DATASPACE 4HEALTH
LUXEMBOURG

Data Flow Diagrams for Oncology

Document ID: DS4H_WP5

Hee Kim (LIH), Mary Roszel (LIH), Maximilian Fünfgeld (LIH), Reka Toth (LIH), Casper Arij Van Den Berg (HRS), Nour-El-Houda Moussaoui (HRS)

Luxembourg Institute of Health (LIH); Robert Schuman Hospitals (HRS)

Date: 10/06/2026

Funding

This project has received funding from the Ministry of the Economy of Grand Duchy of Luxembourg under Grant agreement no 20230505RDI170010392869.

Disclaimer

The documents published by the consortium are intended solely for informational purposes and reflect the views and opinions of the consortium members at the time of publication and within the scope of the Dataspace4Health project (DS4H).

The ideas expressed herein do not necessarily reflect the official policy or position of the funding entity.

Efforts have been made to achieve the relevance of the content; however, the consortium does not make any representation regarding its completeness and accuracy.

This document may be subject to further revision.

This document is intended to be published on the Dataspace4Health website.

Table of Versions

Version n°	Issue Date	Reason for change
0.0	15/03/2025	First draft.
0.1	31/03/2025	LNS is replaced with Luxgen and corresponding comments are made. Dataflow of research outcome to the hospital is removed because it is out of scope. The 0.1 version is shared with all partners.
0.2	06/05/2025	Dataflow back to the HRS has been restored, even though it is not the scope of this project. It is included to show the potential and to assure that the infrastructure is ready for the molecular tumor board. In addition, entity names and corresponding figures are revised.
0.3	26/09/2025	A major revision has been conducted across the text and figures. The key changes include distinguishing the medical data flow from the EHDS governance flow, adding the current data flow state to the "As-Is" diagram, and clarifying the role of the LNDS. Furthermore, future work related to the Molecular Tumor Board is now depicted and explained, while various errors have been corrected and unrelated elements (e.g., LuxGen, HealthNet, EMR) have been removed.
0.4	08/01/2026	IBBL has been removed. Pseudonymisation by LNDS remains optional. The terms “treatment purpose” has been refined as “primary purpose”.
0.5	06/03/2026	Data flow is decoupled into MTB and Research purposes. The PoC data flow is introduced. Redundancies have been eliminated, with common descriptions are noted in Section 2 for reference.
0.6	16/04/2026	The As-Is diagram is decoupled into use cases, with updated color scheme labels. For the MTB case, the SPE has been deplicated, and the deployment location remains unspecified.
0.7	01/06/2026	The LNDS role has been fixed, 'SPE' is now corrected to 'SPE Provider,' and the figure labels have been revised.
1.0	10/06/2026	Final version has been sent to the coordinator.

TABLE OF CONTENTS

TABLE OF FIGURES

Glossary

Abbreviation	Expression
CIViC	Clinical Interpretation of Variants in Cancer
CRC	Colorectal cancer
DC	Data Connector
DFD	Data Flow Diagram
EHDS	European Health Data Space
EHR	Electronic health record
ETL	Extract, Transform, Load
HDAB	Health Data Access Body
HRS	Hôpitaux Robert Schuman
INC	Institut National du Cancer du Luxembourg (National Cancer Institute of Luxembourg)
LDFD	Logical Data Flow Diagram
LIH	Luxembourg Institute of Health
LNDS	Luxembourg National Data Service
LNS	Laboratoire National de Santé (Luxembourg National Health Laboratory)
MTB	Molecular Tumor Board
NGS	Next-Generation Sequencing
PDFD	Physical Data Flow Diagram
PII	Personally identifiable information
PoC	Proof of Concept
SPE	Secure Processing Environment
TBD	To Be Defined
VEP	Variant Effect Predictor

1. INTRODUCTION

Data flow for oncology research and the molecular tumor board (MTB) pose challenges due to the dynamic nature of actors, participants, and the complexity of data. Specifically, health data flow in Luxembourg is disconnected by nature since healthcare providers and healthcare research institutes are different legal entities. The data Flow Diagram (DFD) [1] is a useful tool to understand the dynamics of data flow and to plan for data pipeline implementation. It provides a visual representation of data movement, which enables stakeholders to understand and refine workflows with clarity and precision. In particular, the DFDs presented in this document enhance the understanding of data flows for oncology cases in Luxembourg across systems and institutions.

