health RI
enabling data driven health & life sciences

Layers, pipes and patterns: detailing the concept of data stations as a foundational building block for federated data systems in healthcare

Abstract We describe …

1 Towards federated health data platforms for secondary use

The ambition for a seamlessly connected digital healthcare ecosystem, capable of leveraging vast quantities of patient data remains illusive. Designing and implementing health data platforms is notoriously difficult, given the heterogeneity and complexity of such systems. As a starting point - and to frame the scope of this paper - consider the trade-offs along the following two design criteria.

1.1 Primary vs. secondary health data sharing

First, we distinguish health data platform for primary or secondary health availability [1]. It is well known data systems have different design an performance characteristics depending whether they are built for online transactional processing (OLTP) or online analytical processing (OLAP), as summarized in the Table 1 below (taken from M. Kleppmann and C. Riccomini [2]).

Table 1: Distinguishing characteristics between transactional and analytical systems

Property	Transaction processing systems (OLTP)	Analytic systems (OLAP)
Main read pattern	Small number of records per query, fetched by key	Aggregate over large number of records
Main write pattern	Random-access, low-latency writes from user input	Bulk import (ETL) or event stream
Data modeling	Predefined	Defined post-hoc, either schema-on-read or schema-on-write
Primarily used by	End user/customer, via web application	Internal analyst, for decision support
What data represents	Latest state of data (current point in time)	History of events that happened over time
Dataset size	Gigabytes to terabytes	Terabytes to petabytes

Current efforts to design and implement the European Health Data Space (EHDS) in fact aims to support primary and secondary use in one go [3]. This Herculean endeavour has spawned many initiatives to develop a coherent architecture and support implementation across Europe that ultimately should lead to interoperability in the broadest sense of the word, most notably:

- The Data Space Blueprint v2.0 (DSB2) by the Data Spaces Support Centre ([4]) that serves as a vital guide for organizations building and participating in data spaces.
- The Simpl Programme ([5]) that aims to develop an open source, secure middleware that supports data access and interoperability in European data initiatives. It provides multiple compatible components, free to use, that adhere to a common standard of data quality and data sharing.
- TEHDAS2 ([6]), a joint action that prepares the ground for the harmonised implementation of the secondary use of health data in the EHDS.

We believe, however, that in order to successfully design and implement health data spaces, more detailed analysis and solution patterns are required that distinguish between primary (OLTP) and secondary (OLAP) data use. Although functional components can be shared between these two, it is a matter of the devil being in the details. Hence one of the objectives of this paper is to detail an open, technology agnostic architecture for secondary use, to complement existing efforts and guide the development in the field. Although this paper is mostly informed within the context of the Dutch healthcare sector, the solutions we propose are written with the wider European context in mind.

1.2 Centralized vs. decentralized platforms: what constitutes a federated data system?

A second design criterium pertains to the choice of single-node (centralized) or distributed (decentralized) platforms, which are not only be driven by technical considerations (scalability, elasticity, fault tolerance, latency) but are also strongly dependent on organizational, legal or regulatory requirements such as data residency. The general approach of EHDS and other data spaces is federative by nature, that is, decentralized. For example, DSB2 stresses the need for interoperability and federative protocols within and across data spaces.

Upon closer inspection, however, specific functional components that are foreseen within the EHDS are best characterized as centralized (sub-)systems. As an example, consider the secure processing environments (SPE) as defined in article 50 of the EHDS. Known examples of such SPEs include data platforms provided by national statistics officces (CBS Microdata environment), healthcare-specific national platforms (Finland's Kapseli platform) and Trusted Research Environments (TREs) within the domain of research (see EOSC-ENTRUST for examples across Europe). Given that healthcare data is often vertically partitioned (data elements of the same subject are scattered across various data holders), SPEs provide the most effective means to (temporarily) share, integrate and analyse such data. Hence many SPEs are best described as centralized systems, and thus we need to take into account that data spaces constiute a hybrid architecture that includes both centralized and decentralized components.

