



M5.4 Draft guideline for data enrichment

TEHDAS2 – Second Joint Action Towards the European Health Data Space

21 April 2026

Co-funded by
the European Union



0 Document info

Disclaimer

Views and opinions expressed in this deliverable represent those of the author(s) only and do not necessarily reflect those of the European Union or HaDEA. Neither the European Union nor the granting authority can be held responsible for them.

0.1 Authors

Author(s)	Organisation
Anna Niemeyer	TMF - Technology and Methods Platform for Networked Medical Research, Germany
Beatriz Barros	Sciensano, Belgium
Fredrik Carlsson	Karolinska Institutet, Sweden
Hadjittofi Daniel	National e-Health Authority (NeHA), Cyprus
Menikou Andreas	NeHA, Cyprus
Michael Peolsson	The Swedish e-Health Agency, Sweden
Nienke Schutte	Sciensano, Belgium
Neophytou Marios	NeHA, Cyprus
Stylianides Antonis	NeHA, Cyprus
Aurēlija Usačova	Centre for Disease Prevention and Control, Latvia
Eva Zvirgzdiņa	Centre for Disease Prevention and Control, Latvia
Maria Ahlsén	Karolinska Institutet, Sweden

Acknowledgements

The authors would like to express their sincere gratitude to Matthias Löbe (IMISE, University of Leipzig, Germany), Ana Martin-Moreno (Ministry of Health, Spain) and Rosa Gini for their comprehensive and valuable contributions, as well as to them and the entire Review Board for their careful and thorough editorial review of this user manual. Particular thanks are likewise due to Zdenek Gütter (Ministry of Health, Czech Republic), Samer Schaat (gematik, Germany) and Marije van Melle (Nictiz, Netherlands), whose critical observations and constructive suggestions have substantially improved both the rigour and the clarity of the present work. The authors remain, of course, solely responsible for any errors or omissions that may persist.

0.2 Keywords

Keywords	TEHDAS2, Joint Action, Health Data, European Health Data Space, Data enrichment, HDAB, Health Data Access Body, Data user, Data holder
-----------------	--

0.3 Document history

Date	Version	Editor	Change	Status
18/03/2025	0.1	Anna Niemeyer	Initial document creation	Draft
14/10/2025	0.2	Anna Niemeyer and all contributors	First draft	Draft
01/12/2025	0.3	Anna Niemeyer, Beatriz Barros	Draft to be reviewed by the commission	Draft
28/01/2026	0.4	Anna Niemeyer, Beatriz Barros	Draft to be reviewed by the review board	Draft
14/03/2026	0.5	Anna Niemeyer, Beatriz Barros	Draft to be reviewed by the Consortium	Draft
17/04/2026	0.6	Anna Niemeyer, Beatriz Barros	Document to be submitted for PSG approval	Draft
22/04/2026	0.7	Anna Niemeyer	Document to be submitted for public consultation	Final

Accepted in Project Steering Group on 21 April 2026.

Copyright Notice

Copyright © 2024 TEHDAS2 Consortium Partners. All rights reserved. For more information on the project, please see www.tehdas.eu.

Contents

1 Executive summary.....	4
2 List of abbreviations.....	6
3 Introduction	7
3.1 Target Audience.....	7
3.2 Purpose	8
3.3 Scope and framing.....	9
3.3.1 Topics out of scope of this guideline	11
4 Data Enrichment in the EHDS User Journey	13
4.1 Data pre-processing by the data user	15
4.1.1 Correction of data	16
4.1.2 Data synchronisation and deduplication	17
4.1.3 Data standardisation	17
4.1.4 Data harmonisation	18
4.1.5 Data validation and correction	18
4.2 Zooming in on data enrichment.....	19
4.2.1 What is data enrichment?	19
4.2.2 Data enrichment operations	20
4.2.3 Data enrichment: Illustrative examples relevant to EHDS secondary use	21
5 Data enrichment journey.....	22
5.1 Considerations for data users.....	23
5.2 Considerations for HDABs regarding data enrichment.....	25
5.3 Considerations for data holders regarding enrichment outputs	30
6 Considerations for implementation	34
6.1 Legal framework for sharing enrichment outputs	34
6.2 Operational considerations	37
6.2.1 Possible ways to implement and to perform	38
6.2.2 Processing enriched dataset	38
6.3 Semantic interoperability	39
6.4 Organizational aspects	39
6.5 Practical examples.....	40
6.6 Pitfalls and challenges	41
6.6.1 Usual errors	41
6.6.2 Biases	42
7 References.....	43
8 Annexes	44
8.1 Methodology.....	45
8.2 User journey.....	45
8.3 Glossary	47
8.4 Relation to TEHDAS2 tasks and deliverables.....	52
8.5 Links to the EHDS regulation.....	52
8.6 Use case examples of data pre-processing and enrichment operations.....	53

1 Executive summary

This guide, produced by the Second Joint Action Towards the European Health Data Space (TEHDAS2), clarifies the concept of **data enrichment** within the framework of the European Health Data Space (EHDS). The EHDS aims to advance the secure and smooth use of health data for secondary purposes, such as research, innovation, and policymaking. Since the EHDS Regulation does not define specific legal obligations for enrichment, this guide provides illustrative considerations and practical examples to ensure harmonized procedures across EU Member States.

Within this evolving framework, data enrichment is understood as an optional, user-driven, and value-based activity that may occur during secondary use of health data. The EHDS Regulation does not establish data enrichment as a mandatory compliance requirement, nor does it prescribe uniform governance models for it. Instead, it recognises that enrichment activities may add analytical and scientific value in certain contexts, while leaving Member States discretion to decide whether, and how, such activities are addressed in national frameworks. This guideline therefore does not seek to impose obligations, but to illustrate possible approaches and promote a shared understanding across the EHDS ecosystem. Data enrichment is an optional, user-driven activity. Any practices described in this guide are non-binding and subject to national implementation choices.

What data enrichment is and why it matters

Data enrichment is a crucial step in data analysis, done by authorized users within a Secure Processing Environment (SPE). It adds new details or value to datasets, boosting their accuracy, completeness, or clarity. Data enrichment usually occurs in two ways:

1. **Internal enrichment:** Deriving new information exclusively from the dataset itself (e.g., calculating risk scores from existing clinical measurements or segmenting tumour regions in medical images).
2. **External enrichment:** Adding information from other datasets, to which access was granted upon a data permit and are, therefore, available in the SPE (e.g., appending area-level socioeconomic indicators or environmental exposure metrics to patient data).

Data enrichment differs from **data linkage**, a centralized process managed by the Health Data Access Body (HDAB) or data holder that connects records from various sources before users access the data. External enrichment does not include record-level linkage across datasets unless explicitly authorised as data linkage under applicable governance procedures.

The key roles and the data quality feedback loop

The enrichment process recommended in Rec. 57 can, with corresponding regulation by Member States, create an important feedback loop designed to continuously improve data quality for secondary use. Three main actors could take on complementary roles in this process:

- Data User:** Performs the enrichment within the SPE. If the enrichment is deemed valuable, they may return a description of the enrichment, or the enriched dataset where permitted and appropriate under applicable legal and contractual frameworks, along with Sharing this information with the data holder may should occur under agreed conditions, including where applicable free of charge, cost-recovery or licensing arrangements.
- HDAB:** Acts as the intermediary and permitting body. Where foreseen by national frameworks, the HDAB may facilitate or coordinate the assessment of data enrichment activities.
- Data Holder:** The initial holder of the data. They may engage with HDAB or the data user to make the enhanced dataset available for future use.

2 List of abbreviations

Abbreviation	Description
ADR	Adverse Drug Reaction
AI	Artificial Intelligence
AHIMA	American Health Information Management Association
API	Application Programming Interface
Art	Article
ATC	Anatomical Therapeutic Chemical
CDM	common data model
COVID-19	Coronavirus disease 2019
CPT	Current Procedural Terminology
CT	Computer Tomography
DICOM	Digital Imaging and Communications in Medicine
DPA	Data Processing Agreement
DRG	Diagnosis Related Groups
DSA	Data Sharing Agreement
DUA	Data Use Agreement
EEHRxF	European Electronic Health Record Exchange Format
EHD	Electronic Health Data
EHDS / EHDSR	European Health Data Space / European Health Data Space Regulation
eHDSI	eHealth Digital Service Infrastructure
EMA	European Medicines Agency
EU	European Union
ETL	Extract, Transform, and Load
GDPR	General Data Protection Regulation
HDAB	Health Data Access Body
HDL	high-density lipoprotein
HL7 FHIR R5	Health Level Seven, Fast Healthcare Interoperability Resources standard, Release 5
IDMP	Identification of medicinal products
ICD	International Classification of Diseases
ISO/TC	International Organization for Standardization /Technical committees
JCA	Joint Controllership Agreement
LDL	low-density lipoprotein
LLM	Large Language Model
LOINC	Logical Observation Identifiers Names and Codes
MedDRA	Medical Dictionary for Regulatory Activities
OMOP	Observational Medical Outcomes Partnership
Rec	Recital
RWD	Real World Data
RWE	real world evidence
SNOMED CT	Systematized Nomenclature of Medicine – Clinical Terms
SPE	Secure Processing Environment

Abbreviation	Description
TEHDAS2	Joint Action Towards a European Health Data Space 2

3 Introduction

Advancing health data use in the European Health Union

As part of the European Health Union, the European Union (EU) is advancing the use of health data for secondary purposes, including research, innovation and policymaking. Smooth and secure access to data will drive the development of new treatments and medicines and optimise resource utilisation - all with the overarching goal of improving the health of citizens across Europe.

TEHDAS2, the second joint action Towards the European Health Data Space, represents a significant step forward in this vision. The project will develop guidelines and technical specifications to facilitate smooth cross-border use of health data, and support data holders, data users and the new health data access bodies in fulfilling their responsibilities and obligations outlined in the European Health Data Space (EHDS) regulation.

TEHDAS2 focuses on several critical aspects of health data use.

1. **Data discovery:** findability and availability of health data, ensuring it is accessible for secondary purposes.
2. **Data access:** developing harmonised access procedures and establishing standardised approaches for granting data access across Member States.
3. **Secure processing environment:** defining technical specifications for environments where sensitive health data can be processed safely.
4. **Citizen-centric obligations:** providing guidance on fulfilling obligations to citizens, such as communicating significant research findings that impact their health, informing them about research outcomes and ensuring transparency in how their data is used.
5. **Collaboration models:** developing guidance on collaboration and guidelines on fees and penalties as well as third country and international access to data.

TEHDAS2 will contribute to harmonised implementation of the EHDS regulation through the concrete guidelines and technical specifications. Some of these documents and resources will also provide input to implementing acts of the regulation. Hence, the joint action will increase the preparedness for the EHDS implementation and lead to better coordination of member states' joint efforts towards the secondary use of health data, while also reducing fragmentation in policies and practices related to secondary use.

3.1 Target Audience

The target audience of this guideline includes health data access bodies, data holders and health data users who may be involved in, or affected by, data enrichment activities in the context of secondary use under the EHDS framework.

- **HDABs** are the central authorities managing access to electronic health data for secondary use under the EHDS Regulation. They receive and assess data access applications, issue data permits where applicable, and oversee that access and processing comply with permit conditions, including the use of authorised Secure Processing Environments (SPEs) and output controls. Regarding data enrichment specifically, the EHDS Regulation does not assign HDABs responsibilities in this area. However, in some Member States and where national rules allow, HDABs could play an illustrative role as intermediaries for communicating enrichment-related information between data users and data holders, one possible governance model among others that Member States might consider.
- **Data holders** are responsible for making relevant electronic health data available to HDABs upon request, in accordance with a data permit. The Regulation does not impose any binding obligations on data holders to integrate enriched datasets received from secondary users. Any involvement of data holders in assessing, integrating, or reusing enriched data is therefore optional and subject to national rules and institutional policies. This guideline therefore provides illustrative considerations for data holders that may choose to engage with enrichment outputs communicated via HDABs or other mechanisms, including assessment methodological soundness, traceability, alignment with standards, and potential utility of the enrichment for future secondary use. This guideline provides guidance on how data holders could approach such evaluations in a structured and consistent manner, supporting dataset quality, interoperability, and reusability within the EHDS ecosystem.
- **Data users** are granted access to approved datasets for secondary use under a data permit and are responsible for processing the data within a SPE in compliance with legal, ethical, and technical requirements. While the legal framework does not provide binding rules for enrichment, data users play a central role in performing data enrichment as part of their analytical role, which may include deriving new variables, adding annotations, or integrating additional authorised information to enhance the analytical value of a dataset. This guideline is intended to support data users by clarifying what constitutes data enrichment within the EHDS framework for secondary use and providing guidance on concepts, benefits, and good practices for documentation and communication in alignment with EHDS principles. Taken together, these interactions may, where enrichment is performed and shared, contribute to feedback loops that support gradual improvements in data quality and reuse for secondary use within the European Health Data Space.

3.2 Purpose

Data enrichment is a relevant step of the data analysis lifecycle. It encompasses a range of operations through which data users may enhance the accuracy, completeness, or interpretability of datasets during their analytical work. While enrichment can be valuable in certain contexts, it is not a universal requirement, as many secondary-use workflows proceed without enrichment.

Although the EHDS Regulation does not define specific legal obligations related to data enrichment, it leaves Member States the competence to decide whether and how such activities are addressed within their national secondary use frameworks (Article 51(3)). As such, this guideline is non-binding and exploratory in nature. It does not establish legal obligations, mandatory workflows, or compliance requirements for any actor under the EHDS Regulation. Instead, it presents illustrative examples, considerations, and good practices related to data enrichment, recognising that Member States may adopt different approaches or choose not to regulate data enrichment at all. The purpose of this document is to explain when and why data enrichment can be valuable and how it can work in practice, for those cases where it is relevant. It aims to illustrate the potential benefits of enrichment for data quality and reuse in certain analytical contexts. Using examples, it describes typical enrichment processes and their associated risks (bias, re-identification) and presents good practices. The illustrative considerations and practical examples made in this guideline should not be interpreted as binding governance or procedural requirements, but rather as suggestions for the future design of the regulatory framework by Member States.

Considering this, the goal of this guideline is to support a common understanding across the EU of what constitutes data enrichment and how it can be appropriately managed within the EHDS framework. It discusses:

- What defines data enrichment and the types of operations it may include
- Where data enrichment can occur along the EHDS secondary use process and by whom
- Illustrative considerations and practical examples for Member States on integrating data enrichment into governance and technical frameworks
- Challenges and open questions that may require future consideration.