Another objective is to illustrate data flow between data holders and data users while ensuring security, pseudonymisation, and compliance with data governance and regulations in the European Health Data Space (EHDS) [2]. EHDS technical components will be utilized for data sharing within the Secure Processing Environment (SPE) [3]. Each participant is given their own software called a data connector, while data access and data processing within the SPE are fully governed by the data holder and the Health Data Access Body (HDAB) [4], which does not yet exist at the time of this project. Note that the integration of EHDS components depends on the data use. For treatment purposes, such as the MTB case, the use of the SPE is not mandatory. This document distinguishes two oncology cases: the MTB case for treatment purpose and the oncology research case. The following participants represent the roles and their respective responsibilities involved in operating these use cases in Luxembourg.

1. **Bioinformatician (LIH):** Responsible for advanced genomic data analysis, including variant annotation, the identification of potential drug targets, and other tasks.
2. **Molecular Biologist (IBBL):** Performs genomic sequencing (e.g., next-generation sequencing, NGS) and generates sequencing results in an unstructured text format for further analysis, primarily supporting research and translational use cases.
3. **Molecular Biologists, Pathologists, and Geneticists (LNS):** Operate in a clinical diagnostic context, with a primary focus on molecular diagnostics. They are responsible for generating, validating, and clinically interpreting molecular and pathological data, supporting patient care.
4. **Physician and Case Manager (HRS):** Collect and manage oncological patient data, including medical history, diagnosis, and treatment plans, using the electronic health record (EHR) system.
5. **Trusted Third Party (LNDS)**: Responsible for developing and providing technical mediators, such as an SPE and a pseudonymisation tool [5]. It is required to replace personally identifiable information (PII) with pseudonyms before sharing clinical data for research purposes, for which the service provided by LNDS can optionally be utilized.
6. **MTB Coordinator (INC)**: The Institut National du Cancer du Luxembourg (INC) is the national cancer institute in Luxembourg, playing a central role in organizing and coordinating national tumor boards, including the Molecular Tumor Board (MTB).

2. DATA FLOW DIAGRAM KEY COMPONENTS

Entities, processes, and data stores are three key components of DFD. These elements help visualise how data moves between systems, how it is processed, and where it is stored. Therefore, it is essential to clarify them in detail.

2.1. ENTITIES

An entity represents a source or destination of data that interacts with the system. Entities can be people, organizations, systems, or devices that provide data input to the system or receive data output from it. The EHDS regulates health data usage by distinguishing between primary use (direct care) and secondary use (research/innovation).

- **HRS:** Hôpitaux Robert Schuman (HRS) receives and treats patients, including oncological patients. By doing so, HRS collects and manages patient data, including medical history, diagnosis and treatment plans. For research purposes, patient data are pseudonymised (or anonymized, if applicable) before sharing the data with research institutes, with the aim of ultimately improving patient care.
- **LNS/IBBL:** Laboratoire National de Santé (LNS) is the public health laboratory providing diagnostic genetic counselling and the diagnosis of genetic diseases, while the Integrated Biobank of Luxembourg (IBBL) is a specialized biobank providing a solution for all stages of the biospecimen journey including collection, processing, storage, and distribution. Both the LNS and the IBBL are able to perform next-generation sequencing (NGS) and generate sequencing results.
- **SPE Provider:** Secure Processing Environment (SPE) is one of three technical components of the EHDS [14], providing privacy-preserving health data analysis. SPE can be developed and provided by the EU, HDAB, data holder, or Trusted Third Party provider according to the regulations [7]. Note that the use of SPE is mandated for the EHDS2 for secondary use (research) only. In this project, SPE is used for the research purpose.
- **LIH:** The Luxembourg Institute of Health (LIH) is a biomedical research organization focusing on precision health to improve patient outcomes. The LIH conducts advanced genomic data analysis, including variant annotation and the identification of potential drug targets.
- **LNDS:** The Luxembourg National Data Service (LNDS) is a public agency designed to implement Luxembourg's data strategies. The LNDS facilitates the secure secondary use of data for research and innovation by providing technical solutions like pseudonymisation and the SPE.

2.2. PROCESSES

Processes refer to any transformations applied to the data, showing how it is changed or used within the system. A process manipulates, calculates, or transforms data by taking input data, applying logic, and producing output data. For the oncology research, the core processes (e.g., data integration and advanced genomic data analysis) could be executed in the SPE.