In computer science and engineering, federated data systems (FDS) have emerged as a new paradigm. Recent technological inventions offer important new enablers to to implement FDS:

- Capabilities of edge computing and single-node computing has increased significantly whereby it is now possible to process up to 1 TB of tabular data on a single node thereby enabling large volumes of data processing to be done efficiently on a single data station [7], [8]
- Federated machine learning (or federated learning in short) has matured as a means for training of predictive models, most notable through weights sharing of deep learning networks [9]
- Privacy-enhancing technologies (PETs) such as secure multi-party computation (MPC) significantly improve secure processing across a network of participants and are now sufficiently mature to be used on an industrial scale [10], [11]
- The composable data stack provides a way to unbundle the venerable relational database into loosely components, thereby making it easier and more practical to implement FDS using cloud-based components with microservices, thereby opening up a transition path towards more modular and robust architectures for FDS [12], [13].

The architectural shift from centralized to federated data systems is not merely a technical evolution. Modern approaches to data governance are undergoing a paradigm shift towards federated solutions. As an example, the concept of a data mesh is increasingly being adopted at large corporations. From the perspective of sovereignty and solidarity, we believe that a commons-based, federated approach has distinct benefits in moving towards a more equitable, open digital infrastructure [14].

The ongoing paradigm shift is not without challenges. As noted by Perreira et al. (2023):

> The requirement for specialization in data management systems has evolved faster than our software development practices. After decades of organic growth, this situation has created a siloed landscape composed of hundreds of products developed and maintained as monoliths, with limited reuse between systems. This fragmentation has resulted in developers often reinventing the wheel, increased maintenance costs, and slowed down innovation. It has also affected the end users, who are often required to learn the idiosyncrasies of dozens of incompatible SQL and non-SQL API dialects, and settle for systems with incomplete functionality and inconsistent semantics.

Also, the notion of what constitutes a federated data system needs to be defined more specifically if we are to see the forest for the trees between different instantiations of the same concept. 'Federation' can mean any of the following solution patterns:

- Data federation addresses the problem of uniformly accessing multiple, possibly heterogeneous data sources, by mapping them into a unified schema, such as an RDF(S)/ OWL ontology or a relational schema, and by supporting the execution of queries, like SPARQL or SQL queries, over that unified schema [15];
- Federation within the concext of a Personal Health Train (PHT) refers to the concept by data processing is brought to the (personal health) data rather than the other way around, allowing (private) data accessed to be controlled, and to observe ethical and legal concerns [16];
- Data sharing within a network of research organizations in a TRE, with different types of federations services (localization, access);
- Federation services as defined in the DSSC Blueprint 2.0 pertain to the support the interplay of participants in a data space, operating in accordance to the policies and rules specified in the Rulebook by the data space authority.

What then, is a viable development path out of this creative chaos?

1.3 Data stations as a foundational building block for federated data systems

Inspired by previous calls to action to move towards open architectures for health data systems [17], [18] and the notion of the hourglass model [18], [19], [20], we hypothesize that the concept of a 'data station' can be used as a foundational building block for federated data systems. A data station should provide a set of minimal standards (at the waist of the hourglass), thereby maximizing the freedom to operate between data providers and data consumers within the context of a health data space. Note that this approach has many similarities with the FAIR Hourglass as proposed by E. Schultes [20]. Our approach of data stations, however, aims to complement these approach with focus on secondary data use of routine collected clinical data and the concept op the PHT.

