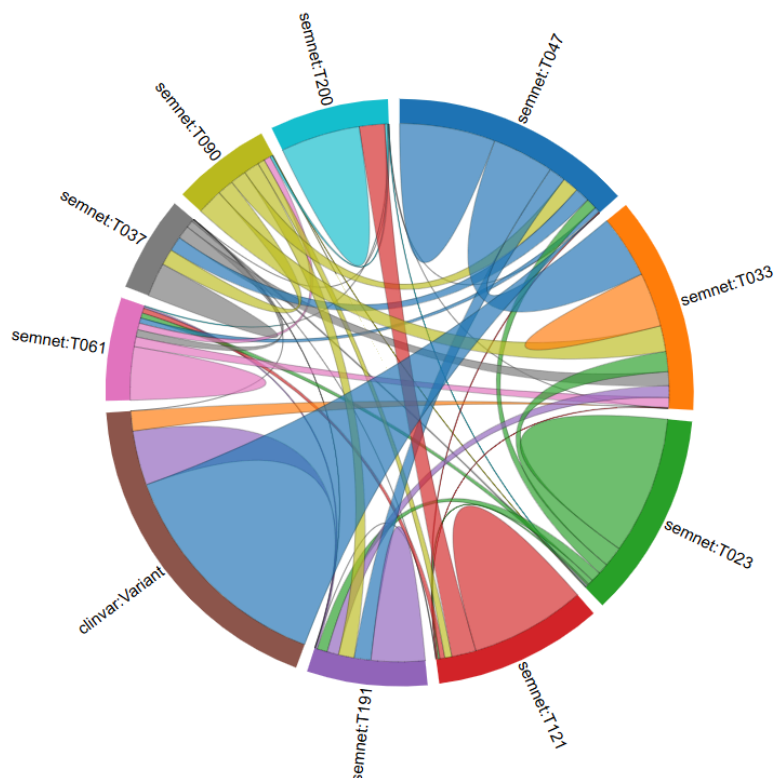
<sup>163</sup> <https://www.ncbi.nlm.nih.gov/gene/>

<sup>164</sup> <https://www.ncbi.nlm.nih.gov/clinvar>

<sup>165</sup> <https://www.nlm.nih.gov/research/umls/index.html>



Figure 10. Class relationships.



The selected biomedical resources (NCBI Gene, ClinVar, UMLS, and Expression Atlas) collectively offer a comprehensive and interoperable foundation for HEREDITARY's goals. NCBI Gene and Expression Atlas provide molecular and expression data; ClinVar contributes curated clinical variant-phenotype associations; and UMLS ensures semantic consistency by harmonising terminology. Their combined use establishes a robust, standards-aligned basis for developing ontologies and representing complex biological and clinical relationships.

## 6.6. Analysis of AI Act Requirement of Completeness for Training Datasets

Katholieke Universiteit Leuven (KUL) contributed a legal-doctrinal analysis of the AI Act's **dataset-quality requirement of "completeness"** under Article 10(3) of the AI Act.<sup>166,167</sup> The work interprets the operative text and surrounding context to outline what, in practice, "complete in view of the intended purpose" could require of training, validation and testing datasets, addressing how providers can evidence compliance.

<sup>166</sup> Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) [2024] OJ L 12.7.2024.

<sup>167</sup> <https://artificialintelligenceact.eu/article/10/>

A paper on this topic, co-authored by Elisabetta Biasin <sup>168</sup> and Donatella Casaburo <sup>169</sup> and currently under review, was presented at the Privacy Law Scholars Conference (PLSC) Europe 2025 <sup>170</sup> on the 23rd of October 2025 under the title "The AI Act's Requirement of Completeness for Training Datasets: Can Historical Data Comply?".

For **T3.6 FAIRification**, these insights provide a governance lens for **data quality** that complements our technical curation. They inform how we document dataset scope and gaps, justify quality enrichment steps, and record assumptions and provenance so that datasets are more reusable and interoperable in line with FAIR principles.