- **Extract Transform Load (ETL):** Structures patient data to ensure it is fit for purpose. For instance, the ETL process in the SPE integrates data from multiple data holders, and then prepares it for a target use case.
- **Transfer data**: Data are transferred from data holders to data users for research purposes. For primary purposes such as the MTB, however, the processed results (i.e., treatment recommendations) must be transferred back to the caregiver (HRS).
- **Genomic sequencing:** Performs the NGS workflow that begins with sequencing to read DNA fragments, followed by alignment to map these fragments to a reference genome. Finally, variant calling identifies any differences.
- **Genomic Data Analysis:** The LIH conducts genomic data analysis in three steps. First, data inputs from multiple data holders are integrated using either a patient identifier or a combination of a pseudonymised patient identifier and a project identifier. Then Variant Effect Predictor (VEP) [10] is conducted to annotate newly discovered variants, and finally researchers interpret what a specific variant means for a patient, such as which drugs might work or what the prognosis might be.
- **Molecular Tumor Board:** According to the Deliverable 6.2 [13], the current MTB is organized by INC. It is manual and ad-hoc, making it slow and inefficient. Patient data are presented via video conference using screen sharing. A comprehensive system integrating complex patient information could be introduced in the future, however, this is outside the scope of the Dataspace4Health project.
- **Pseudonymisation:** Replaces PII with artificial identifiers, known as pseudonyms. For instance, for research purposes, HRS can optionally use the pseudonymisation service provided by the LNDS for concealing patient identifiers, kit identifiers, and any PII.

2.3. DATA STORES

Data stores are repositories where data are maintained and managed. Operating a seamless data flow between storage and processing steps is vital for the effective functioning of the project.

- **EHR:** Centralizes different clinical, administrative and diagnostic data, enabling healthcare professionals to access and manage patient information. The EHR allows for structured data extraction such as in OncoBox format.
- **Scality RING:** is an object storage solution to store massive amounts of unstructured data, including sequencing data and variant results. It provides a scalable and robust solution for managing the ever-growing volume of genomic data [12].
- **Knowledgebase**: Designed to facilitate the clinical interpretation of genetic variants in cancer by linking genomic alterations to evidence-based therapeutic, prognostic, and diagnostic outcomes. The clinical Interpretation of Variants in Cancer (CIViC) [11] can be utilized as an open-access and community-curated knowledgebase.

- **Annotation Database:** Stores genetic variants and associated information such as drug targets. Unlike a file system, it enables researchers to query and retrieve relevant information quickly leading to efficient genomic data analysis.
- **Repository for Integrated Data:** Temporarily stores data inputs from data holders in a secure and remote computing environment.

3. DATA FLOW DIAGRAM

The data flow diagram illustrates the movement of medical data between data holders and data users for advanced sequencing analysis. Each actor has specific tasks and responsibilities in handling data. Figures depicted in this section utilize different color codes to distinguish types of data flows. Data flows highlighted in purple represent the treatment purpose (MTB), while those in blue indicate the research purpose.

1. **Data flow to LNS/IBBL:** HRS transfers biosamples and associated metadata to the LNS for primary use and treatment purposes, while likewise transferring them to the IBBL for research purposes. The biosamples are physically transported and curated in the biobank at each institute.
2. **Data flow to SPE:** Data from multiple data holders (i.e., HRS, LNS, and IBBL) are integrated in the SPE, where advanced genomic analysis is conducted for research purposes. The data user (LIH) attains access to and processes data in the SPE.
3. **Data flow to LIH:** For research purposes, raw data never flow to the LIH (data user), instead, only research results flows to the LIH with approval by the HDAB. In the MTB case, the LIH receives clinical context directly from the HRS rather than through the MTB system. Introducing a middle actor would only increase the risk of errors.
4. **Data flow to HRS:** Unlike the data flow to the LIH, results for treatment purposes (i.e., treatment recommendations through the MTB) flow back to the HRS without approval by the HDAB. On the other hand, the LNS delivers sequencing results (Genetic test reports) to the HRS.
5. **Data flow to LNDS:** Clinical data for research purposes must be pseudonymised before transferring the data. The LNDS offers a pseudonymisation service enabling the data user to link pseudonymised data from different data holders.