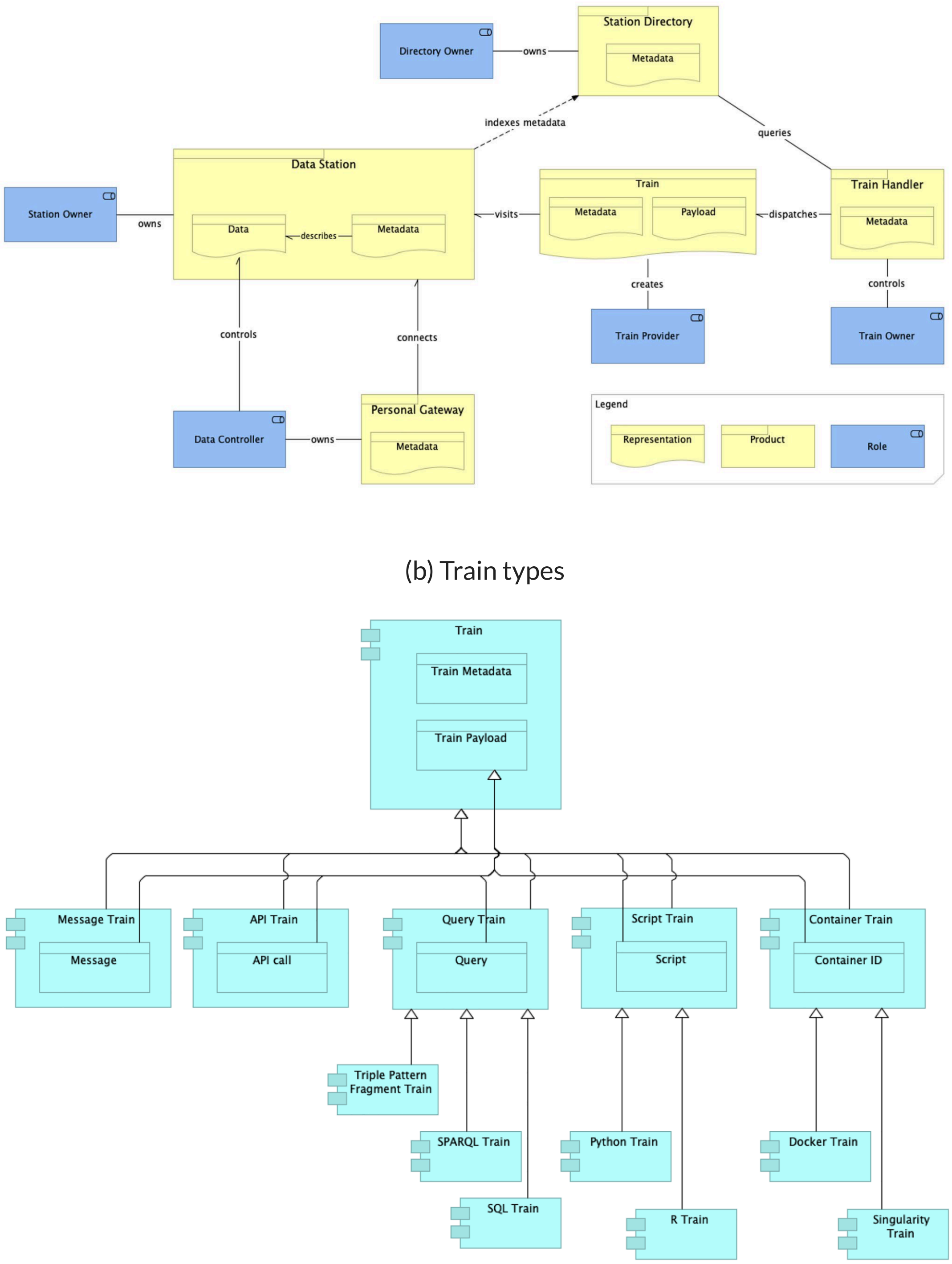
In this paper we take loosely follow an Action Design Research approach [21], [22] to design an open architecture for data station. Our main contributions are:

- Ontology of a data station, that integrates the PHT architecture [16] and the DSSC Blueprint 2.0
- Comparative analysis of existing implementations
- Synthesis of the above into functional and technical description of a data station in Archimate, thereby focusing on two primary patterns [23]:
  - the layers pattern for addressing various aspects of interoperability across the stack
  - the pipes and filters pattern for addressing various solution designs for the extract-transform-load (ETL) mechanisms
- A reference implementation of data stations for federated learning

2 Ontology of a data station

As a starting definition, in the Personal Health Train (PHT) architecture, a Data Station is defined as a software application primarily responsible for making data and its associated metadata available to users under conditions determined by applicable regulations and Data Controllers. The main concepts are shown in Figure 1a, while the various types of trains are shown in Figure 1b.