By promoting a shared understanding of data enrichment, this document seeks to help avoid fragmentation and ensure consistency in national approaches, supporting a coherent and interoperable EHDS ecosystem.

3.3 Scope and framing

This guideline is part of a series developed under the TEHDAS2 project to support the operationalisation of the EHDS Regulation, with a specific focus on Chapter IV concerning the secondary use of health data. The approaches described in this guideline should be read as illustrative options rather than as a prescriptive implementation blueprint.

As outlined in the previous section, data enrichment is one possible step in the data processing pipeline for researchers and authorised data users. In the EHDS context, the concept is introduced in Recital 57 of Regulation (EU) 2025/327 (see box below), which envisions the possibility for health data users to enhance the datasets they access through various corrections, annotations, or other improvements, for example, by supplementing missing or incomplete data to increase their accuracy, completeness, or quality. While Recital

57 recognises the value of data enrichment and suggests some general principles, the Regulation itself does not define specific legal obligations regarding this process. Consequently, it falls within Member State competence to establish the legal basis, procedures, and governance mechanisms for data enrichment within their national frameworks for secondary use, as indicated in Article 51(3) of the EHDS Regulation. Given this decentralised responsibility, this guideline provides illustrative considerations and practical examples for HDABs, data holders, and data users on how to approach data enrichment in a way that supports coherent implementation and interoperability across the EHDS ecosystem.

EHDS Regulation – Recital 57

“Health data users who benefit from access to datasets provided for under this Regulation could enrich the data in those datasets with various corrections, annotations and other improvements, for instance by supplementing missing or incomplete data, thus improving the accuracy, completeness or quality of the data in the datasets.

(...) To support the improvement of the initial database and further use of the enriched dataset, Member States should be able to establish rules for the processing and the use of electronic health data containing improvements related to the processing of those data. The improved dataset should be made available free of charge to the original health data holder together with a description of the improvements. The health data holder should make the new dataset available, unless it provides a justified notification to the health data access body for not doing so, for instance in cases in which the enrichment by the health data user is of low quality.”

Importantly, although Recital 57 refers to enhanced datasets, please note that the practical considerations in this document often favour sharing the methods, documentation, code, or analytical approaches underlying enrichment activities rather than transferring enriched datasets themselves. This reflects practical realities of the SPE model and Member State’s flexibility in determining how enrichment outputs are communicated and governed. Different Member States may adopt different approaches, some may facilitate dataset sharing, others may prioritise documentation or code sharing, and some may not formally govern enrichment at all. Sub-task 5.3.1 contributes operational guidance by bridging the legal vision with hands-on analytical practice, illustrating what data enrichment may look like in different contexts and how it could be documented and communicated. The scope of this deliverable focuses on illustrative considerations related to:

- **Who:** the data user authorised to process data for secondary use;
- **What:** data enrichment operations carried out to improve the dataset during analysis;
- **When:** during data processing within a SPE, under a valid data permit.

Building on this framing, the document explores the definition and examples of data enrichment operations (Chapter 4), examines the roles and responsibilities of actors involved in the EHDS framework (Chapter 5), and concludes with illustrative considerations for integrating data enrichment into governance and technical frameworks, including future challenges and open questions (Chapter 6).

Readers should note that the existence, form, and governance of data enrichment practices may vary significantly between Member States, depending on national legal choices, technical infrastructures, and institutional arrangements. Data users are therefore advised to familiarise themselves with the relevant regulations for data enrichment in the country from which the dataset is accessed or provided to them. In case such rules have not yet been formulated, it is advisable to enquire with the respective HDAB about the possibilities for data enrichment and the associated conditions.

Beyond its operational dimension, data enrichment also raises broader questions related to scientific practice, openness, and the reuse of research outputs. In many areas of data science, the ability to reuse enriched datasets, intermediate analytical outputs, or derived variables is considered an important component of reproducibility, transparency, and cumulative knowledge generation. Within the current EHDS framework, however, the emphasis is placed primarily on controlled access to source datasets and the production of analytical results, while the reuse of enriched datasets themselves remains limited. While it is beyond the scope and remit of this guideline to address these systemic aspects, they are of clear relevance and should be carefully considered in ongoing and future discussions on the evolution of secondary use practices in Europe.

3.3.1 Topics out of scope of this guideline

To create a target guide focus on data enrichment, the following topics are out of scope of this work and will not be considered in this document:

1. Data Linkage

Data linkage refers to the process of connecting or matching records that relate to the same individual or entity across different datasets. This linkage is typically achieved through unique identifiers or matching algorithms, with the purpose of creating a single, richer combined dataset. In contrast, data enrichment involves adding new information, such as additional attributes or derived variables to an existing dataset, usually by merging it with external or contextual sources, thereby enhancing its analytical depth.

Example of data linkage: A research team submits an application to a HDAB requesting to link electronic health records (category a) with a national mortality register (category I). The HDAB or the relevant data holder performs the linkage and provides the researcher with the resulting, already linked dataset that unifies care trajectories and outcomes. The researcher does not receive the separate original datasets or access to the linkage keys.

Example of data enrichment: After receiving an authorised dataset, a data user working within a SPE supplements it with additional socio-economic indicators (category b), such as

area-level deprivation scores, sourced from publicly available data. This step enhances the dataset’s contextual value but is performed solely within the user’s analysis environment, under the terms of the data access permit.

Distinction in the EHDS workflow

- Within the EHDS framework, the key distinction between linkage and enrichment lies in when these processes occur and who is responsible for them. Data linkage is a **pre-access**, centralised operation. It must be explicitly requested during the data access application and is carried out by the HDAB or the designated data holder under regulated conditions. The technical and governance checks, such as ensuring data minimisation, matching accuracy, and legal basis, are conducted before any dataset is made available to the user. Only the final, linked dataset is provided to the applicant. Data enrichment, by contrast, is a **post-access**, analytical operation performed by the authorised data user within the SPE. It occurs after the approved dataset has been provided and involves user-driven steps to enhance or contextualise data as part of the permitted research activities.

In summary, data linkage is a formal, pre-access process managed by HDABs or data holders and subject to regulatory assessment (as defined in Article 68(1)(b) of the EHDS Regulation), whereas data enrichment is a user-led analytical process conducted during data processing under an existing data permit. For this reason, data linkage falls outside the scope of this guideline. Its operational and governance aspects will be covered in a separate deliverable, the “Guideline for Health Data Access Bodies on Linkage of Health Datasets” (forthcoming, consultation wave 3, May 2026).

2. Critical errors in a dataset

Another topic outside the scope of this guideline concerns the identification and correction of critical errors in datasets. Critical errors are substantial inaccuracies or inconsistencies within a dataset that prevent the data user from carrying out the intended data processing under an approved data permit. These errors differ fundamentally from enrichment activities: they represent foundational data integrity issues that must be addressed before any analytical processing can take place, and they typically cannot be corrected independently by the data user as they often require verification against original data sources or technical adjustments within the data infrastructure. When a data user encounters a potential critical error, the appropriate course of action is to report the issue to the HDAB. Management and correction of critical errors involve dedicated quality assurance processes governed by HDABs and data holders in line with the broader EHDS framework and therefore fall outside the scope of this guideline.

Table 1. Examples of critical errors and their resolution pathways

Type of critical error	Description	Resolution pathway
Linkage failure	Missing or incorrect pseudonymisation codes prevent	Re-execute or validate linkage procedure; issue corrected dataset

	records from different datasets from being matched correctly.	
Logical inconsistency	Dataset contains contradictions (e.g. diagnosis date before birth date, male patient with pregnancy record).	Review source systems; correct or flag inconsistent records before re-release
Systematic coding error	Misclassification or mislabelling of diagnosis or treatment codes due to mapping or system failure.	Validate coding logic; correct mappings and regenerate dataset
Missing core variables	Key variables listed in the data permit are absent, preventing analysis.	Retrieve missing data from source or provide documented justification and updated dataset
Corrupted or incomplete data files	Technical errors during extraction or transfer render the dataset unreadable or incomplete.	Re-extract data, verify file integrity, and provide corrected version

3. Reporting of significant findings

As established in TEHDAS2 Milestone 8.2 document “Guideline to Health Data Access Bodies on implementing the obligation of notifying the natural person on a significant finding from the secondary use of health data”, clinically significant findings are defined as observations arising from data analysis that have a direct and meaningful impact on patient care, diagnosis, treatment, or prognosis. Such findings are relevant only when they carry tangible clinical implications that could influence medical decision-making or therapeutic options. Not all research outcomes meet this threshold, as many analytical results remain of scientific or statistical interest without bearing direct clinical consequence.

Within the EHDS framework, the identification and reporting of significant findings take place after the processing and analysis of health data, and therefore beyond the data enrichment phase covered in this document. The communication of significant findings, including whether, when, and how patients or healthcare professionals should be informed, is governed by dedicated legal and ethical procedures. For detailed rules and operational guidance on this subject, readers should refer to the forthcoming TEHDAS2 Milestone 8.2 document “Guideline to Health Data Access Bodies on Implementing the Obligation of Notifying the Natural Person on a Significant Finding from the Secondary Use of Health Data”, which provides dedicated operational guidance for reporting and notification processes.

Accordingly, this guideline does not address any requirements or procedures related to the detection, reporting, or management of significant findings and the overall topic of significant findings falls outside the scope of this guideline.

4 Data Enrichment in the EHDS User Journey

In the EHDS Regulation, the secondary use of health data follows a structured workflow designed to ensure secure, lawful, and efficient access while maintaining data quality and integrity. Recital 57 recognises that, during analysis in an SPE, data users may improve datasets through corrections, annotations or other enhancements and that Member States may establish arrangements for sharing such improvements. The EHDS Regulation does not

impose a mandatory enrichment workflow. Therefore, this chapter describes typical stages of secondary use and illustrates where enrichment can occur, while recognising that practical procedures may differ across Member States and HDAB implementations.

In general, a data user's journey can be summarised in four main stages:

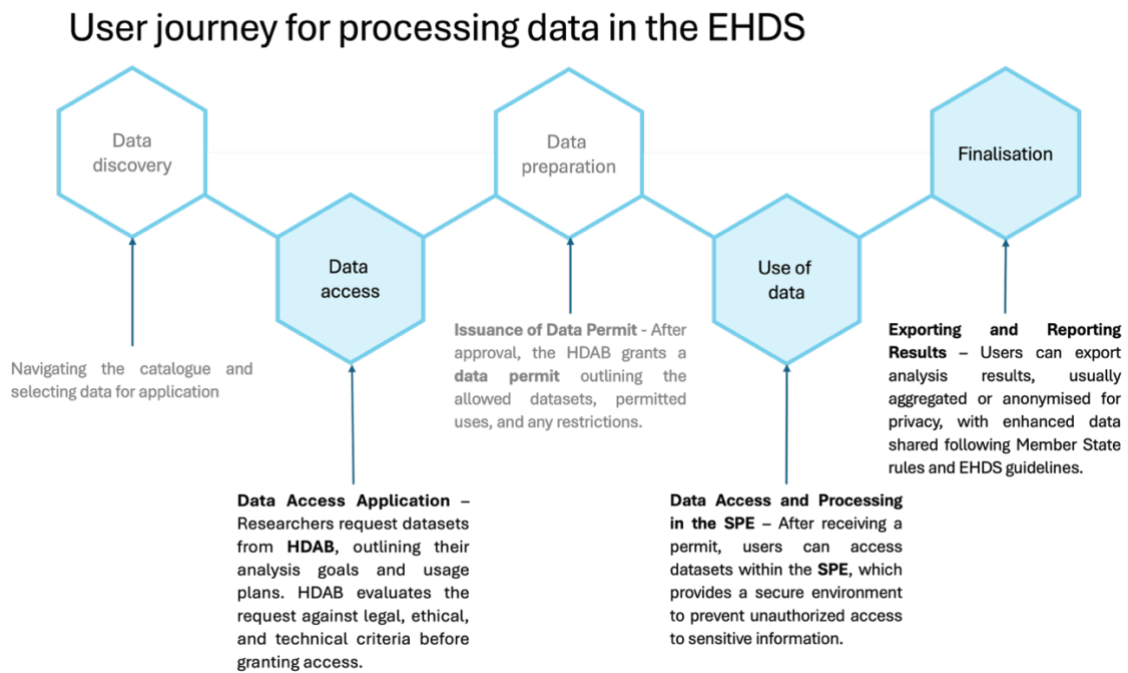
1. **Application for Data Access** – Applicant (prospective data user) submits an application to the relevant HDAB. The application specifies the datasets required, the purpose of the analysis, and any planned data operations, including intentions to link or enrich data. The HDAB evaluates the request against legal, ethical, and technical criteria before granting access.
2. **Issuance of Data Permit** - Once approved, the HDAB issues a data permit that defines the scope, conditions, and limitations of data use. The permit specifies which datasets the user may access, the authorised operations, and any restrictions applying.
3. **Data Access and Processing in the SPE** - Following permit issuance, the data user gains access to the requested datasets within a SPE. The SPE ensures that all processing occurs in a controlled and secure environment, preventing unauthorised access to sensitive data.
4. **Exporting and Reporting Results** - After analysis, the data user may export results, subject to the conditions of the data permit. Outputs are typically aggregated, anonymised, or otherwise processed to ensure compliance with privacy and security requirements. Documentation of enrichment methods and, where legally and technically feasible, related outputs may be communicated to the HDAB and, where relevant under national rules, to the original data holder, in line with Recital 57 of the EHDS Regulation.

Within this EHDS workflow, different types of data operations occur at distinct stages and involve different actors. After the issuance of a data permit, certain foundational operations are carried out by the HDAB and the data holder to ensure that the data are accessible and compliant with regulatory requirements. These operations include preparing, compiling, pseudonymising, anonymising, or minimising datasets as needed. They may also include formal data linkage when explicitly requested and approved. Importantly, the Regulation does not require the HDAB or data holder to tailor the dataset to the specific analytical preferences of the data user beyond what is needed to fulfil the permit and legal safeguards. Such preprocessing, including cleaning, standardisation, or other preparation tasks, is typically specific to the planned analysis and therefore falls within the responsibility of the data user. Once the data user accesses the datasets in the SPE, they may perform a range of operations, including project-specific data preparation, cleaning, transformation, and, where relevant, data enrichment. Enrichment consists of adding value, context, or completeness to the datasets, for example by supplementing missing information, integrating permitted external data, or annotating variables to improve interpretability and analytical utility.