3.1. CURRENT DATA FLOW

The current status of the level-0 logical data flow is depicted in Figure 1 as the “As-Is Diagram”. A data pipeline does not exist for the MTB. Data for the MTB are presented via video conference with screen sharing. The INC organizes and coordinates national tumor boards, including the MTB. From 2019 to 2024, 34 MTB sessions were organized in Luxembourg [6]. This number indicates that an MTB session in Luxembourg is held approximately every 7-8 weeks, a frequency that could be improved in the future. On the other hand, data sharing with research institutes (LIH) happens exclusively for specific research projects only. Pseudonymised data are shared in an ad hoc manner.

Numerous challenges are presented in the current data flow. The LIH is currently not involved in the MTB for providing bioinformatics consultation. Moreover, there is no comprehensive system integrating cancer patient data. To address these challenges, Dataspace4Health develops a technical means as a Proof of Concept (PoC) to pave a path for precision oncology. Data flow diagrams of two use cases are also proposed in Figures 1-4, however, implementing these use cases is beyond the scope of the Dataspace4Health project.

3.2. LOGICAL DATA FLOW DIAGRAM (LDFD)

A Logical Data Flow Diagram (LDFD) focuses on what a system does, not on how it is implemented. It is a high-level overview that illustrates the data flow between entities. The purpose of an LDFD is to identify the business requirements, independent of the technology used. This helps stakeholders to agree on the core logic and functionalities of a system without understanding the technical details.

3.2.1. LEVEL-0 LDFD

The Level-0 Logical Data Flow Diagram, also known as the Context Diagram, provides the highest-level view of the entire system. At this level, processes are not shown. This diagram's main goal is to identify the boundaries of each entity and how data flows between entities. It is known as a tool for defining the project scope and ensuring everyone understands the big picture before diving into more detailed views.

Two use cases of precision oncology are depicted in the bottom layer of Figure 1. Patient data are initially collected at the hospital site (HRS), and then transferred to other entities for either clinical treatment purpose or research purpose. Advanced genomic analysis is subsequently conducted by the LIH. The data flow for treatment purposes (MTB) allows the sharing of personally identifiable information (PII), whereas the research purpose data flow must pseudonymised PII data. Optionally, the pseudonymisation service provided by the LNDS can be used, for which PII is de-identified on a per-project basis. This allows for integrating data in the SPE using a combination of project ID and pseudonyms. In the research data flow in the context of the EHDS, the research outcomes are only transmitted to the data user (LIH) following formal approval from the HDAB.

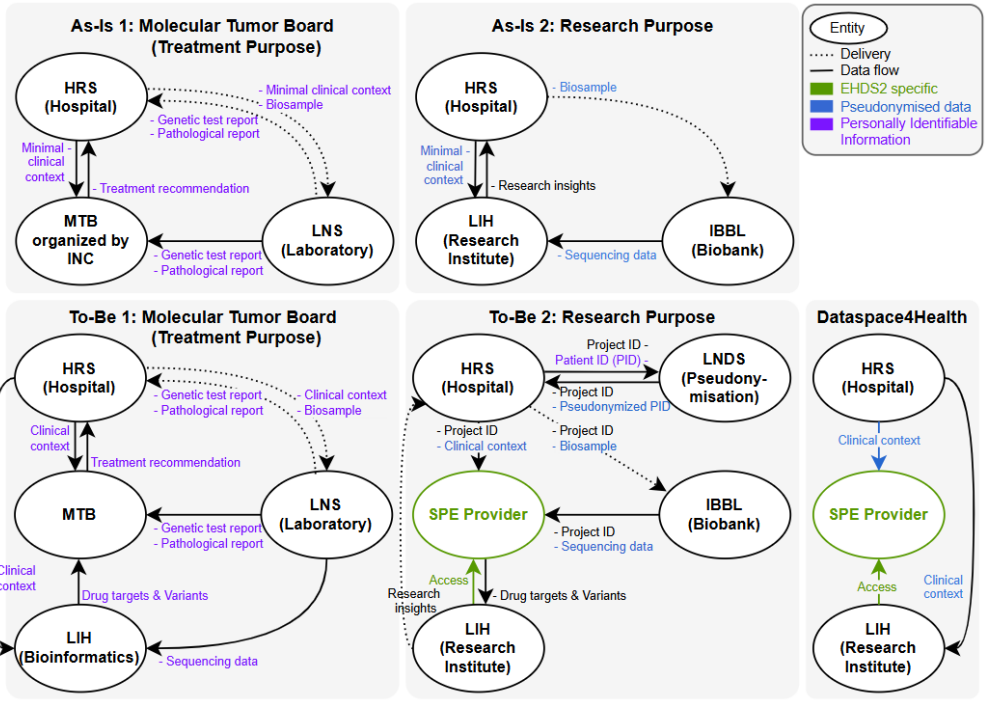

Figure 1. Level-0 Logical Data Flow Diagram.