Figure 1: High-level overview of the Personal Health Train (PHT) architecture



(a) Main roles and components

(b) Train types

Taking the concepts from L. O. Bonino da Silva Santos, L. Ferreira Pires, V. Martinez, J. Moreira, and R. Guizzardi [16], Table 2 maps the PHT architecture to the DSSC Blueprint 2.0 (DSB2). Some mappings are relatively evident. For example, the concept of Data and Metadata as defined in PHT is subsumed in the concept of a Data Product in DSB2. Less evident, is the mapping of the notion of a Train:

> A Train represents the way data consumers interact with the data available in the Data Stations. Trains represent a particular data access request and, therefore, each train carries information about who is responsible for the request, the required data, what will be done with the data, what it expects from the station, etc.

to Value Creation Services in DSB2 that includes data fusion and enrichment, collaborative data analytics and federated learning. More specifics of this mapping will be provided in the Archimate specification.

Table 2: Mapping the key concepts from the PHT architecture [16] to the concepts of the DSSC Blueprint 2.0.

Component PHT	mapping to DSSC Blueprint 2.0 concepts
• Data Station	• Data Space Building Block
• Data • Metadata	• Data Product
• Data Controller	• Data Rights Holder
• Station Controller	• Data Product Provider
• Personal Gateway	• included in Participant Agent Services
• Station Directory	• included in Federation Services • Catalogue provisions and discovers offerings of data and services in a data space
• Directory Owner	• Common Intermediary provides federation services that are common to all participants of the data space
• Train	• Value Creation Services
• Train Provider	• Service Provider
• Train Handler	• specialization of Data Space Component that realises the Train Value Creation Service
• Train Owner	• included in Service Provider as most generic role • concept of Intermediary (specialization of SP) is closer to definition of Train Owner

TO DO:

- Explain that FAIR Data Point specification only contain part of the functionality of a data station, namely just the metadata and catalog.

- Initiative has started to extend the paper by L. O. Bonino da Silva Santos, L. Ferreira Pires, V. Martinez, J. Moreira, and R. Guizzardi [16] into a full specification, named the FAIR Data Train
- One of the key questions of this paper is to detail the ‘data conformity zone’ as defined in the Cumuluz canvas as the functionality through which the data station is populated

3 Comparative analyses of existing data stations

This section provides a detailed comparative analysis using the concepts of the PHT architecture of how the “data station” concept is realized in existing federated data systems:

- the Dutch PLUGIN federated network [24] as an example of data stations that are focused on federated learning [9]
- the Swiss SPHN network [25] as an example of a data station that uses graph databases both for the data and metadata
- the Fair Data Cube [26] as an example of a graph-based data station combined with federated learning
- the Datastation-as-a-Service as defined by the Zorginstituut for federated analytics using privacy-enhancing technologies [27]
- TO DO: data station as defined by Cumuluz

3.1 Data stations in PLUGIN

3.2 Data stations in SPHN

3.3 Data stations in the Fair Data Cube

3.4 Datastation-as-a-Service in KIK-V

4 Synthesis of solution patterns patterns for data stations

We take Klepmann (2017) as our starting point, who states that “Many applications today are *data-intensive*, as opposed to *compute-intensive*. Raw CPU power is rarely a limiting factor for these applications—bigger problems are usually the amount of data, the complexity of data, and the speed at which it is changing.”