## 7. Challenges

- **Limited availability of datasets for FAIRification.** This commonly stems from two main issues:
  - **Data privacy.** There are cases where data privacy and/or signed consents do not allow for the full extent of FAIRification (e.g., storing data in a mature long-term archival). In such scenarios, some FAIRification threads may be limited, reducing their scope to Tier C. This also rules out simpler open-data routes, and emphasises the importance of controlled-access archives such as EGA/FEGA and future Health Data Access Body (HDAB)-mediated access under the EHDS.
  - **Unfamiliarity with the FAIR principles.** Despite FAIR not being a synonym of "open" or "public" data, these interpretations of the FAIR principles persist and persuade data controllers to be overly cautious. In this sense, responses to the FAIRification efforts by the respective dataset controllers are varied, reducing the amount of proprietary datasets that can be used as FAIRification use-cases.
- **Coordination in an international consortium.** Working across multiple institutions, languages and governance cultures has made it difficult to keep all contributors aligned and informed, especially when FAIRification is not a core deliverable for most partners.
- **Legal bottlenecks.** For the BaCo dataset, EGA submission required institutional review and signature of EGA's DPA, as well as a fresh assessment of pre-GDPR informed consent by the METC. These steps, although necessary, added several months of delay between technical readiness and the start of actual data deposition.
- **Standards and infrastructure still in flux.** Core components of the ecosystem, such as the GDI standards, HealthData@EU infrastructure and national HDABs, are still being designed or rolled out, so there is no fully stable

---

<sup>168</sup> <https://orcid.org/0000-0001-9090-3315>

<sup>169</sup> <https://orcid.org/0000-0002-3548-6755>

<sup>170</sup> <https://plsc-europe.eu/plsc-2025/>

end-to-end stack to target. This makes some HealthDCAT-AP fields (e.g., HDAB references) impossible to fill at present.

- **Evolving findability tools and overlapping approaches.** Within HEREDITARY there are parallel efforts around Beacon, polystore-style architectures (i.e., HDN) and catalogue-based discovery. Aligning these without duplicating work is non-trivial and requires continuous coordination.
- **Limitations of current EGA metadata model for multimodal datasets.** For example, the lack of native support for HealthDCAT-AP views, which hinders automatic metadata harvesting by data.europa.eu. This constrains how richly BaCo and similar cohorts (Tier B) can be described until the FEGA model is available.
- **Intensive effort needed to map free text to ontologies.** Data dictionaries for BaCo and other datasets started as free-text spreadsheets with heterogeneous coding. Cleaning, normalising and mapping fields and values to recommended ontologies is labour-intensive and requires domain input from clinicians as well as bioinformaticians. Depending on the amount of datasets subject to FAIRification during the lifespan of the process, this approach may not scale.
- **Variable data quality and documentation across partners.** Some datasets lacked clear dataset boundaries, stable identifiers, or separate technical documentation. In several cases the first HealthDCAT-AP record or Zenodo landing page created under T3.6 is the only public documentation. This heterogeneity complicates harmonisation and reuse.
- **Administrative centralisation of FAIRification work in T3.6.** At present, most updates to HealthDCAT-AP records and metadata mappings have to pass through a single T3.6 contributor, which does not scale as more datasets come on board. This motivates the creation of further guidelines and examples, to serve as training materials so that local data stewards can FAIRify and maintain their own records.
- **Uncertain long-term sustainability of third-party catalogues.** Key Tier C components such as the European Health Information Portal (HIP) and its HealthDCAT-AP editor are project-driven services whose long-term funding model is still being clarified. That is the case even when they are advertised as central EU-wide entry points for HealthDCAT-AP metadata. This raises questions about where HEREDITARY records should ultimately be hosted if pilot services are restructured. Therefore, a delicate balance between relying on external service providers and the technical overhead of creating a bespoke HEREDITARY catalogue appears. This will be extremely relevant for the long-term sustainability plan expected at "Final exploitation plan and final IPR strategy" (D8.6) led by Task 8.3.
- **Resource and time constraints for creating example or sample datasets.** HealthDCAT-AP and some discovery platforms encourage providing "Sample Datasets" or synthetic examples. Preparing privacy-preserving samples for

sensitive human data is costly and outside the capacity of many data providers within the HEREDITARY timeframe. It is for that matter that, for the HealthDCAT-AP records, we treated the data dictionaries as the current "dataset samples".