3.2.2. LEVEL-1 LDFD

The Level-1 Logical Data Flow Diagram breaks down the entity in the Level-0 LDFD into granular subprocesses. It illustrates how data flows move not only between these entities identified in the Level-0 diagram but also how data are transformed internally in each entity. Figure 2 illustrates the data flow between various processes, represented by rectangular icons, while entities is indicated by light blue boxes. As MTB is currently organized by the INC in Luxembourg, the entity for hosting the MTB system could be INC. In the absence of a HealthNet endpoint at the INC, the LIH could be an alternative option for hosting the MTB system. The detailed descriptions of these processes have been previously provided in Section 2.2. This level provides more detailed understanding of the major functions in each entity and how they interact between one another.

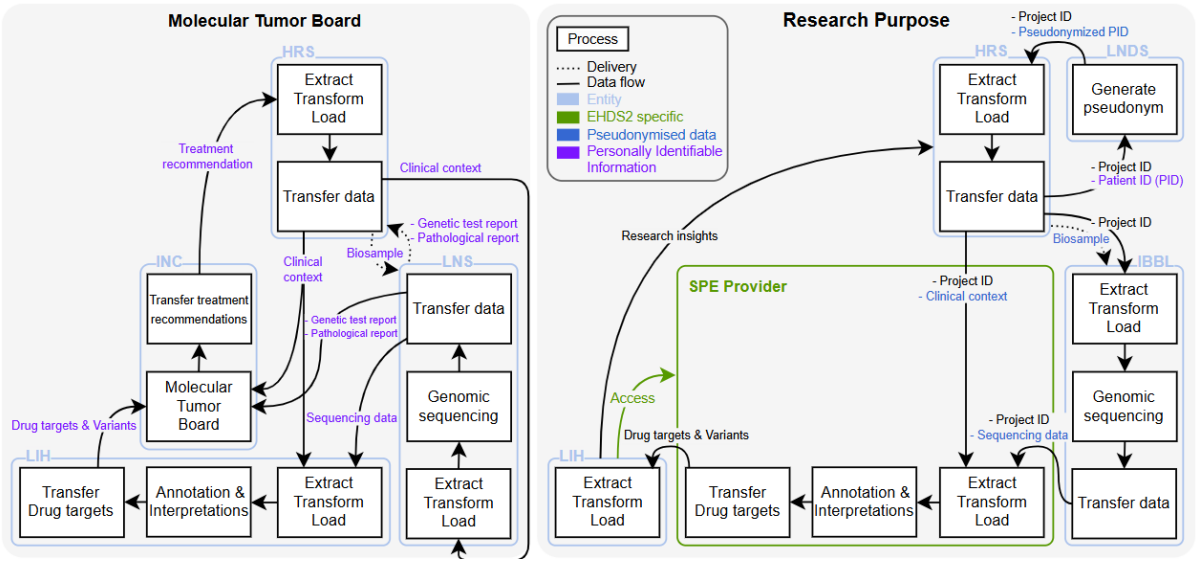

Figure 2. Level-1 Logical Data Flow Diagram

3.3. PHYSICAL DATA FLOW DIAGRAM (PDFD)

The Physical Data Flow Diagram (PDFD) adds technical specifications that were left out of the LDFD and focuses on how databases are actually integrated. It shows the specific hardware involved, demonstrating the technology and infrastructure that support the data flow. Together, the LDFD and the PDFD provide a concrete and comprehensive understanding of data flow from the data sources to the data destinations.

3.3.1. LEVEL-0 PDFD

Similar to the LDFD, the Level-0 PDFD provides an overview of the physical storage systems. It serves as a bridge between the conceptual requirements and the technical implementation. Internal processes are not yet visible at Level-0, nonetheless, the diagram shows how data moves between entities and their databases. The detailed descriptions of these databases have been previously provided in Section 2.3. Figure 3 illustrates the data flow of patient data and analysis results among all associated entities and data stores. Each entity manages specific data in a designated database, and the data moves sequentially from the EHR database at the HRS to support for oncology care in MTB context and to support research activities.