Generically, we want:

Reliability	Scalability	Maintainability
tolerating hardware & software vaults	Measuring load & performance	Operability, simplicity & evolvability
human error	Latency percentiles, throughput	

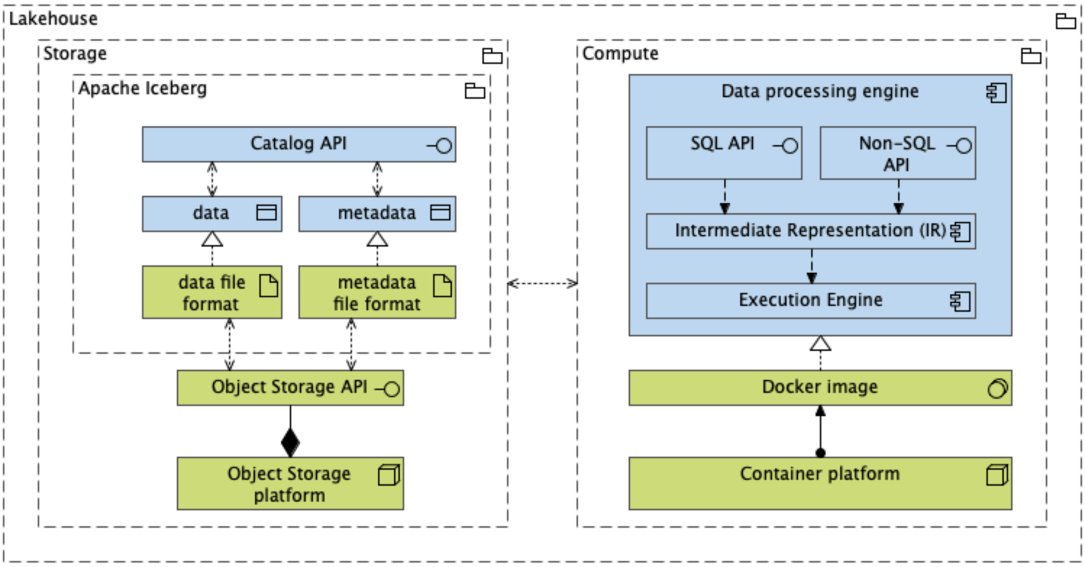
We focus on analytical data systems, with different patterns from transactional data systems.

4.1 Detailing the layers of a data station

TO DO: provide detailed layers, and explain how interoperability works across the layers:

- Storage layer (technology): all reference architectures stipulate use of S3-compliant blob storage
- Data and metadata (application): resides in the data station
- We propose to move towards open table formats, that is, Apache Iceberg, whereby storage and compute can be separated

Figure 2: Solution of a minimal lakehouse that sits at the core of a data station



4.2 Detailing the data conformity zone

TO DO: explain that

- data conformity zone is essentialy a lakehouse pattern
- the architecture of a lakehouse has stabilized and converged towards:
  - **Colum-oriented storage and memory layout:** Apache Arrow ecosystem, including Apache Flight
  - **Late-binding with logical data models most suited for analytics:** ELT pattern with zonal architecture
    - *staging zone:* hard business rules (does incoming data comply to syntactic standard), change data capture
    - *linkage & conformity zone:* concept-oriented tables, typically following a data vault modeling principle, ascertain referential integrity across resources, with tables per concept and linking tables. Mapping to coding systems. Entity resolution for record linkage at the subject level
    - *consumption zone:* convenient standardized views like an event table (patient journey, layout for process mining) with uniformity of dimensions using a star schema

4.3 Detailing the trains

TO DO: explain

- difference between centralized and distributed federated learning (causes lots of confusion)
- basically Train is a generalization of all types of computes
- difference between
  - Train for secondary use, which usually with batch-wise, less strict latency requirements
  - Train for primary use, like API call and messaging, with stricter latency requirements. This also includes deployment of AI for inference

5 Reference implementation with current open source software (OSS) components

5.1 Data stations from the composable data stack

- Python and SQL(-like) languages as the *de facto* standard for analytical processing i.e. the most commonly used analytical scripting languages. Where necessary, using Intermediate Representations (IR), any analytical query can be transpiled to the target engine of choice
- Single-node compute capable of efficiently processing up to 1 TB of data within tens of seconds (polars, DuckDB), so we do away with distributed processing
- Open table formats (Iceberg, Hudi, Delta) and open file formats (parquet, AVRO)