The scope of this chapter is focused specifically on the operations performed by the data user within the SPE, with particular emphasis on data enrichment. Subsequent sections will provide definitions, classify types of enrichment procedures, and illustrate examples of

enrichment activities. Later chapters will then address how enrichment-related information, including methods, documentation, code, and where legally and technically feasible, related outputs, may be communicated to the HDAB and data holder, and how such information may be integrated into the EHDS framework for future use.

Figure 1: Journey of a data user



Importantly, while Recital 57 envisions the possibility of enriched datasets being made available for future use, the practical realities of data holder infrastructure often constrain how enrichment feedback is operationalised. Many data holders, particularly those managing data from electronic health record systems or other operational infrastructures, face significant technical and organisational constraints that make integration of enriched datasets into their systems difficult or infeasible. These constraints may include legacy system architectures, operational dependencies, data model rigidity, or insufficient IT resources. For this reason, enrichment feedback mechanisms that prioritise sharing methods, documentation, code, and executable procedures rather than enriched datasets themselves reflect practical realities and institutional capacities. Data holders can assess and potentially benefit from methodological contributions without needing to integrate new data into operational systems.

4.1 Data pre-processing by the data user

As introduced in the previous section, once a data user gains access to the requested datasets in the SPE, a variety of operations can be performed on the data. For clarity and practical alignment with EHDS workflows, it is useful to distinguish between three complementary types of operations:

- **Data preparation** - making existing data usable and consistent for analysis (e.g. cleaning, formatting, harmonising)
- **Data enrichment** - adding new derived or contextual information to enhance the dataset
- **Data linkage** - person-level matching or joining across datasets, which remains outside the scope of this guideline unless separately authorised under the data permit

While both data preparation and enrichment occur within the SPE, they serve distinct purposes. This section focuses on data preparation as foundational to enrichment; the detailed discussion of data enrichment, including its definition, types of operations, and examples, is provided in Section 4.3.

Data preparation encompasses operations that ensure the dataset is fit for analysis. This typically includes cleaning (addressing missing values, outliers, inconsistencies), formatting (standardising variable structures), harmonising (aligning coding schemes or units across sources), and validating (checking data integrity before proceeding to analytical steps). Because each research project has unique analytical objectives, project-specific data preparation is normally a first step in the analysis workflow. This ensures the dataset is properly structured for the intended analytical operations, providing a reliable foundation for subsequent enrichment activities where relevant.

4.1.1 Correction of data

Purpose: The primary goal of data correction is to ensure that health data is accurate, consistent and aligned with established standards.

Description: Data correction refers to the process of identifying and rectifying errors, inconsistencies or inaccuracies. The aim is to maintain precision and consistency, enabling better patient outcomes and adherence to regulations [1].

Process: This process includes correcting incorrect data entries, addressing missing or incomplete data and standardizing values. By improving the accuracy, completeness, and internal consistency of datasets, data correction enhances their usability for research, public health initiatives, and policymaking [2], and supports alignment with applicable regulatory and quality requirements.

EHDS examples:

- An example of data correction is standardizing date formats like “DD/MM/YYYY” in a dataset into a uniform format to avoid misinterpretation.
- The measured height of the patient, based on several entries, averages approximately 1.75 m. One entry reported a height of 1.85 m, which appears to be a data entry error. This outlier can be replaced with the mean value calculated from the remaining valid measurements for the same patient.

4.1.2 Data synchronisation and deduplication

Purpose: Data synchronisation and deduplication is used to ensure data consistency two or more datasets to improve the precision of the statistical results by eliminating redundancies. [3, 4]

Description: Process of identifying and eliminating duplicate data records within two or more datasets to retain a single record of the information to ensure data consistency and avoid redundancies.

Process: The process entails detecting duplicate records within two or more datasets, for example, based on predefined criteria, such as unique identifiers for each record, eliminating duplicates and retaining a single record of the information.

EHDS examples:

- Consolidated datasets from multiple clinics contain duplicate records of the same disease for a single patient. The redundant records for the patient are eliminated to retain a single entry for the diagnosis.

Note: Data deduplication may also be a standard activity of data holder during data preparation, consolidation (see M6.1 Draft guideline for health data holders on making personal and non-personal electronic health data available for reuse for further information). However, there still may be duplicated data in datasets that were only revealed by the data user.

4.1.3 Data standardisation

Purpose: Data standardisation ensures that several datasets can be integrated into a single, unified dataset, allowing researchers to analyse data from various sources seamlessly [7] and enables the unified interpretation and processing of data.

Difference with harmonisation: while harmonisation seeks to align and reduce differences among existing standards, standardisation aims to create a single, uniform standard that eliminates variations [8]

Description: Converting data into a uniform format using an agreed convention. This can include the standardisation of units of measurement, date formats or address formatting (e.g. converting "St." to "Street").

Process: The process includes application of predefined standards and rules – like converting units of measurement to fit a certain format, converting date formats and categorical values to have the same format across the whole dataset.

EHDS examples:

- Researchers merge datasets from several healthcare institutions that use different disease coding classifications (ICD-10, SNOMED CT etc.) using resources such as the

OMOP CDM to facilitate these mappings and merging. Researchers must determine which classification they'll use and then update the whole dataset accordingly.

- Dataset might record "gender" as a binary value (male/female), while another dataset might use a more inclusive set of values (male/female/other).
- Two laboratories use different units that need to be converted. For example, to convert total cholesterol, LDL and HDL from mg/dl to mmol/l.

4.1.4 Data harmonisation

Purpose: Harmonisation and standardisation both aim for data uniformity. Standardisation focuses on conformity, while harmonisation emphasizes consistency.

Description: Data harmonisation is the process of standardising and aligning different data sets from different sources to ensure a uniform and consistent presentation [9]. By harmonising data, data quality can be improved, and reliable analyses can be enabled. It is therefore particularly important in areas, where accurate and comparable data is crucial.

Process: Data harmonisation can often be an iterative process, that involves collecting datasets from various sources, assessing their quality and structure, and defining variables to be harmonised. The process includes mapping or merging data elements to align with a common framework across all data sources. The harmonised dataset is then validated and requires ongoing maintenance as new data emerge [9].

4.1.5 Data validation and correction

Purpose: The primary purpose of data validation is to ensure the accuracy, consistency, and quality of data before it is processed and used in various applications. It aims to identify and prevent errors, inconsistencies, or missing values, thereby improving the reliability of data-driven decisions.

Description: Data validation is the process of systematically verifying data against established or predefined rules, standards, data models and expectations to determine its validity, integrity, and compliance with specific requirements. These criteria typically include data type correctness (e.g., numeric fields contain only numbers), value range accuracy (e.g., age values within realistic limits), format consistency (e.g., dates in a standard format), logical coherence between data elements, conformance to a schema and completeness of required information [10].

Process: When analysing the data, the user defines validation rules based on data standards and expected patterns. These rules are applied to specific variables, involving checks for missing values, data type verification, and ensuring values fall within expected ranges. If errors are found, the user applies the necessary corrections [11].

EHDS examples:

- Ensuring that measurements are within physiologically possible ranges.

- Validating that ICD-10 codes are valid and current.
- Verifying that dates of death are not in the future and occur after the patient's date of birth.
- Ensuring that billing codes in insurance claims align with the diagnosis codes in medical records (e.g., ICD-10 codes matching the procedure performed).
- Validation of medication data according to the corresponding FHIR profiles from the EEHRxF.

4.2 Zooming in on data enrichment

4.2.1 What is data enrichment?

Data enrichment refers to the process of enhancing an existing dataset by adding new information, context, or derived value to increase its analytical usefulness. Enrichment goes beyond preparing the data for analysis (e.g., cleaning or structuring) and instead aims to expand the informational content of the dataset so that it can support deeper or more refined insights.

For the purposes of this guideline, the distinction between internal and external enrichment is used as an illustrative aid rather than a strict technical or legal categorisation. In practice, enrichment activities exist along a continuum of data combination practices, differing in terms of data origin, granularity, identifiers used, and potential re-identification risk. Some forms of external enrichment may rely on joins or mappings that are technically similar to linkage operations. Importantly, where external enrichment involves record-level matching or joining across datasets to create person-level connections, it should be treated as data linkage for governance purposes unless national rules explicitly provide otherwise. The distinction introduced below is therefore intended to support conceptual clarity and practical discussion, while recognising that governance implications may still need to be assessed on a case-by-case basis, depending on the characteristics of the enrichment performed and applicable national rules.

- **Internal enrichment:** This involves deriving new information directly from the dataset itself. Examples include creating new variables, classifications, indices, or temporal indicators that increase the analytical value of the data. Internal enrichment does not introduce information from external sources, but transforms and augments the dataset through computation, summarisation, or segmentation.
- **External enrichment:** In the context of the EHDS framework, this would involve appending additional information from other authorised datasets available to the user in the SPE. Such enrichment typically adds supplementary attributes or contextual elements (for example, area-level socioeconomic indicators or environmental exposure values) without creating a fully merged individual-level dataset. While these operations may technically involve joins or mappings, they are characterised here by their lower granularity and contextual nature, rather than by the absence of linkage mechanisms.

Why is data enrichment important?

Within the EHDS, health data collected primarily for direct patient care, administrative, or regulatory purposes (primary use) are made available for reuse (secondary use). These datasets are subject to strict data minimisation principles during access authorisation, meaning only the necessary data elements strictly required for data request purposes and permitted by the HDAB are provided to the data user. Consequently, the dataset included in the data permit should contain all variables present in the source data and necessary for the approved research purpose. However, it may still lack derived variables, analytical constructs, or contextual information that are not part of the original dataset but would be needed to further tailor it to the specific needs of the data user's research question. It is therefore unlikely that the dataset, as originally collected for primary purposes, is a perfect fit for a new analytical purpose without further user-driven processing. Data enrichment can be an important step at this stage to supplement and enhance the available data, enabling researchers and analysts to fill gaps, add relevant contextual factors, and improve data completeness and quality within the constraints of the access permit.

By enriching the data in a SPE, users can refine their datasets by appending additional information from approved internal or external sources, harmonising values, and deriving new variables that augment the analytical potential. However, enrichment activities can alter the data's risk profile and must be carried out in a manner consistent with data protection requirements and the conditions of the access permit. Thus, data enrichment is about thoughtfully expanding the dataset's information content in a manner that supports valid and efficient analysis, while remaining mindful of the safeguards and constraints that govern secondary use.

4.2.2 Data enrichment operations

As introduced in the previous section, data enrichment operations consist of deliberate steps taken by the data user to increase the analytical value, completeness, and context of a dataset once it is accessed in the SPE. Most enrichment is project-specific; only some enrichment outputs may be suitable for wider reuse, subject to quality assessment and national arrangements.

Internal enrichment operations

Internal enrichment refers to operations that derive new information exclusively from the dataset itself, without incorporating external sources. Typically, operations may include but are not limited to:

- **Feature derivation:** Calculating new variables or indicators based on existing data fields, such as risk scores computed from combinations of clinical measurements.
- **Temporal aggregation:** Summarizing time-series data to capture trends or episode-level summaries (e.g., summarizing hospital visits into disease progression phases)
- **Semantic tagging:** Applying standardized clinical codes or ontologies internally for more precise data description and searchability.

- **Temporal and Longitudinal Transformations:** Generating temporal variables or event sequences that allow for trajectory or time-to-event analyses (e.g. time from diagnosis to treatment, follow-up periods, sequence of interventions).

Internal enrichment leverages the existing data to create new perspectives or measures while remaining fully within the bounds of the dataset provided in the SPE.

External enrichment operations

External enrichment involves adding authorised information from other datasets available to the user in the SPE. Typically, operations may include but are not limited to:

- **Contextual augmentation:** Adding epidemiological, environmental, or socio-economic variables available in other datasets to enrich the understanding of patient or population cohorts.
- **Supplementing missing or sparse information:** Filling gaps in datasets by appending approved variables from other sources that are not directly linked but provide useful context such as adding population-level prevalence data, reference ranges, or aggregated health metrics relevant to the cohort.
- **Aggregated data appending:** Using population-level statistics or summary metrics matched by geographic or demographic identifiers to enhance dataset context.
- **Annotation:** Adding descriptive information, coding references, or clinical annotations that improve interpretability. Another example is creating a novel segmentation operation on a dataset of CT scans.

4.2.3 Data enrichment: Illustrative examples relevant to EHDS secondary use

Use case 1: Derivation of Composite Clinical Scores from EHRs

Data Categories Involved: (a) EHR data, (e) administrative data

Operation Type: Internal enrichment

Description: Using electronic health records (EHRs), a researcher calculates a patient risk score (e.g., a frailty index) by combining existing fields (diagnoses, medication use, lab results) into a single summary metric applicable for cohort stratification. The enrichment is internal, based solely on the variables present in the authorised EHR, generating new actionable information for population health analytics.

Use case 2: Annotation and Segmentation of Medical Images

Data Categories Involved: (a) EHR data, (h) device-generated health data, (m) clinical studies

Operation Type: Internal enrichment

Description: A researcher working with CT scan images obtained during routine clinical care uses advanced software to segment and annotate tumour regions. New labels indicating tumour boundaries and tissue types are derived and added to the dataset without sourcing new external data. This process enriches the imaging dataset, facilitating downstream analysis for AI model training or statistical evaluation, as shown in recent studies on biomedical image annotation and segmentation.

Use case 3: Augmenting Patient Records with Socio-Economic and Environmental Indicators

Data Categories Involved: (a) EHR data, (b) socio-economic and environmental determinants, (k) public health registries, (c) aggregated resource data

Operation Type: External enrichment

Description: A study enriches individual patient records by adding neighbourhood-level deprivation scores, pollution exposure metrics, or regional health indices derived from publicly available external datasets in the SPE. These indicators are appended using non-identifiable grouping (e.g., by postal code) rather than direct linkage, as highlighted in EHDS frameworks and RWE research. Such augmentation remains subject to the authorised research purpose, applicable national rules, and must be evaluated for re-identification risk, particularly where the granularity of indicators is high or cohort sizes are small.

Use case 4: Integrating Omics Profiles and Device Data in Population Cohorts

Data Categories Involved: (f) genomic data, (g) omics data, (h) device data, (p) research cohorts

Operation Type: External enrichment

Description: A researcher studying diabetes uses genomics and metabolomics datasets, together with wearable device activity tracking, to enrich a clinical registry cohort. Each dataset remains separate, but summary characteristics (e.g., mutation counts, average activity levels) are appended to the main cohort dataset for enhanced multi-modal analysis. This represents a high-risk edge case involving multiple sensitive data streams and complex derivation choices. Careful consideration must be given to how these summaries are constructed, as choices in aggregation or derivation can significantly influence both re-identification risk and the validity of downstream analyses. Only aggregated or non-identifying attributes should be appended where this reduces re-identification risk. Such enrichment may require separate authorisation and case-by-case governance assessment by the HDAB, depending on the specifics of the data combination, the sensitivity of the omics and device data involved, and applicable national rules.