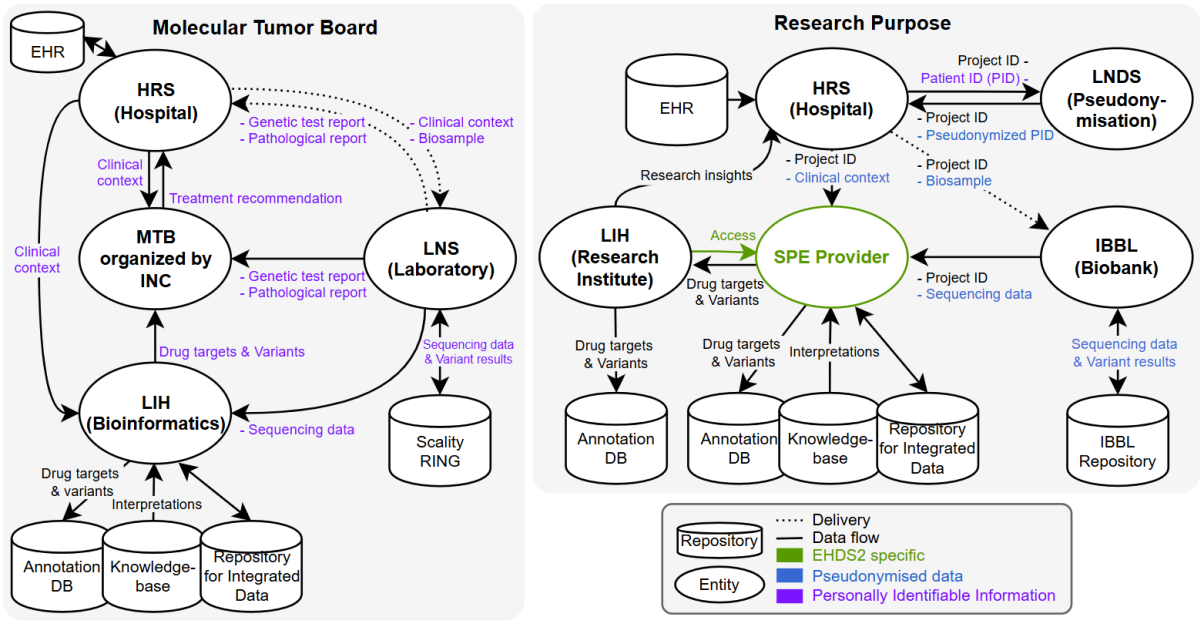

Figure 3. Level-0 Physical Data Flow Diagram.

3.2.2. LEVEL-1 PDFD

The Level-1 PDFD is the most detailed representation of the storage systems and the major processes. Each entity from Level-0 PDFD is unfolded into specific processes and data flows among these processes and storage systems are depicted. Figure 4 shows how data moves, detailing which inputs enter a process and where the processed outputs are stored in specific databases. The detailed descriptions of these processes and repositories have been previously provided in Section 2.2 and Section 2.3. This diagram is a useful blueprint for technical teams, providing the necessary details for system design and implementation.