- **Time-consuming access granting model at institutions.** The [assessment of HEREDITARY datasets](#) and subsequent communication with project partners acting as DACs highlighted the extremely arduous process they go through to assess data access requests when there is no standard access method in place. This is the common case where proprietary datasets are siloed and hosted in their respective institutions. By using BaCo as a concrete example of added short-term value and expanded time to focus on collaborators' primary research, T3.6 aims to demonstrate benefits that may help overcome scepticism for Tier B FAIRification and above in subsequent project years.
- **Heterogeneous standards across data sources.** As an international, multi-disease consortium, HEREDITARY partners use different local formats, coding systems and ontologies (or none at all) to represent their metadata (e.g., clinical variables). This heterogeneity makes horizontal harmonisation difficult, increases the cost of mapping to recommended standards and limits practical interoperability between datasets. Albeit T3.6 tries to mitigate this laborious task relying on clinicians and data curators, institutional, project and service constraints may prevail.

## 8. Discussion

The FAIRification activities in HEREDITARY have primarily improved the **Findability of key consortium datasets**, while also laying foundations for better **Accessibility, Interoperability and Reusability** in line with the FAIR Principles. By preparing the submission of the BaCo dataset to EGA, and by drafting HealthDCAT-AP records for RadboudUMC's HBS, UNIPD's ALS cohort and UNITO's ALS baseline datasets, Task 3.6 has moved several HEREDITARY resources from being **essentially invisible outside their host institutions** to being ready for **inclusion in European catalogues** and archives that are explicitly designed for cross-border reuse. Even before further refinements are done, ontology-backed data dictionaries, clear access policies and catalogue records already provide tangible benefits: they enable discovery of these resources, clarify governance, reduce ad-hoc email exchanges about "what the dataset actually contains", and give data controllers ready-to-use material for grant applications and institutional impact assessments.

At the same time, the work has exposed the **practical limits of what can be achieved within one project cycle**, especially when dealing with sensitive human data and a heterogeneous consortium. Only one dataset (BaCo) could realistically progress to **Tier B FAIRification** within the scope of D3.6. But, as is often the case when dealing with human sensitive data, there were prolonged delays in order to resolve ethical and legal questions which arose when preparing the dataset for submission. For other datasets, strong governance constraints and limited partner capacity meant that we had to shift from an initial focus on data transfer towards a **metadata-only Tier C approach**. This

work centred on documenting datasets through Zenodo and HealthDCAT-AP catalogue records, alongside future study-level documentation at BioStudies.

The [BaCo case study](#) illustrates both the **value and the friction of deep FAIRification**. Ontology mapping of free-text clinical fields, cross-linking human and microbial components via BioSamples, and designing an EGA+FEGA-ready metadata model are labour-intensive steps, but they yield a dataset that is more intelligible to external users and far easier to integrate into federated infrastructures such as the 1+MG Framework. For the collaborators at RadboudUMC, this investment is not purely altruistic: having a **well-described, archive-ready cohort** strengthens future funding applications, eases DAC maintenance, asserts dataset long-term sustainability, and supports institutional downstream studies. In the future, they can point to robust governance, clear reuse conditions and alignment with European standards rather than to an internal spreadsheet and local file share system.

An important distinction is that **discoverability does not equate to open access**. By using EGA/FEGA for sensitive genomic data and HealthDCAT-AP for high-level descriptions, HEREDITARY can expose rich metadata and standardised access information without weakening local access controls or ownership. This privacy-preserving discoverability of project datasets is not only aimed towards a selfless scientific devotion, but also to increase institutional collaboration, research outputs and funding sustainability of each participating team. This is directly in line with the design of the EHDS, where HealthDCAT-AP is the recommended profile for describing datasets that remain under HDAB- or controller-mediated access.