5 Data enrichment journey

Unlike other operations within the secondary use of health data, data enrichment is not legally defined in the EHDS Regulation in terms of a structured framework, detailed roles, or formal responsibilities.

Recital 57 recognises the potential for users to enrich datasets and states that Member States may establish national rules or procedures for processing and using enriched data, leaving operationalisation largely to national discretion. Considering this, this chapter explores possible ways that Member States could organise voluntary feedback on data enrichment activities within the EHDS secondary use context. It builds on the previous chapter's definitions and examples of enrichment operations and considers practical implications for enrichment activities, while maintaining consistency with the SPE model, permit conditions, and applicable safeguards. This chapter does not prescribe workflows or define binding obligations for data users, HDABs, or data holders.

To structure the discussion, this work considers possible roles and interactions across three main actor domains (recognising that Member States may choose different approaches):

1. **Data user:** Performs enrichment within the SPE. Where national rules allow, they may share enrichment-related information (methods, documentation, code, or outputs) with the HDAB or data holder.
2. **HDAB:** Acts as the intermediary responsible for evaluating access requests, issuing data permits, and facilitating secure access to datasets. Where established by national rules, HDABs may facilitate communication about enrichment activities, though this is not a regulatory requirement.
3. **Data holder:** The original custodian of the datasets. Where national rules provide for it, they may receive enrichment-related information from users or the HDAB and determine how such information may be integrated or stored.

By exploring potential interactions among these actors, this chapter illustrates possible approaches, not requirements, for organising enrichment activities and feedback. The objective is to provide practical examples and considerations, rather than prescriptive guidance, for how Member States might choose to address enrichment within the EHDS framework.

Note: Any sharing of enrichment outputs must respect GDPR and EHDS governance safeguards (including SPE controls, permit conditions, and logging). Where relevant, Member States may also need to address IPR and trade secret considerations (e.g., through contractual terms or restrictions on export). Enrichment activities should not undermine existing protective measures.

5.1 Considerations for data users

Building on the previous chapter, data enrichment is an activity that can be performed by the data user within the SPE after accessing the datasets approved under a data permit. Once enrichment activities are complete, the data user must consider whether and to what extent, these operations should be communicated or shared, for example, to support transparency, reproducibility or potential reuse by future researchers.

1. **Should the enrichment be communicated?**

Communication of data enrichment may add value for future data users, but it is not a legal requirement under the EHDS regulation. Reporting data enrichment activities helps ensure that other researchers can reproduce the analytical workflow and understand the transformations applied to the dataset. Secondary use of health data is designed to promote reuse and cumulative knowledge generation. Therefore, dedicated mechanisms for reporting data enrichment activities could add significant value, allowing future data users of the same dataset to benefit from additional derived variables, annotations, or contextual information. However, reporting enrichment is an additional administrative burden, so data users should evaluate the importance and potential utility of the enrichment before deciding to report it.

Key questions a data user could consider include:

- **Relevance beyond the current project:** Is the enrichment specific to my research question, or does it have broader applicability that could benefit future studies using the same dataset?
- **Significance of effort and outcome:** Does the enrichment represent a substantial analytical effort, generating valuable new information or variables that substantially increase the dataset's utility?
- **Regulatory and ethical considerations:** Can the enrichment activities be reported and shared without compromising the data permit, ethical approvals or data minimisations principles?

If these considerations indicate potential added value for future research, the user could be encouraged to share the enrichment with HDAB through appropriate data enrichment communication channels.

2. How could enrichment activities be communicated?

Where Member States establish mechanisms for reporting enrichment, the use of structured templates and proportionate documentation is recommended to support interoperability and reduce burden.

Illustrative examples for elements that could be included in enrichment reporting include:

- Purpose and rationale: Explain why the enrichment was needed for the analysis and why it may also hold value for future research. This should include any limitations that might prevent broad reuse.
- High-level description of the enrichment: Summarise what was added, derived, or annotated (e.g., new variables, risk scores, image segmentation labels, contextual attributes).
- Detailed description of enrichment operations: Provide a clear and structured description of the steps undertaken, including the operations performed, methods, tools, and algorithms used and the logic behind derivations or annotations. This should allow HDABs and data holders to understand the process.

- Key outputs: Provide a list of the new dataset attributes created, including their meaning, format, and intended interpretation.
- Proportionality of reporting: Trivial enrichments (e.g., simple categorisation, permutation of variables for analysis) may require only brief documentation while substantial enrichments (e.g., linkage to authorised external datasets, algorithm-derived features, domain expert annotations) should include a more detailed technical package.

Where national rules provide for it, enrichment documentation prepared by data users may be shared with the HDAB and data holder to support transparency, reproducibility, and future dataset assessment. Well-documented enrichment packages can become valuable reusable artefacts within the EHDS, enhancing dataset discoverability, reducing duplicated analytical effort, and contributing to cumulative scientific knowledge. Member States may consider whether and how enrichment documentation should be retained, shared, or integrated into dataset metadata for future users.

Looking ahead, Member States may wish to explore how enrichment documentation could be integrated into broader platforms for submitting and managing analytical outputs. Some EU initiatives, such as the QUANTUM project, are developing mechanisms to capture structured feedback from data users on dataset adequacy and analytical experiences (fit-for-purpose mechanisms). Where feasible and legally sound, coordinating enrichment documentation with such feedback mechanisms could enhance usability and help capture insights from enrichment activities in a unified system, though such coordination remains a matter of national choice.

In this context, it is also relevant to note that, under the EHDS framework (Article 74), health data users may be considered data controllers with respect to the permitted dataset. In practice, however, their ability to exercise typical controller prerogatives, such as long-term control over datasets or reuse of processed data, may be limited by the conditions of the data permit, including restrictions on data retention and reuse. This creates a specific governance configuration that differs from traditional data controllership models and may require further clarification at national or operational level.

5.2 Considerations for HDABs regarding data enrichment

HDABs are primarily responsible for facilitating and coordinating lawful, secure, and proportionate access to datasets within SPEs and for controlling the exports of data. With regards to data enrichment, where Member States establish feedback mechanisms to report data enrichment activities, HDABs could act as intermediaries for sharing enrichment notes, documentation and, if applicable, enriched data.

Where data enrichment outputs are proposed for export, HDABs may consider whether export conditions remain met in line with SPE rules and permit conditions. This section presents illustrative considerations and options for how HDABs could interact with enrichment activities, without implying new legal obligations or monitoring duties. The focus is on enabling transparency, knowledge-sharing, and safe handling of enrichment outputs where relevant.

The resulting potential workflows are detailed in Figure 2. This figure illustrates one possible enrichment feedback pathway in which the HDAB acts as an intermediary and enrichment-related information is eventually transferred. However, Member States may establish different approaches, including:

- Direct communication between data user and data holder (with or without HDAB intermediation)
- Sharing of enrichment methods, documentation, and code only (without dataset transfer)
- No formal enrichment feedback mechanism at all

The figure shown below represents three illustrative pathways. The specific approach adopted will depend on national legal frameworks, institutional capacity, and national discretionary choices.

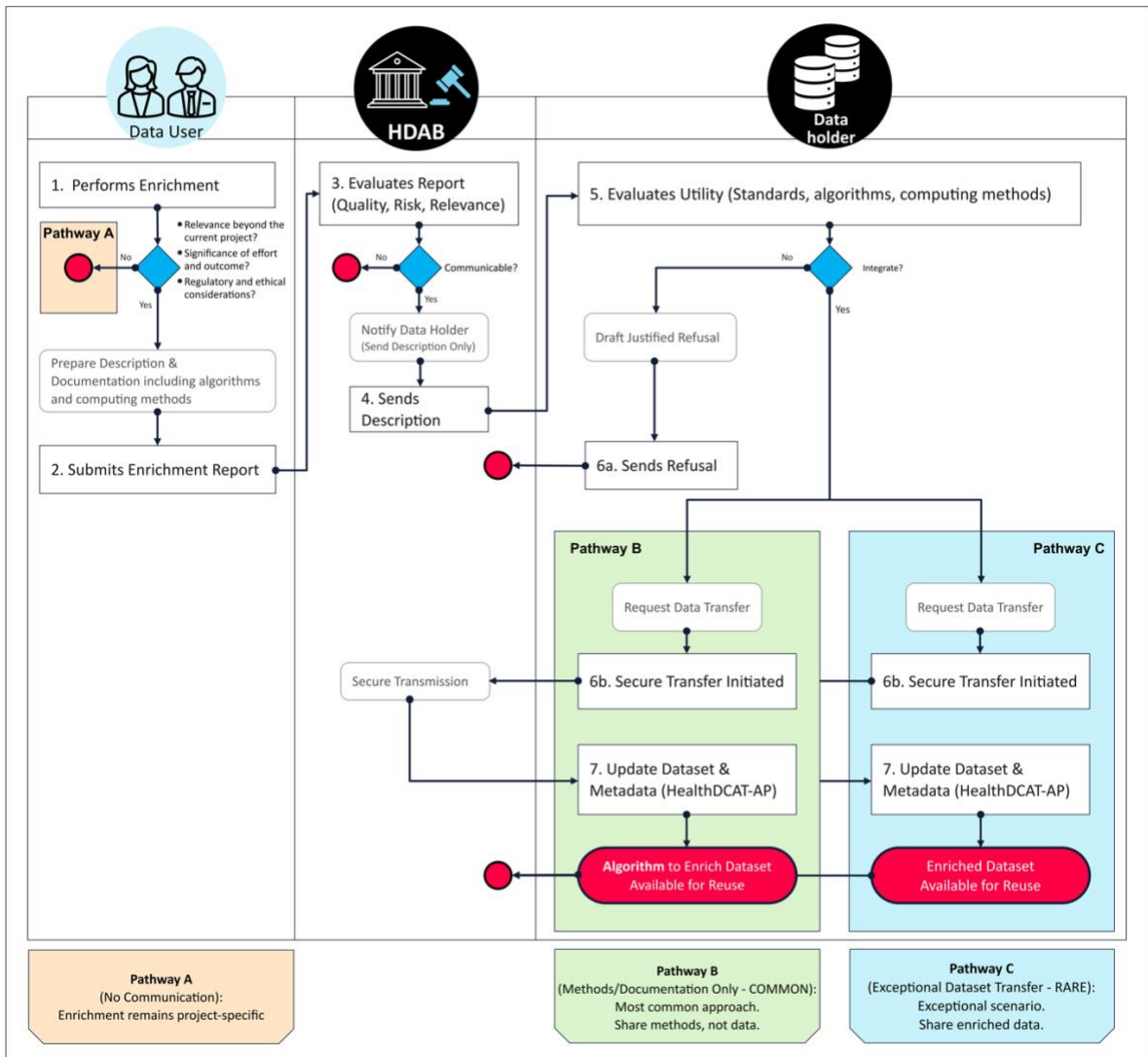
Pathway A represents scenarios where enrichment remains project-specific: when documentation is inadequate, the scope too narrow or the methodology insufficiently generalisable, the process concludes without notification to the data holder and the enrichment remains confined to the original project context.

Pathway B constitutes the probably most common approach; wherein well-documented and potentially valuable enrichment methodologies may be shared with data holders through the HDAB. Crucially, this pathway would involve the transfer of methods, code and documentation only, not the enriched dataset itself, enabling data holders to evaluate and independently apply the methodology to their own datasets without any data transfer occurring.

Pathway C represents an exceptional scenario that may occur rarely: dataset transfer is considered only when the enrichment demonstrates exceptional value, the data holder explicitly requests it and national regulations explicitly permit such transfer. This pathway would necessitate formal controller-to-controller legal agreements and would require that the data holder possesses adequate infrastructure capacity to receive and process the enriched data.

The following diagram illustrates these three pathways and their respective decision points.

Figure 2: Example for potential workflows of data enrichment communication



5.2.1 Assessing export of enriched data: illustrative considerations

In the EHDS framework, datasets are accessed under strict governance, technical measures and data protection workflows, including pseudonymisation, anonymisation, or other minimisation techniques applied before data are made available in the SPE. Most enrichment operations occur entirely within the SPE and are never exported and therefore do not trigger additional control points. However, where enriched datasets are proposed for export, HDABs could consider whether the export conditions continue to be met, consistent with SPE rules and the terms of the data permit. Enrichment can change the risk profile of a dataset, because:

- New variables can increase granularity.
- New combinations of attributes can create more unique individual profiles.
- Derived information can inadvertently reveal sensitive patterns not present before.

- Additional contextual detail may unintentionally strengthen re-identification pathways.

Why enrichment can increase re-identification risk.

Practical EHDS examples

- Rare diseases – Rare disease datasets often contain very small cohorts. When enrichment adds derived clinical phenotypes, unique genetic variants, or detailed longitudinal disease trajectories, individual profiles can become extremely distinctive. Even if the original dataset passed anonymisation thresholds, the enriched version may no longer do so due to the heightened uniqueness of cases.
- Paediatric oncology – Children with specific cancers (e.g., neuroblastoma, retinoblastoma) form highly localised and small populations. Enrichments such as additional clinical markers, treatment response indicators, or granular timelines make these individuals even more distinguishable. The combination of rarity, geography, and new variables can raise identifiability risks significantly.
- Genetic and genomic data – Enrichment with derived variables such as polygenic risk scores, rare variant annotations, or inferred disease predispositions can produce extremely unique feature combinations. Even when genetic data are minimised or aggregated at the source, derived enrichments may inadvertently reintroduce uniqueness that undermines the anonymity of the dataset.

These examples illustrate why HDABs must treat enriched datasets as potentially new risk objects, requiring a fresh evaluation before any decision on export.

Therefore, the transmission of enriched data sets should also be excluded in further regulations of the Member States wherever possible. Should it be necessary and justifiable in exceptional cases to export enriched data sets from the SPE, then it should be made mandatory that HDAB verifies that the enriched dataset still complies with anonymisation or pseudonymisation standards, before considering authorising the export of an enriched dataset from the SPE. In other words, the risk of re-identification must be re-evaluated after enrichment and should not assumed to be unchanged.