5.2 Container-based trains as the most versatile solution

5.3 Bringing it all together for federated analytics & machine learning (FL)

- Local data stations are conceptualized as serverless lakehouses
  - Local ELT pipelines
  - Decentralized (pre-)processing, including quality control upon ingest
  - ...
- For horizontally partioned data, we can apply FL techniques where only aggregated results are combined centrally
- For vertically partitioned data, we need an intermediate/temporary zone for linking the data
- For both horizontally and vertically partitioned data, we can choose to add PETs, most specifically MPC, as an extra security measure
  - Horizontally partitioned data: one-shot FL
  - Vertically partitioned data:
  - linkage in the blind
  - reversible pseudonimization
- standardized approach to mapppings [28]

6 Parking lot

- Difference with data mesh: mesh of domains, federation is in the same domain. Underlying technology of a data station, however, is functionally identical
- UMCU: CQRS pattern for separately optimizing read/write patterns
- DSSC Blueprint: FL subsumed in value adding services

Table 3 lists known examples of existing health data platform architectures along these two trade-offs.

Table 3: Broad categorization of health data platforms

	primary	secondary
centralized	openHIE [29], Digizorg, Nordics	kapseli, Mayo, …
decentralized	RSO Zuid Limburg, Twiin portaal, …	many federated analytics research networks such as x-omics programme and EUCAIM

Bibliography

[1] F. Cascini, A. Pantovic, Y. A. Al-Ajlouni, V. Puleo, L. De Maio, and W. Ricciardi, “Health Data Sharing Attitudes towards Primary and Secondary Use of Data: A Systematic Review,” *eClinicalMedicine*, vol. 71, p. 102551, May 2024, doi: 10.1016/j.eclinm.2024.102551.

[2] M. Kleppmann and C. Riccomini, *Designing Data-Intensive Applications, 2nd Edtion (Early Release)*. O'Reilly, 2026.

[3] “European Health Data Space Regulation (EHDS).” Accessed: Jun. 09, 2025. [Online]. Available: https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_en

[4] “Data Spaces Blueprint v2.0.” Accessed: Jun. 09, 2025. [Online]. Available: https://dssc.eu/space/BVE2/1071251457/Data+Spaces+Blueprint+v2.0+-+Home

[5] “Simpl Programme.” Accessed: Jun. 09, 2025. [Online]. Available: https://simpl-programme.ec.europa.eu/

[6] “TEHDAS2.” Accessed: Jun. 09, 2025. [Online]. Available: https://tehdas.eu/

[7] M. Raasveldt and H. Mühleisen, “DuckDB: An Embeddable Analytical Database,” in *Proceedings of the 2019 International Conference on Management of Data*, Amsterdam Netherlands: ACM, Jun. 2019, pp. 1981–1984. doi: 10.1145/3299869.3320212.

[8] F. Nahrstedt, M. Karmouche, K. Bargieł, P. Banijamali, A. Nalini Pradeep Kumar, and I. Malavolta, “An Empirical Study on the Energy Usage and Performance of Pandas and Polars Data Analysis Python Libraries,” in *Proceedings of the 28th International*

Conference on Evaluation and Assessment in Software Engineering, Salerno Italy: ACM, Jun. 2024, pp. 58–68. doi: 10.1145/3661167.3661203.

[9] N. Rieke *et al.*, “The Future of Digital Health with Federated Learning,” *npj Digital Medicine*, vol. 3, no. 1, pp. 1–7, Sep. 2020, doi: 10.1038/s41746-020-00323-1.

[10] “The PET Guide,” 2023. Accessed: Jan. 22, 2025. [Online]. Available: https://unstats.un.org/bigdata/task-teams/privacy/guide/

[11] “From Privacy to Partnership,” Jan. 2023.