In addition, the work underscores the **need for a more coherent, end-to-end infrastructure for human-derived data**. Complex, sensitive, and multimodal datasets are quickly becoming common in EU-driven initiatives, yet they lack a fully encompassed system that: (1) supports dataset discovery, controlled access and long-term sustainability; (2) faithfully represents their metadata; and (3) is compatible with international standards, such as 1+MG Framework's (e.g., HealthDCAT-AP) and GA4GH's (e.g., Beacon). This is what we refer to here as Tier A FAIRification, and what we expect to achieve by implementing the FEGA Metadata Model with project datasets like BaCo in the next reporting period.

Beyond the HEREDITARY-owned datasets, **ONTO's FAIRification pipelines** over LinkedLifeData Inventory and **CNAG's genomic data quality protocol** show how the project is also investing in reusable infrastructure and quality frameworks. These efforts operationalise FAIR principles at scale: LLDI and its Airflow-based workflows demonstrate how heterogeneous **public resources can be FAIRified** in a repeatable way, while CNAG's QC and harmonisation toolchains provide concrete, shareable practices for ensuring interoperable, high-quality genomic data across nodes.

In parallel, **KU Leuven's legal-doctrinal analysis of the AI Act's completeness** requirement for training datasets adds a governance layer to these technical workflows. It clarifies how documenting dataset scope, gaps, provenance and quality-enrichment steps can support future compliance with Article 10(3) when HEREDITARY datasets or workflows are reused in high-risk AI systems.

On this basis, our main recommendations are:

- Data-producing projects should **plan for FAIRification in the early stages**, budget for the effort required based on their datasets, and adopt community standards rather than local ad-hoc formats.
- To raise awareness of FAIR datasets, with the aim to change the culture around data reuse in the scientific community. Institutional data stewards should be **empowered and trained** to create and maintain their own dataset records (e.g., at EGA/FEGA and data.europa.eu), rather than relying solely on central project teams.
- European initiatives should prioritise the **long-term sustainability and interoperability of end-to-end FAIR solutions** so that investments in FAIRification, like those made in HEREDITARY, remain usable beyond the lifetime of individual projects.

## 9. Next steps

- Extend Tier B and Tier C FAIRification to additional HEREDITARY datasets, using BaCo, HBS and the ALS cohorts as templates and prioritising those with clear governance and high reuse potential.
- Maintain existing Tier C FAIRified datasets, adding more details as they evolve.
- Hold WP3-wide discussions on FAIRification approaches to integrate HealthDCAT-AP dataset records and EGA/FEGA-ready metadata into WP3's multimodal semantic platform.
- Track and adopt emerging European standards and infrastructure. For example, the evolving FEGA metadata model, GDI requirements, and HealthData@EU tools, updating mappings and catalogue records as these mature.
- Embed the proposed FAIRification workflows and 1+MG Framework recommendations in HEREDITARY's Data Stewardship Wizard (DSW) instance,<sup>171</sup> so that future datasets are "born FAIR" and collected directly against standard schemas. This will be coordinated with T1.4 "Ethics management and data management plan".
- Produce further training materials for partners describing how to FAIRify their datasets. The format of these may vary, from internal documentation to a project's FAIR Cookbook<sup>172</sup> recipe.
- Work with WP8 "Exploitation, Innovation, Communication and Dissemination" to create a "Datasets" section on the HEREDITARY website's Open Hub,<sup>173</sup> listing project datasets with stable landing pages and URIs.

<sup>171</sup> <https://ega.dsw.elixir-europe.org/wizard/projects/1d2ecd58-f102-437b-95e2-46e3c107a6ac>

<sup>172</sup> <https://faircookbook.elixir-europe.org/content/search-wizard.html>

<sup>173</sup> <https://hereditary-project.eu/open-hub>



- Work with FEGA and GDI teams on developing the FEGA Metadata Model and implementing it towards a viable Tier A FAIRification alternative that incorporates 1+MG Framework's recommendations natively along EHDS evolving standards and services (e.g., automatic harvesting of dataset records by data.europa.eu).
- Support CNAG in finalising and operationalising the genomic data-quality protocol over HEREDITARY datasets, reusing its QC and harmonisation outputs in D3.7 and in T3.4's beaconisation activities.
- Apply ONTO's LLDI- and Airflow-based FAIRification pipelines to additional public and project-relevant resources, and document them as reusable workflows for future projects.
- Continue close collaboration with other HEREDITARY tasks (DMP revision in D1.2, future FAIRification reporting in D3.7, HDN development in D3.2, long-term sustainability plan with Task 8.3) to ensure that the project's data management planning, technical platforms and discovery services are aligned and that lessons from BaCo and Tier C activities are reflected in later project phases.