In such cases, it would be advisable for the HDAB assessment to take the following measures into account:

- Evaluate identifiability risks introduced by new derived variables, contextual attributes, annotations, or reconstructed longitudinal patterns
- Assess whether enrichment creates new variable combinations with higher singling-out potential or that could facilitate inference attacks
- Verify continued compliance with data minimisation and alignment with the authorised purpose under the data permit

Only if the enriched dataset meets the required protection standards an export authorisation should even be considered. Additional safeguards may also be necessary to maintain an acceptable risk profile. To maintain the security and confidentiality of datasets, the HDAB could assess the re-identification risk in a practical and robust manner using modern statistical, cryptographic and infrastructural methods without directly viewing the data.

5.2.2 HDAB considerations for sharing enrichment outputs

As mentioned in the previous section, data users should communicate to the HDAB a description of the results and outputs produced within the SPE (Article 61). Ideally, this process should also include a dedicated channel for communicating data enrichment operations, in line with the criteria discussed in Section 5.1.

When the data user would communicate an enrichment, the HDAB should conduct an evaluation based on the information provided. The purpose of this review is twofold:

1. Assess the quality, completeness, and relevance of the enrichment documentation.
2. Determine whether the enrichment should be communicated to the original data holder(s).

Importantly, this evaluation should balance the potential scientific or operational value of the enrichment with the need to avoid unnecessary administrative pressure on HDABs and data holders, already among the most heavily tasked actors in the secondary use framework. Therefore, the process should follow principles of proportionality, efficiency, and harmonisation, supported by streamlined procedures across Member States.

Possible outcomes of the HDAB evaluation

Where Member States establish mechanisms for enrichment reporting, HDABs may need to make decisions about whether enrichment documentation warrants communication to data holders. The following scenarios illustrate possible approaches:

Communication to the data holder(s): If enrichment is well-documented, technically sound, and potentially applicable beyond the specific project, an HDAB might choose to communicate it to relevant data holder(s). This would not imply any obligation for the data holder to update datasets but could support knowledge-sharing and continuous improvement.

No communication required: If enrichment is narrowly scoped, inadequately documented, or dependent on project-specific assumptions that limit wider reuse, an HDAB might decide not to inform the data holder(s). This approach keeps the process proportionate and avoids unnecessary workload.

Where HDABs choose to assess enrichment documentation, the following criteria are presented as illustrative examples of factors that *could* guide such evaluation. These are not mandatory requirements but rather examples of how Member States might structure proportionate decision-making (Table 2).

Table 1: Illustrative criteria for enrichment evaluation

Type of criteria	Guiding questions
Completeness and clarity of documentation	Are the purpose and rationale of the enrichment clearly explained?
	Is the process described in a transparent, step-by-step manner?
	Are methods, tools, and algorithms adequately described?
	Are the new dataset attributes listed and interpreted?
Scientific and practical relevance	Does the enrichment have potential value for future reuse beyond the specific project?
	Is it aligned with recognised scientific, clinical, or public health methodologies?
Applicability and generalisability	Is the enrichment understandable and reproducible by other experts?
	Does the enrichment avoid excessive reliance on project-specific parameters?
	Could it reasonably apply to other studies using the same dataset or similar datasets?
Impact on data protection and risk profile	Does the enrichment change the identifiability risk for individuals?
	Has it introduced new or derived variables that increase granularity?
Administrative proportionality	Is the significance of the enrichment sufficient to justify involving the data holder(s)?
	Would communicating the enrichment meaningfully support data holder quality improvement or enrich catalogue metadata?
Multi-source considerations	If multiple data holders contributed to the enriched dataset, which ones should be informed?
	Does the enrichment apply equally to all sources, or only to a subset?

Design principles for enrichment workflows: Member States choosing to establish enrichment procedures should design them carefully to avoid creating bottlenecks for HDABs and data holders, who already carry extensive responsibilities under the EHDS. A well-defined, proportionate workflow supported by templates, clear decision rules, and harmonised criteria can help ensure that scientifically meaningful enrichments are captured while keeping processes streamlined and predictable. However, such frameworks remain optional and should reflect national priorities and institutional capacity.

5.3 Considerations for data holders regarding enrichment outputs

In the vision expressed in Recital 57 EHDS Regulation, when a data user produces a meaningful enrichment that improves the analytical value of a dataset, the underlying

enrichment concept should be made available to the original data holder. Ideally, this communication would consist of a detailed description of the enrichment process and, where appropriate, the corresponding code or algorithm, rather than through the transmission of the enriched dataset itself. The data holder may consider whether this enrichment could be applied to improve datasets available for secondary use, or, in certain cases, inform enhancements to primary data systems. This approach would support the broader objective of strengthening the reusability, quality, and interoperability of datasets within the EHDS, while maintaining a clear separation between methodological innovation and data transfer.

However, in practice, transferring enriched data to data holders would require careful consideration of the safeguards governing the secondary use environment, since it would imply data flows from the SPE to the HDAB and onwards to the data holder that are not envisaged in the current guidelines. The current versions of the guidelines on SPE (7.1. and 7.4), pseudonymisation (7.2) and record linkage (7.5) do not permit such data flows, and any attempt to reintroduce enriched data would create additional responsibilities, new risk-management steps and potential privacy risks. For this reason, only enrichment descriptions and, where relevant, executable procedures or code should be communicated, and any reintegration of the resulting outputs into operational data environments should occur, if at all, under clearly justified and tightly controlled conditions.

Communication pathways for enrichment documentation: Member States may establish different approaches for how enrichment documentation reaches data holders. In some national frameworks, the HDAB may act as an intermediary, reviewing enrichment reports against criteria such as those outlined in Section 5.2, assessing methodological soundness and broader utility, and forwarding relevant documentation to data holder(s) where warranted. In other frameworks, data users or other parties might communicate enrichment documentation directly to data holders, or enrichment reporting may not be formally structured at all. Where such communication does occur, only methodological descriptions and, where appropriate, associated code or implementation logic would be shared; no enriched dataset would be transmitted. This allows data holders to assess potential value, feasibility, and resource implications without receiving or processing additional data from the SPE.

Given that responsibilities for data holders regarding enrichment are not defined in binding terms in the EHDS Regulation, each Member State is encouraged to specify national rules, processes, and expectations for how enrichment descriptions and algorithms are to be evaluated and, where appropriate, implemented. A clear national framework, whether or not it involves formal HDAB review, can help ensure consistent, proportionate processes without creating excessive administrative burdens on data holders and HDABs, who already manage significant obligations under the secondary use regime.

Evaluation considerations for data holders

Upon receiving the enrichment description (and, where applicable, code) from the HDAB, the data holder should determine whether the enrichment, if applied internally, would merit integration into its datasets for secondary use or, in certain cases and subject to national rules, whether specific elements could even inform improvements in primary data systems.

The evaluation would accordingly focus on the conceptual and technical merits of the enrichment approach rather than on the reuse of any specific enriched dataset, in line with the standard model that prioritises the circulation of methods over the transfer of data. Applying the following criteria could support a structured evaluation:

- **Methodological soundness:** Evaluate whether the enrichment is based on clear, reproducible, and technically or scientifically valid methods. The description should allow the data holder, or another independent expert, to understand how the enrichment was carried out and assess whether it follows recognised good practices.
- **Traceability and documentation quality:** All transformations, derivations, and new variables must be traceable. The enrichment description should identify what was added or transformed, why these steps were taken but also the tools, algorithms, or external sources used or any assumptions or thresholds applied.
- **Alignment with standards and terminologies:** Determine whether the enrichment is consistent with relevant standards. Standard-aligned enrichments tend to offer higher added value and improve interoperability.
- **Contribution to data quality:** The data holder should assess whether the enrichment improves completeness, consistency, accuracy, structure, or harmonisation. For instance, transforming free-text medication fields into structured terminology may significantly improve future usability.
- **Broader applicability:** The enrichment should ideally provide value beyond the original research context. A general-purpose improvement, such as deriving standard clinical indicators or harmonising variable naming conventions, has greater potential for reuse than a narrowly scoped transformation.

If the data holder would decide not to integrate the enrichment, it should provide the HDAB with a justified explanation, grounded in the evaluation criteria above. Reasons may include lack of broader relevance, insufficient methodological transparency, incompatibility with standards, or potential privacy risks. Additionally, one important point is also the feasibility of integrating or updating the dataset. In many situations, for instance when dealing with data originating from electronic health record systems or other operational infrastructures, it may simply not be technically or organisationally possible to incorporate enrichment descriptions, executable procedures, code nor the enriched dataset. Limitations in data models, system architecture, or clinical audit trail requirements may prevent any modification of the original records. Even when the enrichment is of high quality, the data holder may conclude that it can only be used within a secondary-use context and cannot replace or update the primary data collection. Nevertheless, data enrichment may provide important insights into systematic errors within the original dataset or reveal recurring weaknesses such as inconsistent coding practices, structurally missing fields, or patterns of misclassification. Even if the enriched dataset cannot be integrated into the operational data environment, these findings can help the data holder identify underlying issues in upstream data collection processes and improve the quality of future data flows. In this way, the value of enrichment extends beyond the dataset itself, contributing to broader system-level improvements.

Secure transmission (where applicable)

In the exceptional case where a data holder determines that an enrichment is valuable and decides to integrate an enriched dataset into its systems, such transfer would need to follow secure data exchange protocols agreed between the HDAB and the data holder, ensuring that all technical and legal safeguards are respected in accordance with EHDS requirements. However, as discussed earlier, such dataset transfer represents a significant exception and may not be feasible or permitted under current SPE and output control guidelines in many Member States. Should such integration occur, the data holder may document the decision, update its dataset records, and revise the dataset description to reflect the enrichment, making the updated dataset findable for future secondary use. However, the decision to integrate, store, or make enriched datasets available for future use remains entirely within the data holder's discretion and is subject to national rules and institutional capacity.

How to describe an enriched dataset

Under the EHDS framework, each dataset available for secondary use must have a unique metadata record in the corresponding national dataset catalogue, reviewed at least every year, should reflect the information about the dataset. When a data holder decides, under the applicable legal and technical framework, to make a revised dataset available for secondary use, the corresponding metadata should be updated accordingly.

To manage this update, the versioning properties in HealthDCAT-AP should be used to link the new enriched dataset to the previous version. Specifically:

- **isVersionOf** should be used to link the new enriched version of the dataset to its original version.
- **version** will be updated to reflect the new version number of the enriched dataset.

Example 1: The original dataset is version 1.0. A data user performs enrichment, and the resulting enriched dataset is now version 2.0. The metadata for version 2.0 would be updated as follows:

- **Version:** "2.0"
- **isVersionOf:** "1.0"

This way, version 2.0 is clearly identified as the enriched version of version 1.0, and future data users will know that they are working with the updated dataset. Therefore, this versioning ensures that users can track the evolution of the dataset over time and identify which version of the dataset is the most current.

In addition to using the *version* and *isVersionOf* properties to track dataset versions, it is also important to fill the *versionNote* property. The *versionNote* provides a description of the differences between the current version and the previous one, offering context to data users about the nature of the update. When updating the metadata for an enriched dataset, the *versionNote* property should be used to explain why the new version was created and what changes were made. In the specific context of the EHDS, this would involve describing the

data enrichment performed and the verifications made by the data holder. This helps ensure that future data users understand the specific improvements made in the new version.

Example 2: The original dataset is version 1.0 of a patient medical record dataset, and the data enrichment involves adding semantic annotations using the SNOMED CT vocabulary. The versionNote for the enriched dataset 2.0 could be written as:

- **Version:** "2.0"
- **isVersionOf:** "1.0"
- **VersionNote:** *"This dataset was enriched during the analysis associated with the EHDS data permit [data permit identifier]. The enrichment process involved adding semantic annotations to the variables related to medical conditions and procedures, using the SNOMED CT taxonomy. These annotations were meticulously verified by the data holder to ensure their accuracy and compliance with relevant standards. The data holder assessed the enrichment and confirmed that it improves the dataset's utility for future secondary use. As a result, the enriched dataset has been approved for future requests and is now considered the latest version for re-use."*

Note: For complete information on the use of HealthDCAT-AP, consult *D5.1 - Guidelines for Data Holder on Data Description*.

6 Considerations for implementation

The previous chapters have defined what data enrichment entails and where it fits within the EHDS user journey (Chapter 4) and have explored in depth the potential data enrichment journey, highlighting the key actors involved and their respective roles in illustrative frameworks. As noted, data enrichment is not legally defined or binding under the EHDS Regulation; therefore, Chapter 4 presents a conceptual and operational perspective on how a data enrichment framework could be established in the context of secondary use of health data.

This chapter builds on that foundation by examining practical considerations for implementing data enrichment activities. In practice, and as discussed in previous sections, data enrichment feedback is most commonly operationalised through the sharing of documentation, methods, code, or executable procedures rather than through the transfer of enriched datasets. The considerations below should be read in that light. This chapter addresses a range of domains, including legal, operational, and organizational aspects, and highlights challenges and opportunities for integrating data enrichment into the EHDS, supporting the long-term goal of enabling secure, efficient, and high-quality reuse of health data.

6.1 Legal framework for sharing enrichment outputs

The legal requirements for sharing enrichment-related information depend on what is being shared.

1) Sharing enrichment documentation, methods, or code

Where enrichment feedback consists of documentation, methodologies, algorithms, or executable code, rather than data itself, this is not a transfer of personal data and does not trigger the same data processing agreements as data transfers. In such cases, the data holder may receive methodological information that they may choose to apply to their own datasets as the original controller. Standard contractual frameworks for such knowledge-sharing may be simpler than those required for personal data transfers, though intellectual property, trade secret protection, and attribution considerations may still apply depending on national rules and institutional agreements.

2) Transfer of enriched datasets (exceptional cases)

If, in exceptional cases and under clearly defined national rules, enriched datasets themselves are transferred to the data holder, the legal framework becomes more complex. Such transfers would constitute controller-to-controller data transmissions rather than processor-controller relationships. In these cases, appropriate legal or contractual instruments would be necessary to define:

- Data protection obligations and safeguards applicable to the enriched dataset
- Permitted uses and any restrictions on reuse
- Security measures and liability arrangements
- Data provenance and documentation of processing history
- Intellectual property and ownership considerations

The data permit issued by the HDAB is a legally binding instrument that governs how data can be used within the SPE, including specific conditions and limitations. However, the EHDS Regulation does not currently contain explicit provisions related to data enrichment or the feedback of enriched datasets to data holders. Where Member States choose to establish frameworks for enrichment feedback, national legal instruments should clarify whether and under what conditions enrichment documentation or code may be shared, whether and under what conditions enriched datasets (as a rare exception) may be transferred, the roles and responsibilities of actors involved in such exchange and the data protection and security requirements applicable to each scenario.