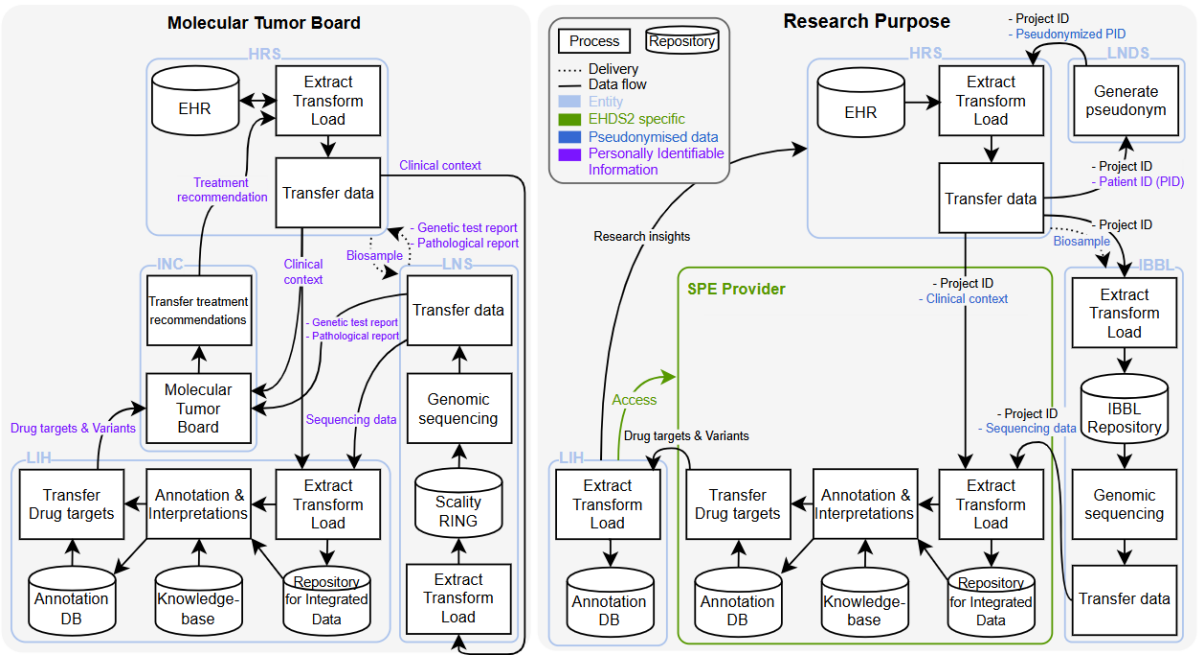

Figure 4. Level-1 Physical Data Flow Diagram.

4. CONCLUSION

Data Flow Diagrams in this document demystify the data movements for precision oncology with two use cases in Luxembourg. By visually representing the flow of clinical and genomic data through the logical and physical layers of the system, we have established a shared understanding of the processes and technical infrastructure required.

We believe that this document not only clarifies how data flows through entities and is stored in dedicated storage systems, but also envisions the work needed to bridge the gap between current manual, ad-hoc workflows and a digitalized approach. We hope that it serves as a roadmap for the continued digitalization of oncology workflows in Luxembourg, strengthening the MTB as well as accelerating research, and thus ultimately improving patient care.

REFERENCES

1. Data-flow diagram. Retrieved April 1, 2025, from https://en.wikipedia.org/wiki/Data-flow_diagram
2. European Health Data Space Regulation (EHDS). (2025). REGULATION (EU) 2025/327 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 11 February 2025 on the European Health Data Space and amending Directive 2011/24/EU and Regulation (EU) 2024/2847. Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202500327
3. Secure Processing Environment. Retrieved April 1, 2025, from https://www.lnds.lu/secure-processing-environment/
4. Health Data Access Bodies – Community of Practice. Retrieved April 1, 2025, from https://health.ec.europa.eu/ehealth-digital-health-and-care/ehds-action/projects-supporting-ehds/health-data-access-bodies-community-practice_en
5. Identifier-matching and Pseudonym Management Service (IPMS) by INCERT and LNDS. Retrieved April 1, 2025, from https://howto.lnds.lu/cards/identifier_matching_and_pseudonym_management_service/
6. INC activities. Retrieved September 23, 2025, from https://institutcancer.lu/en/nos-activites
7. EHDS Regulation (EU) 2025/327, Recital 77.
8. XML-OncoBox. Retrieved April 1, 2025, from https://www.xml-oncobox.de/de/Home/FuerZentren
9. OnkoZert by the German Cancer Society (DKG). Retrieved April 1, 2025, from https://www.onkozert.de/en/
10. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. The ensembl variant effect predictor. Genome biology. 2016 Dec;17:1-4.
11. Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC, Danos AM, Ainscough BJ, Ramirez CA, Rieke DT, Kujan L, Barnell EK. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nature genetics. 2017 Feb;49(2):170-4.
12. Scality's Object Storage at the National Health Laboratory. Retrieved September 16, 2025, from https://itnation.lu/news/le-stockage-objet-de-scality-au-laboratoire-national-de-sante/
13. Dataspace4Health Deliverable 6.2 Functional Processes -Research PoC. Dataspace4Health February. 2026.
14. Dataspace4Health Deliverable 5.3 A Technical Reference Architectureto Interconnect the Care andResearch Dataspaces. Dataspace4Health July. 2025