## 10. Milestone verification

**Milestone 7 "Data FAIRification"**, due at M24, is verified by D3.6 and not D3.10 as per the project amendment.<sup>174</sup> This deliverable provides evidence of milestone completion by documenting the design and first implementation (M7–M24) of practical FAIRification workflows for HEREDITARY's participating data resources, aligned with 1+MG/GDI and EU/EHDS approaches, and focused on measurable improvements in findability and metadata interoperability without changing access governance.

In particular, milestone 7 is verified by providing the following evidence of FAIRification work delivered by T3.6:

- **A concrete, partner-aware, and tiered FAIRification strategy.** This approach was implemented as follows: (1) Tier C FAIRification for three datasets (see [Section 6.2.2](#)); (2) ongoing Tier B FAIRification for a controlled-access cohort via submission to EGA (see [Section 6.3](#)).
- **A justified selection of discovery and metadata platforms.** It reports a structured landscape analysis (13 platforms assessed) to select an EU-aligned cataloguing route for HEREDITARY (see [Section 6.2.1](#)).
- **Objective validation of metadata quality and FAIR exposure.** For Tier C datasets, this includes DCAT-AP/HealthDCAT-AP compliance checks (validator/SHACL), metadata quality scoring (see [Figure 2](#)), and FAIR-Checker evaluation of the EHDS FDP exposure (see [Figure 3](#)).

---

<sup>174</sup> <https://hereditary.dei.unipd.it/groupoffice/index.php?r=files/file/download&id=1549>

- **FAIRification methods for scale-up.** D3.6 records reusable FAIRification components and next steps, including ongoing FEGA metadata model alignment work, a genomic data-quality protocol, and reusable FAIRification pipelines (e.g., LLDI-oriented workflows) to extend FAIRification beyond the initial implementation examples.



## References

Key	Reference
1	Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. <i>Scientific data</i> , 3(1), 1-9.
2	Gaignard, A., Rosnet, T., De Lamotte, F., Lefort, V., & Devignes, M. D. (2023). FAIR-Checker: supporting digital resource findability and reuse with Knowledge Graphs and Semantic Web standards. <i>Journal of Biomedical Semantics</i> , 14(1), 7.
3	Tolonen, H., Saso, M., Unim, B., Palmieri, L., Schutte, N., Peyroteo, M., ... & Bogaert, P. (2024). European Health Information Portal: a one-stop shop for health information. <i>European Journal of Public Health</i> , 34(Supplement 1), i29-i34.
4	Rueda, M., Leist, I. C., & Gut, I. G. (2024). Convert-Pheno: A software toolkit for the interconversion of standard data models for phenotypic data. <i>Journal of Biomedical Informatics</i> , 149, 104558.

## Annexes

Number	Name
1	Introduction to the benefits of participating in HEREDITARY Task 3.6
2	HEREDITARY Findability Platforms Landscape
3	Turtle dataset record of the Healthy Brain Study (HBS)
4	Turtle dataset record of the UNIPD's Longitudinal ALS Cohort
5	Data Access Policy of UNIPD's Longitudinal ALS Cohort
6	Data Dictionary of UNIPD's Longitudinal ALS Cohort
7	Turtle dataset record of UNITO's ALS Demographic and baseline data
8	Data Access Policy of UNITO's ALS Demographic and baseline data
9	Data Dictionary of UNITO's ALS Demographic and baseline data
10	Video Tutorial: HealthDCAT-AP Editor
11	BaCo data dictionary - Original fields
12	BaCo data dictionary - Term Mapping
13	BaCo data dictionary - Value Mapping