Existing European initiatives provide illustrative examples of alternative approaches. For instance, the Federated European Genome-phenome Archive (FEGA) enables the storage and controlled reuse of genomic datasets across national nodes. In this model, data are made available through Data Access Committees, and datasets, including processed or derived data, can be reused under governed conditions across multiple research projects. Similar federated approaches are being implemented in initiatives such as the European Genome-phenome Archive (EGA) and the One Million Genomes initiative.

While these models differ from the current EHDS implementation, they illustrate how controlled reuse of enriched or processed datasets could support reproducibility, scientific accountability, and efficient reuse of resources, while maintaining strong governance and data protection safeguards. These examples may provide useful reference points for

informing ongoing and future discussions on the operationalisation of data enrichment within the EHDS.

This highlights the need for Member States to develop clear national legal guidance on enrichment feedback mechanisms, distinguishing between the simpler case of sharing methodological information and the more complex case of transferring data in enriched form.

Proposed legal solutions and governance framework on Member State level

1) Legal and contractual considerations

To accommodate data enrichment feedback within the EHDS framework, Member States might consider:

- Integrating enrichment-related clauses into existing instruments such as the data permit or the controller–processor agreement (as defined under Article 74), specifying whether and under what conditions enrichment documentation or outputs may be shared.
- Developing a dedicated Data Use Agreement (DUA) or similar instrument between the HDAB and the data user, where applicable, to clarify enrichment rights, permitted transformations, output sharing rules, and return mechanisms.

These legal instruments should distinguish between the simpler case of sharing enrichment methods and documentation (which does not involve personal data transfer) and the exceptional case of transferring enriched datasets (which would require additional safeguards and controller-to-controller agreements).

2) Governance framework for enrichment feedback

Where Member States establish mechanisms for enrichment feedback, a governance framework for a feedback loop may help structure the process. Such a framework could address:

- How data users may submit enrichment documentation, methods, or (in exceptional cases) enriched datasets
- Quality criteria and validation procedures for assessing enrichment submissions
- Conditions under which a data holder may choose to consider integration of enrichment-related information into its systems
- Intellectual property, licensing, and attribution arrangements
- Metadata and data provenance tracking to document the origin and processing history of enrichments

3) Data protection and security considerations

Enrichment feedback mechanisms must comply with GDPR and EHDS safeguards:

- Where enrichment documentation or code is shared (not personal data), standard data protection obligations for that information still apply (e.g., confidentiality, proper handling of trade secrets or proprietary methods).
- Where enriched datasets are exceptionally transferred, the transfer must be governed by appropriate controller-to-controller agreements that specify data protection obligations, security measures, permitted uses, and retention periods.
- All enrichment feedback should include comprehensive documentation of data provenance and processing history, enabling data holders to understand the origin and modifications applied to enriched information.
- Output controls and security measures applicable within the SPE should inform decisions about what enrichment outputs (if any) may leave the SPE and under what conditions.

Member States are encouraged to develop clear national guidance on these legal, governance, and data protection requirements, tailored to their institutional arrangements and national legal frameworks. Such guidance would help ensure that enrichment feedback mechanisms, where they are established, operate transparently, consistently, and in compliance with EHDS requirements.

6.2 Operational considerations

As Member States prepare for the operationalisation of the EHDS, the integration of data enrichment processes into national and cross-border data infrastructures will become essential. Future implementation of data enrichment should align with the EHDS Regulation, especially Article 51 and Recital 57, supporting secure, lawful and value-driven reuse of health data.

A forward-looking approach to data enrichment must account for both flexibility in execution and harmonisation across actors (health data users, HDABs, and data holders). It is anticipated that enrichment will evolve beyond local or project-based activities to become an embedded element of the data lifecycle. This implies that enriched data will increasingly need to be managed through structured workflows, using standards such as HealthDCAT-AP for metadata, and HL7 FHIR and SNOMED CT for semantic interoperability. Automation capabilities for detecting critical errors, standardising terminologies, and supporting audit trails can support traceability and compliance with data protection requirements.

Member States and HDABs should anticipate that enrichment activities, particularly where they involve complex computations, external data integration, or the processing of sensitive data types, may have implications for SPE infrastructures. These may include computational resources, as enrichment operations (e.g., algorithm-driven feature engineering, machine learning-based annotations) may require additional computational capacity depending on dataset size and enrichment complexity. Also, where enrichment activities were not initially anticipated in a data permit, consideration should be given to whether the permit's conditions adequately cover the enrichment operations being performed, or whether permit amendment or clarification is necessary to ensure compliance.

These considerations should inform Member State planning for enrichment integration, ensuring that SPE infrastructures are adequately resourced and that data permits are sufficiently clear about what processing activities are authorised.

6.2.1 Possible ways to implement and to perform

Implementation of data enrichment can follow a modular approach, adaptable to the capabilities of data users and institutional maturity across EU Member States. Several paths can be envisaged:

Enrichment within Secure Processing Environments (SPEs): Data users perform enrichment tasks such as annotation, standardisation, or deduplication within SPEs, ensuring that enriched outputs remain secure and compliant. These enrichments may include the use of validated AI tools (e.g., image segmentation or NLP-based concept extraction).

Workflow Integration: Enrichment processes are embedded in existing data workflows, especially during data cleaning, integration and preparation stages. Tools like pipeline automation, transformation logging and enrichment validation rulesets can support consistency.

User-led contributions with HDAB coordination: Health data users may voluntarily return their enrichment concepts to the HDAB, following documentation standards. HDABs act as intermediaries, validating documentation (not data quality) and relaying improvements to data holders.

Federated implementation models: Harmonised interoperability standards are critical for data enrichment in any context where multiple data sources, holders, or users are involved. This includes cross-border scenarios as well as domestic situations with multiple data holders or user consortia. Standards such as SNOMED CT, ATC, Identification of medicinal products (IDMP), and LOINC, supported by mapping services and terminology servers, enable consistent enrichment across different datasets and ensure that enrichments remain interpretable and reusable across the EHDS ecosystem. Member States implementing enrichment mechanisms should consider how standardisation supports the reproducibility and portability of enrichment operations.

Use of APIs and Metadata Registries: HealthDCAT-AP-based dataset registries and APIs may support programmatic exchange of enriched datasets, including versioning, annotation references and traceability attributes.

Audit and reproducibility mechanisms: Each enrichment process should be traceable through logs or metadata annotations to allow future reuse, verification, or rejection by data holders EHDS Regulation Article 51(3).

6.2.2 Processing enriched dataset

Processing enriched datasets in the EHDS ecosystem involves maintaining both semantic integrity and legal compliance, while supporting reuse and transparency. Key considerations include:

- **Semantic traceability:** Enriched datasets must retain references to the original dataset (e.g., via `isVersionOf` and `versionNote` in HealthDCAT-AP). This ensures that downstream users understand the source, method and scope of changes.
- **Validation of enrichments:** While HDABs do not evaluate the data quality of enrichments, data holders are expected to review the methodological soundness, traceability and alignment with standards before accepting the dataset for reuse.
- **Version control:** All enrichment operations must support dataset versioning, with structured metadata describing the changes. This may involve integration with dataset catalogues and workflow tools supporting reproducible updates.
- **Security and lawful processing:** The transmission and use of enriched datasets must follow SPE protocols, data minimisation principles and user consent models where applicable. Data enrichment logs should be ready for audit.
- **Impact assessment and reuse:** Data holders must assess the added value of the enrichment, whether it improves completeness, enables interoperability or addresses systemic errors. Enrichments that meet these criteria should be incorporated and made discoverable for future secondary use.
- **Exclusion of low-value enrichment:** Outputs that lack reproducibility, transparency, or general applicability should not be included. These include unverified model outputs or subjective manual annotations.

Reference: EHDS Regulation (EU) 2025/327, Article 51(3); HealthDCAT-AP usage guidelines (D5.1); Buczek et al., 2023 [1]; Eurostat data validation framework (2025) [10].

6.3 Semantic interoperability

Semantic interoperability is a foundational requirement in the EHDS ecosystem, enabling health data to be accurately interpreted and meaningfully exchanged across systems, languages and national borders. It ensures that the meaning of data elements is preserved during sharing and reuse by aligning with standardized terminologies, ontologies and data models such as SNOMED CT, LOINC, ICD, and HL7 FHIR. [13-16]

In the context of secondary use, while the underlying datasets may not always fully conform to shared semantic standards, aligning enriched data with common frameworks can enhance understandability, comparability, and analytical value. Therefore, metadata describing datasets, including enriched outputs, should follow EHDS requirements such as HealthDCAT-AP [17], which facilitate harmonization and linkage between datasets. Robust semantic interoperability not only enhances data quality and usability but also underpins safe, ethical and effective health data exchange across Europe [12, 15].

6.4 Organizational aspects

Several organizational considerations support the proposal that enriched data sets are not returned to the data holder via the HDAB. Creation of real-world evidence (RWE) from real

world data (RWD) implies indeed enriching the origin data with new variables, via a process called 'operationalisation', 'measurement' or 'phenotyping'. [18]

- First, RWD are not static datasets. They usually evolve in time and many with thousands of new records added every day. The enrichment per se is only done on a specific snapshot of the datasets and is not valuable per se, because it is not reproducible.
- Second, often a snapshot of RWD is requested and stored in the SPE from different data banks and from different data controllers, therefore, enrichment is executed through data linkage across data banks based on pseudonymised identifiers. In this case, it is not clear who should receive the enriched dataset anyway.
- Third, the origin datasets can be enormous in volume and so may be the enriched datasets. Handling this volume is unfeasible.
- Fourth, replicating such huge datasets, removing the copy from the SPE and storing it with the data holder contradicts every principle of data minimisation and the responsible handling of data sources.
- Fifth, documentation and reproducibility can be attained in a leaner manner (see below).

Two exceptions to this proposal are foreseen: 1) the case when the dataset to be enriched is a one-time dataset, so that its enrichment is final and reusable per se 2) the case when a material error in the origin data are found, and return can be the information of the error and a recommendation to improve data collection.

In the next section we list illustrative considerations and practical examples that make data enrichment transparent and reproducible.

6.5 Practical examples

When data enrichment is performed via a computer program, the recommendation is to return the program itself, instead of returning the output. This is consistent with the most recent considerations in the field of generation of RWE [19-21]. This recommendation is articulated as follows

1. a clear documentation of the rules applied by the data holder(s) when creating the snapshot(s) that are being enriched should be shared; this can be articulated in the language of the metadata of each dataset, as restrictions in time and/or in values of the variables; this ensures that the input of the computer program can be reproduced;
2. the computer program should be designed stepwise, and documentation of each step should be returned to the data holder, including metadata of every intermediate dataset; if possible, dummy data of each intermediate dataset should be provided as part of the metadata;
3. if possible, the program itself should be shared as open-source software, and dummy data should be provided that allow running the program and verify its execution;

4. Container virtualization (lightweight, self-contained units that encapsulate application code, runtime environment, libraries, and configuration, and are executed on a host system), for example Docker¹ can be used to reproduce and store the environment where the computer program has been developed and executed.

Not only this process allows documentation of the data enrichment but allows the data holder to replicate it if deemed useful.

6.6 Pitfalls and challenges

6.6.1 Usual errors

1) Data user level

Inadequate documentation: Failing to document the enrichment process (data sources, transformation steps, methodologies) reduces transparency and reproducibility, making it difficult for HDABs, data holders, or future users to assess or reuse the enrichment.

Increased re-identification risk: Enrichment operations that add new variables, combine attributes, or incorporate external data can inadvertently increase the risk of re-identifying individuals, potentially undermining existing anonymisation or pseudonymisation measures. Careful evaluation of enrichment's impact on privacy safeguards is essential before sharing any enrichment outputs.

Over-reliance on automation: Automated enrichment tools may introduce errors if not properly validated, particularly with unstructured, multilingual, or domain-specific data. Manual validation steps remain important.

Scope: Enrichment intended for a specific analytical purpose may not generalise well to other research questions or datasets, yet inadequate documentation of limitations may lead to inappropriate reuse.

Limited persistence of enriched outputs: Where enriched datasets or intermediate analytical outputs are not retained or reusable beyond a single project, the long-term scientific value of enrichment activities may be reduced. This can create challenges for reproducibility, validation of results, and cumulative knowledge generation, particularly in complex or resource-intensive analyses.

2) HDAB level

Evaluation burden without clear criteria: Without harmonised assessment criteria, HDABs may struggle to consistently evaluate enrichment submissions, risking either bottlenecks in processing or inconsistent decision-making across Member States.

Unclear governance boundaries: Where enrichment feedback mechanisms are not clearly defined in national rules, HDABs may be uncertain about their role, responsibilities, and the scope of their involvement in reviewing or facilitating enrichment communication.

¹ <https://www.docker.com>

Data permit ambiguity: If data permits do not explicitly address whether enrichment is anticipated or authorised, disputes may arise about whether enrichment operations comply with permit conditions.

3) Data holder level

Poor metadata and provenance management: Not updating or harmonizing metadata after enrichment or failing to track the origin and processing history of enriched data, can lead to misinterpretation, misuse, or loss of traceability in future uses.

Technical infrastructure: Many data holders operate with outdated or insufficient infrastructure. Receiving, validating, and potentially integrating enriched datasets under such conditions can lead to inefficiencies, data loss, security vulnerabilities, or compliance failures.

Resource and capacity constraints: Assessing, validating, and potentially reintegrating enriched datasets requires technical expertise and staff capacity that may not be available, particularly in smaller organisations or resource-constrained settings.

Unclear ownership and IP frameworks: Without clear agreements on intellectual property, licensing, and attribution for enrichments, data holders may be uncertain about their rights to reuse or modify enriched outputs or may face disputes over ownership.

6.6.2 Biases

Selection Bias: Enrichment may disproportionately represent certain populations (e.g., urban, digitally literate), leading to skewed insights and underrepresentation of vulnerable or marginalized groups.

Algorithmic Bias: Machine learning models used in enrichment (e.g., for data imputation or classification) may reflect or amplify existing societal or systemic biases.

Cultural and Linguistic Bias: Enrichment processes that rely on language-specific tools or assumptions may misrepresent data from minority or multilingual populations.

Temporal Bias: Combining data from different time periods without accounting for changes in medical practice, coding standards, or population health can distort trends and lead to inaccurate conclusions.

Confirmation Bias: Enrichment choices may be influenced by preconceived hypotheses, leading to selective inclusion of data that supports expected outcomes while excluding contradictory evidence.

Uneven re-identification risk: The increased granularity of enriched datasets may disproportionately raise the re-identification risk for small, unique, or otherwise identifiable population groups. This uneven distribution of privacy risk requires careful consideration in risk management and ethical oversight.

Trust and Transparency Deficits: Lack of transparency in how enriched data is used can reduce participation and trust, skewing datasets toward more engaged or informed populations and introducing systemic bias.

7 References

- [1] L. Buczek, F. Azar, J. Bauzon, K. Batra, C. Murphy, and S. Wahi-Gururaj, "The Data Error Criteria (DEC) for retrospective studies: development and preliminary application - PubMed," *Journal of investigative medicine : the official publication of the American Federation for Clinical Research*, vol. 71, no. 4, April 2023, doi: 10.1177/10815589231151437.
- [2] N. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research - PubMed," *Journal of the American Medical Informatics Association : JAMIA*, vol. 20, no. 1, 2013, doi: 10.1136/amiajnl-2011-000681.
- [3] K. Y. Yigzaw, A. Michalas, J. G. Bellika, K. Y. Yigzaw, A. Michalas, and J. G. Bellika, "Secure and scalable deduplication of horizontally partitioned health data for privacy-preserving distributed statistical computation," *BMC Medical Informatics and Decision Making 2016 17:1*, vol. 17, no. 1, 2017, doi: 10.1186/s12911-016-0389-x.
- [4] P. Powell and I. Smalley. "Title."in Series Title IBM. [Online]. Available: <https://www.ibm.com/think/topics/data-deduplication>. Accessed: 29. June 2025.
- [5] P. Powell and I. Smalley. "Title."in Series Title IBM. [Online]. Available: <https://www.ibm.com/think/topics/data-consolidation>. Accessed: 29.06.2025.
- [6] K. Harron, "Data linkage in medical research," *BMJ Medicine*, vol. 1, no. 1, 2022, doi: 10.1136/bmjmed-2021-000087.
- [7] M. Gal and D. L. Rubinfeld, "Data Standardization," *SSRN Electronic Journal*, 2019, doi: 10.2139/ssrn.3326377.
- [8] B.-M. Schmidt *et al.*, "Definitions, components and processes of data harmonisation in healthcare: a scoping review," *BMC Medical Informatics and Decision Making 2020 20:1*, vol. 20, no. 1, 2020, doi: 10.1186/s12911-020-01218-7.
- [9] C. Cheng *et al.*, "A General Primer for Data Harmonization," *Scientific Data 2024 11:1*, vol. 11, no. 1, 2024–01–31 2024, doi: 10.1038/s41597-024-02956-3.
- [10] Eurostat. "Data validation." European Commission,. <https://ec.europa.eu/eurostat/web/main/data/data-validation> (accessed 29.06.2025).
- [11] M. Di Zio *et al.*, "Methodology for data validation 1.0," *Essnet Validat Foundation*, 2016.
- [12] *Regulation (EU) 2025/327 of 11 February 2025 on the European Health Data Space and amending Directive 2011/24/EU and Regulation (EU) 2024/2847*, E. Union L 2025/327, 2025.
- [13] *What is SNOMED CT?*, SNOMED International, 2025. [Online]. Available: <https://www.snomed.org/what-is-snomed-ct>. Accessed: 02.07.2025.
- [14] *LOINC*, catalogue Regenstrief Institute, (1994)2025. [Online]. Available: <https://loinc.org/>. Accessed: 02.07.2025.
- [15] *Eleventh Revision of the International Statistical Classification of Diseases and Related Health Problems (ICD-11) Digital Version*, Technical document WHO, 30 November 2021 2021. [Online]. Available: <https://www.who.int/standards/classifications/classification-of-diseases>. Accessed: 02.07.2025.
- [16] *FHIR® Release 5: Standard for Trial Use.*, Health Level Seven International (HL7), 2023, March 26 2023. [Online]. Available: <https://build.fhir.org/history.html> build.fhir.org. Accessed: Retrieved June 18 2025.
- [17] *HealthDCAT-AP – Unofficial Draft 22 December 2023*, Catalogue P. Derycke, 2024. [Online]. Available: <https://healthdcat-ap.github.io/>. Accessed: 02.07.2025.
- [18] F. Dornie *et al.*, "Standardised and Reproducible Phenotyping Using Distributed Analytics and Tools in the Data Analysis and Real World Interrogation Network (DARWIN EU)," *Pharmacoepidemiology and Drug Safety*, vol. 33, no. 11, 2024, doi: 10.1002/pds.70042.

[19] S. V. Wang and A. Pottegård, "Building transparency and reproducibility into the practice of pharmacoepidemiology and outcomes research," *American Journal of Epidemiology*, vol. 193, no. 11, 2024/11/04 2024, doi: 10.1093/aje/kwae087.

[20] J. Tazare *et al.*, "Sharing Is Caring? International Society for Pharmacoepidemiology Review and Recommendations for Sharing Programming Code," *Pharmacoepidemiology and Drug Safety*, vol. 33, no. 9, 2024, doi: 10.1002/pds.5856.

[21] J. Weberpals and S. Wang, "The FAIRification of research in real-world evidence: A practical introduction to reproducible analytic workflows using Git and R - PubMed," *Pharmacoepidemiology and drug safety*, vol. 33, no. 1, 2024, doi: 10.1002/pds.5740.

8 Annexes

Annex number	Annex title
8.1	Methodology
8.2	User journey
8.3	Glossary
8.4	Relation to TEHDAS2 tasks and deliverables
8.5	Links to the EHDS regulation

8.6	Use case examples of data pre-processing and enrichment operations
-----	--

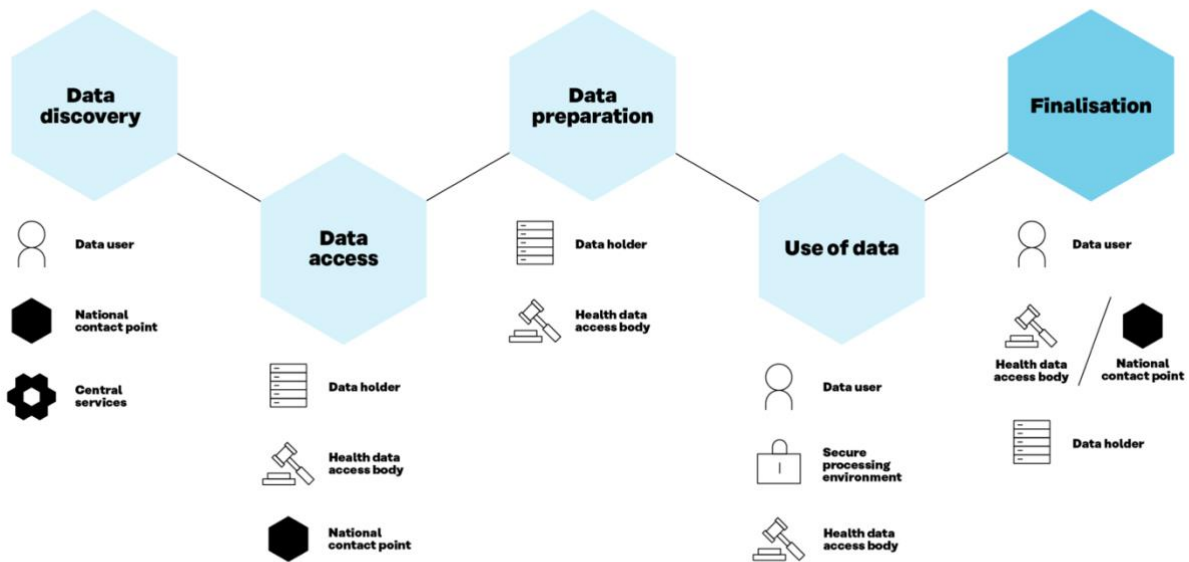
8.1 Methodology

The contributors participated according to their promised commitments, ensuring a collaborative and thorough development process. The following steps were conducted:

- **Desk research** was performed by all contributors. During this process, relevant information was collected from expert organisations, related programmes and entities, such as the Community of Practice, QUANTUM, TEHDAS1.
- **Working meetings** – Regular working meetings were conducted to discuss and outline the key components and structure of the short guide.
- **Responsibilities** for the individual thematic sections were assigned based on available expertise. The responsible partners drew up an initial draft, which was then developed and agreed upon by all participants through repeated commentary and discussion. and finished by a designated contributor.
- **Artificial Intelligence (AI)** tools were employed to support the development of this document, particularly OpenAI’s ChatGPT and DeepL. ChatGPT was used for summary and suggestive purposes. All output generated by ChatGPT was reviewed, edited, and finalised by the contributors. None of the text in the document was independently written by AI or LLM. DeepL was used to improve the English text and to overcome any linguistic barriers. The suggested improvements were reviewed, and subject-specific consistency was ensured.
- **Consultations with DG SANTE** – Four meetings with representatives from DG SANTE were organised to ensure alignment with regulatory requirements and to gather expert feedback.
- **Consultations with related TEHDAS2 tasks** in which alignments between the guidelines were ensured.

8.2 User journey

When a data userⁱ applies for electronic health data for secondary use purposes, such as research and innovation activities, education, and policy-making, within the European Health Data Space (EHDS), the user journey consists of several stages (see Figure 1). Access for certain purposes (public or occupational health, policy-making and regulatory activities, and statistics) is reserved for public sector bodies and Union institutions (see Chapter IV, Art. 53(1) and 53(2)). Figure 1: EHDS user journey consists of five main phases: data discovery, data access, data preparation, use of data and finalisation.



Data discovery

Before being able to use the data, the user needs to investigate whether the data needed is available, and whether it is available in the necessary format for the secondary use purpose. This phase is called data discovery. Datasets available in the EU can be found in a metadata catalogue at <https://qa.data.health.europa.eu/>. Once the data discovery is completed, the user can begin the process of applying for the data.

Data access

In the data access phase, the user fills in and submits a dedicated and standardised data access application form or a data request to a health data access body (HDAB)ⁱⁱ. The user must complete the information required in the form, upload necessary documents, and provide justifications as needed.

Data access application form is used when the user seeks to use personal level data. **Data request** is for cases when the user wants to apply for anonymised statistical data.

Data preparation

During this phase, the data holder(s)ⁱⁱⁱ deliver(s) the necessary data to the HDAB, which starts to prepare the data for secondary use. Techniques for pseudonymisation, anonymisation, generalisation, suppression, and randomisation of personal data are employed. The data minimisation principle (as per the GDPR) must be respected to ensure privacy.

Use of data

In this phase, the user performs analyses based on the received data for the purpose defined in the application phase. Analysing personal level data must be performed in a secure processing environment^{iv}. The duration of this phase is specified in the Regulation (Art 68(12)).

Finalisation

This last phase of the user journey concerns data user's duties regarding analysis outcomes derived from secondary use of data. Data user must publish the results of secondary use of health

data within 18 months of the completion of the data processing in a secure processing environment or of receiving the requested health data. The results should be provided in an anonymous format. The data user must inform the health data access body of the results. In addition, the data user must mention in the output that the results have been obtained by using data in the framework of the EHDS.

8.3 Glossary

Central glossary: *TEHDAS2_Glossary_for_Milestones_and_Deliverables.docx*

Table 2: Key terminology in this guideline

Term	Definition
Anonymisation	The process by which personal data is altered in such a way that a data subject can no longer be identified directly or indirectly. (Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information, Recital 52; EHDS Regulation, Recital 92)
Data controller	A data controller is a person or organisation that determines the purposes and essential means of the processing of personal data. The role of the data controller can be shared by several people or organisations. In that case, they are defined as joint controllers. The controller is accountable and responsible for establishing a lawful data processing workflow and observing the rights of data subjects. (GDPR Article 4(1)(7)).
Data extraction	Data extraction is the process of retrieving data from its source dataset. Structured data extraction involves extracting data from datasets that are already organised in predefined formats. Unstructured data extraction pertains to extracting data from databases handling unstructured formats such as PDFs, images, or free text. There may be one or more different data sources from which data extraction may be required.
Data linkage	The process of combining datasets “from several sources on one topic or data subject” (ISO 5127:2017, 3.1.11.12,). This can be done using unique identifiers, probabilistic methods, or a combination of techniques. Access is only provided to electronic health data that is “adequate, relevant and limited to what is necessary in relation to the purpose of processing indicated in the health data access application by the health data user and in line with the data permit issues pursuant to Article 68.” (EHDS Regulation, Article 66(1)). Data minimisation applies to all stages of the data lifecycle.

Term	Definition
Data minimisation	<p>A principle mandating organisations to only collect, store and process the minimum necessary amount of personal data for a specific purpose. This principle is fundamental under GDPR and relevant to the tasks outlined in EHDS. (GDPR Article 5(1)(c)).</p> <p>Access is only provided to electronic health data that is "adequate, relevant and limited to what is necessary in relation to the purpose of processing indicated in the health data access application by the health data user and in line with the data permit issues pursuant to Article 68." (EHDS Regulation, Article 66(1)).</p> <p>Data minimisation applies to all stages of the data lifecycle</p>
Data permit	<p>An administrative decision issued to a health data user by a Health Data Access Body to process certain electronic health data specified in the data permit for specific secondary use purposes based on conditions laid down in Chapter IV of the EHDS regulation (Art. 2(2v)).</p>
Data preparation	<p>Data preparation is the process in which an organisation (in this case the data holder) transforms and organises raw personal or non-personal health data into one or more datasets (either in individual-based or aggregated form), to comply with a data permit or a data request approval issued by a Data User and approved by the competent Health Data Access Body.</p>
Data Processing	<p>Any operation or set of operations which is performed on personal data or on sets of personal data, whether by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction. (GDPR Article 4(2))</p>
Data Provision	<p>The stage in the data user journey where prepared health data is made accessible to authorised users for secondary purposes.</p>
Data quality	<p>Data quality means the degree to which the elements of electronic health data are suitable for their intended primary use and secondary use; (EHDS Article 2 (2)(z))</p>
Data quality and utility label	<p>Data quality and utility label means a graphic diagram, including a scale, describing the data quality and conditions of use of a dataset. (EHDS Article 2 (2)(aa))</p>
Dataset	<p>A structured collection of electronic health data. (EHDS Article 2(2)(w))</p>
Dataset Catalogue	<p>A collection of dataset descriptions, arranged in a systematic manner and including a user-oriented public part, in which information</p>

Term	Definition
	concerning individual dataset parameters is accessible by electronic means through an online portal. (EHDS Article 2(2)(y))
Data Subset creation	<p>Dataset subset contains only selected records, variables or elements from a larger dataset while maintaining its key characteristics and relationships.</p> <p>Data subset creation refers to the process of extracting the specific portion of a larger database based on defined criteria, to support a particular analysis or creation of a statistical format. This evolves extraction of relevant observation and variables for the specified purpose.</p>
Electronic health data	Personal or non-personal electronic health data (EHDS Article 2(2c)).
Health data access application	An application seeking to access personal-level electronic health data for secondary use in an anonymised or a pseudonymised format (EHDS Article 67).
Health Data Access Body	<p>Member State-designated authority that facilitates the secondary use of electronic health data. HDABs assess the information provided by the health data applicant and decide on health data requests and access applications, authorise and issue data permits, obtain data from data holders and make data available in SPE. HDABs systematically track the data request and data access applications received and the data permits issued. As per Article 58 of the EHDS regulation, HDABs are required to publicly list information on the data permits issued. (EHDS Article 55 and Recital 52)</p> <p>The HDAB duties include:</p> <ul style="list-style-type: none"> – Publishing the data dataset catalogue; – Evaluating health data access applications; – Maintaining records on data access applications and decisions; – Inform citizens on the use of data, the conditions under which data are made available and on how their rights are protected and safeguarded, respectively; – Receiving, preparing and compiling the requested datasets when requested, and properly anonymising or pseudonymising them; – Preserving the confidentiality of intellectual property rights and trade secrets; – providing access to electronic health data to health data users pursuant to a data permit in an SPE;

Term	Definition
	<ul style="list-style-type: none"> - Supervising and enforcing the compliance of data holders and data users; <p>If a Member State has provided for the right to opt out pursuant to Article 71 to be exercised through the (coordinating) health data access bodies, the relevant health data access bodies shall provide public information about the procedure to opt out and facilitate the exercise of that right.</p>
Health data applicant	A natural or legal person submitting a health data access application or a data request to a Health Data Access Body for the purposes referred to in Article 53 of EHDS.
Health data holder	Any person, organisation or public body involved in healthcare, care services, health-related products, wellness apps or health(care) research, that has the right to process data for health care provision or for public health purposes, reimbursement, research, policy making, official statistics or patient safety. This includes, for example, hospitals, insurers, research institutes and EU institutions. For a more detailed definition: EHDS regulation, Article 2(2)(t)).
Health data request	A request to access data in an anonymised statistical format for the purposes referred to in EHDS Article 53. (EHDS Regulation, Article 69)
Health data user	<p>A natural or legal person, including Union institutions, bodies, offices or agencies, which has been granted lawful access to electronic health data for secondary use pursuant to a data permit, a health data request approval or an access approval by an authorised participant in HealthData@EU. (EHDS Article 2(2u))</p> <p>The rights and responsibilities of health data users include:</p> <p>Accessing and processing electronic health data exclusively in accordance with an issued data permit, an approved health data request, or access approval from the relevant authorised participant within HealthData@EU.</p> <p>Ensuring that electronic health data processed within secure processing environments is not shared or disclosed to third parties who are not explicitly identified in the data permit.</p> <p>Refraining from re-identifying or attempting to re-identify natural persons from the electronic health data they have obtained,</p> <p>Publicly disseminating results or outputs from secondary use within 18 months following either the completion of electronic health data processing in the secure processing environment or upon receipt of responses to health data requests,</p>

Term	Definition
	<p>Informing the health data access body of any significant finding related to the health of the natural person whose data are included in the dataset,</p> <p>Cooperating fully with health data access bodies to facilitate the effective performance of their supervisory tasks.</p>
Non-compliance	Any failure to comply with any requirement under the Union harmonisation legislation or under this Regulation; ((EC) No 765/2008 and (EU) No 305/2011)
Non-personal electronic health data	Electronic health data other than personal electronic health data, including both data that have been anonymised so that they no longer relate to an identified or identifiable natural person (the 'data subject') and data that have never related to a data subject. (EHDS Regulation, Article 2(2b))
Open data	<p>Data in an open format that can be freely used, re-used and shared by anyone for any purpose.</p> <p>'Open format' means a file format that is platform-independent and made available to the public without any restriction that impedes the re-use of documents; ((EU) 2019/1024 Open Data Directive)</p>
Open (data) database	Publicly accessible digital data that anyone can freely use, reuse, and redistribute for any purpose.
Personal electronic health data	Data concerning health and genetic data, relating to an identified or identifiable natural person, processed in an electronic form. (EHDS Regulation, Article 2(2a))
Pseudonymisation	The processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure non-attribution to an identified or identifiable person. (GDPR Article 4(5))
Secondary use	Processing of electronic health data for the purposes set out in Chapter IV of EHDS Regulation, other than the initial purposes for which they were collected or produced. (EHDS regulation, Article 2(2e))
Secure Processing Environment (SPE)	An environment in which access to electronic health data can be provided in following a data permit. An SPE is subject to technical and organisational measures and security and interoperability requirements. Specifically allowing access to only those persons listed in the permit, as well as user authentication, authorisation, restricted data handling, logging and the compliance monitoring of respective security measures. (EHDS regulation, Article 73)

Term	Definition
Synthetic data	Data that is artificially generated. The concept of synthetic data generation is to take an original data source (dataset) and create new, artificial data, with similar statistical properties from it.
Tabular data	Data organised in a structured format of rows and columns, where each row represents a single record or entity, and each column represents a specific attribute or variable. This structure is commonly found in spreadsheets or relational databases, making it easy to store, query, and analyse. Tabular data is often used for structured datasets where relationships between variables are well-defined.

8.4 Relation to TEHDAS2 tasks and deliverables

Connections to other tasks in TEHDAS2 are described in the table below.

Table 3: Connections to other tasks in TEHDAS2

Description
Guidance for health data holders on their duties to describe data
Technical specification on the national metadata catalogue
Guidance for health data holders on making personal and non-personal electronic health data available for reuse
Guideline on how to use data in a Secure Processing Environment
Guidance on linkage of health datasets
Guideline to Health Data Access Bodies on implementing the obligation of notifying the natural person on a significant finding from the secondary use of health data

8.5 Links to the EHDS regulation

Relevant EHDS recitals and articles (for references)

The main source for this guideline is

Recital (57)

(57) Health data users who benefit from access to datasets provided for under this Regulation could enrich the data in those datasets with various corrections, annotations and other improvements, for instance by supplementing missing or incomplete data, thus improving the accuracy, completeness or quality of the data in the datasets. Health data users should be encouraged to report critical errors in datasets to health data access bodies. To support the improvement of the initial database and further use of the enriched dataset, Member States should be able to establish rules for the processing and

the use of electronic health data containing improvements related to the processing of those data. The improved dataset should be made available free of charge to the original health data holder together with a description of the improvements. The health data holder should make the new dataset available, unless it provides a justified notification to the health data access body for not doing so, for instance in cases in which the enrichment by the health data user is of low quality. It should be ensured that non-personal electronic health data are available for secondary use. In particular, pathogen genomic data hold significant value for human health, as shown during the COVID-19 pandemic during which timely access to and sharing of such data proved to be essential for the rapid development of detection tools, medical countermeasures and responses to public health threats. The greatest benefit from pathogen genomics efforts will be achieved when public health and research processes share datasets and cooperate to inform and improve each other.

8.6 Use case examples of data pre-processing and enrichment operations

This section provides examples of data enrichment provide from the major contributors' own practice or projects. All examples are from the period prior to the EHDSR and are intended to illustrate the spectrum and possibilities for data enrichment.

Example 1: Harmonisation & Standardisation

Context: Combining two different Hodgkin Lymphoma datasets one from USA patients (DatasetA) and one of Spanish Patients (DatasetB), both with ~200 registries in each one and ~100 fields related with diagnostic parameters, treatment, response, follow up to ~15 years. Those datasets needed to be both harmonized and standardized before any analysis to be performed.

Standardisation approach: Firstly, regarding on the subclassification of all the cases included, originally the subclassification was made in free text and was standardised on the ICD classification at the time of the analysis (Table 1)

Table1: Standardization of Hodgkin Lymphoma classification between datasets

Original HL subtype DatasetA	ICD-10-CM Final version	Original HL subtype DatasetB
Unspecified HL	C81.90 Hodgkin lymphoma, unspecified, unspecified site	Linfoma de Hodgkin NS
HL	C81 Hodgkin lymphoma.	Linfoma de Hodgkin de predominio linfocítico nodular, localización no especificada
Lymphocyte predominant	C81.0 Nodular lymphocyte predominant Hodgkin lymphoma.	Linfoma de Hodgkin de predominio linfocítico nodular, ganglios
CH Nodular sclerosis	C81.1 Nodular sclerosis Hodgkin lymphoma.	Linfoma de Hodgkin clásico tipo esclerosis nodular

Mixed cellularity	C81.2 Mixed cellularity Hodgkin lymphoma.	Linfoma de Hodgkin clásico de celularidad mixta
Lymphocyte rich	C81.4 Lymphocyte-rich Hodgkin lymphoma.	Linfoma de Hodgkin clásico rico en linfocitos
Lymphocyte depletion	C81.3 Lymphocyte depleted Hodgkin lymphoma	Linfoma de Hodgkin clásico con depleción linfocítica

Harmonisation approach: Secondary, we faced the **harmonisation** of some fields, one of the most **paradigmatic examples** was the adjustment of the complete hematic formula, as the way of expressing the values differed greatly between datasets and even within the same dataset due to several factor such as: using different units of measurement, different reference range values or even relative expression of the values (Table2).

These discrepancies need several steps to be able to compare both datasets. First, differences in the measure unit need the mathematical conversion to a common one. Second, it is necessary to use a conversion factor to make the different range values compatible.

Standardisation refers to the process of establishing uniform norms or standards for products, services, or processes. The goal is to ensure that all elements meet certain quality and consistency criteria. This facilitates interoperability, safety, and efficiency, as all participants follow the same rules and specifications. Imposes a uniform set of norms or specifications that everyone must follow.

Harmonisation, on the other hand, involves aligning and coordinating different norms, regulations, or practices so that they are compatible with each other, even if they are not identical. Harmonization aims to reduce differences and conflicts between various systems or regulatory frameworks, allowing for greater coherence and collaboration without imposing a single standard. Adjusts and coordinates different norms or practices to make them compatible, without the need for complete unification.

Table2: Harmonisation of Hemograms of Hodgkin Lymphoma at diagnosis between datasets (example) (* differences in range value; ** differences in measure units)

Hemogram DatasetA			Common Hemogram		Hemogram DatasetB		
Haematological Parameters	(U)	Range	(U) Conversion	Range Adjusted	Haematological Parameters		(U)
Platelet count	(x 10 ⁹ /L)	45-676 *		*	<u>Platelet count</u>	(x10 ⁹ /L)	150-400*
White Cell count	(mm ³)**	4000-10000*	**	*	<u>White Cell count</u>	(X10 ⁹ /mm ³)**	4.5-10.5*
Haemoglobin	(g/dl)**	2,4-14.3*	**	*	<u>Haemoglobin</u>	(g/ml)**	13.8-17.2*
Neutrofiles	(%)	60-77*		*	<u>Neutrofiles</u>	(%)	45-75*
Lymphocytes	(%)	13-27*		*	<u>Lymphocytes</u>	(%)	20-50*
Monocytes	(%)	0.2-0.95*		*	<u>Monocytes</u>	(%)	2-13*
Eosinophils	(%)	2-10*		*	<u>Eosinophils</u>	(%)	1-4*
Basophils	(%)	0-2			<u>Platelet count</u> <u>Basophils</u>	(%)	0-2

Example 2: Synchronization and Deduplication

A future Scenario: Cross-border Secondary Use of Health Data – Synchronizing and Deduplicating Clinical Records for Research Purposes

A public health research institute in Cyprus is conducting a multinational study on adverse drug reactions (ADRs) associated with a newly approved antihypertensive medication. The research project seeks to identify rare side effects that may not have been evident during clinical trials. To build a robust dataset, the Cypriot Health Data Access Body (HDAB) submits a formal request for health data to the corresponding HDABs in several EU Member States, including Germany, under the framework of the European Health Data Space (EHDS). To support the study, the German HDAB extracts data and provides access to pseudonymised health records from patients who have been prescribed the medication in question. These records include structured clinical data such as medication codes, recorded side effects, lab test results, and relevant diagnoses. The information is provided in internationally accepted formats, for example HL7 FHIR, and accompanied by metadata that records the source institution, data entry timestamp, clinical coding system used and the healthcare professional who originally validated the data.

Once the data is securely transmitted, it is integrated into the Cypriot research environment. The integration process is governed by the principles of semantic interoperability and relies on automated services to translate, map, and align different clinical terminologies. For example, while German data contributors use SNOMED CT for coding symptoms, Cypriot systems may rely on subsets of ICD-10 for similar clinical entries. Without semantic alignment, this would lead to misclassification or duplication of data entries. A recurring problem observed is the duplication of adverse event records: a symptom such as “swelling in the lower limbs” might appear under different codes depending on the local documentation protocol, leading the analytics engine to interpret the same symptom as two distinct events.

To address this, semantic interoperability services are deployed to reconcile terminology differences, tools such as OMOP CDM can be useful here. Mapping dictionaries and translation algorithms help standardize the codes, ensuring that medically equivalent concepts are merged appropriately. In one instance in a patient previously visited doctors in both countries, an adverse drug reaction coded as "Oedema of lower extremities" in SNOMED CT and as "localized swelling" in ICD-10 was flagged during the deduplication process. Although they were recorded in different countries and coded differently, the semantic engine correctly identified them as referring to the same clinical outcome, thereby preventing unnecessary duplication in the dataset.

Deduplication is not only performed at the concept level but also involves record-level comparison. If two entries show identical values and temporal alignment, the system considers the older record as superseded or archived. If slight discrepancies exist, such as a follow-up measurement differing from the original result, the entry is preserved as a separate data point, since it may reflect clinical progression. Throughout the process, HDABs ensure that all data transfers and uses respect GDPR principles, including data minimization and lawful secondary use. Additionally, patients' rights are protected via national opt-out systems and project-level data governance measures. The use of technologies such as FHIR resources for data structuring, master patient index systems for linking records and metadata registries for source tracking allows the research team to build a harmonized, high-quality dataset (information models, ontologies or tools such as OpenEHR, for registering in the EHR—are also used for data structuring). This synchronization and deduplication framework not only ensures the accuracy and consistency of the data but also improves the reliability of research outcomes, allowing policymakers to make informed decisions regarding drug safety.

The success of this cross-border data sharing relies on close coordination among HDABs. Their role is essential in supporting semantic, technical, organizational, and legal alignment, from applying standard terminologies and interoperability protocols to coordinating workflows and upholding compliance with GDPR and the EHDS framework. These layers of alignment are vital to achieving high-quality, secure, and lawful secondary use of health data across Europe.