[12] P. Pedreira *et al.*, “The Composable Data Management System Manifesto,” *Proceedings of the VLDB Endowment*, vol. 16, no. 10, pp. 2679–2685, Jun. 2023, doi: 10.14778/3603581.3603604.

[13] “The Composable Codex.” Accessed: Oct. 16, 2024. [Online]. Available: https://voltrondata.com/codex.html

[14] J. Krewer and Z. Warso, “Digital Commons as Providers of Public Digital Infrastructures,” Nov. 2024. Accessed: Jun. 15, 2025. [Online]. Available: https://openfuture.eu/publication/digital-commons-as-providers-of-public-digital-infrastructures

[15] Z. Gu *et al.*, “A Systematic Overview of Data Federation Systems,” *Semantic Web*, vol. 15, no. 1, pp. 107–165, Dec. 2022, doi: 10.3233/SW-223201.

[16] L. O. Bonino da Silva Santos, L. Ferreira Pires, V. Martinez, J. Moreira, and R. Guizzardi, “Personal Health Train Architecture with Dynamic Cloud Staging,” *SN Computer Science*, vol. 4, Oct. 2022, doi: 10.1007/s42979-022-01422-4.

[17] G. L. Mehl *et al.*, “A Full-STAC Remedy for Global Digital Health Transformation: Open Standards, Technologies, Architectures and Content,” *Oxford Open Digital Health*, vol. 1, p. oqad18, Jan. 2023, doi: 10.1093/oodh/oqad018.

[18] D. Estrin and I. Sim, “Health Care Delivery. Open mHealth Architecture: An Engine for Health Care Innovation,” *Science (New York, N.Y.)*, vol. 330, no. 6005, pp. 759–760, Nov. 2010, doi: 10.1126/science.1196187.

[19] M. Beck, “On the Hourglass Model,” *Communications of the ACM*, vol. 62, no. 7, pp. 48–57, Jun. 2019, doi: 10.1145/3274770.

[20] E. Schultes, “The FAIR Hourglass: A Framework for FAIR Implementation,” *FAIR Connect*, vol. 1, no. 1, pp. 13–17, Jan. 2023, doi: 10.3233/FC-221514.

[21] M. Sein, O. Henfridsson, S. Purao, M. Rossi, and R. Lindgren, “Action Design Research,” *MIS Quarterly*, vol. 35, pp. 37–56, Mar. 2011, doi: 10.2307/23043488.

[22] J. Venable, J. Pries-Heje, and R. Baskerville, “FEDS: A Framework for Evaluation in Design Science Research,” *European Journal of Information Systems*, vol. 25, no. 1, pp. 77–89, Jan. 2016, doi: 10.1057/ejis.2014.36.

[23] F. Buschmann, R. Meunier, H. Rohnert, P. Sommerlad, and M. Stal, *A System of Patterns*, vol. 1. in Pattern-Oriented Software Architecture, vol. 1. Wiley, 1996.

[24] D. Kapitan, F. Heddema, A. Dekker, M. Sieswerda, B.-J. Verhoeff, and M. Berg, “Data Interoperability in Context: The Importance of Open-Source Implementations When Choosing Open Standards,” *Journal of Medical Internet Research*, vol. 27, no. 1, p. e66616, Apr. 2025, doi: 10.2196/66616.

[25] “SPHN - Swiss Personalized Health Network (SPHN).” Accessed: Jun. 09, 2025. [Online]. Available: https://sphn.ch/

[26] X. Liao *et al.*, “FAIR Data Cube, a FAIR Data Infrastructure for Integrated Multi-Omics Data Analysis,” *Journal of Biomedical Semantics*, vol. 15, no. 1, p. 20, Dec. 2024, doi: 10.1186/s13326-024-00321-2.

[27] “KIK-V x GERDA,” Apr. 2024. [Online]. Available: https://populationhealthdata.nl/wp-content/uploads/2024/07/Whitepaper-GERDA-x-KIK-V_-databeschikbaarheid-door-hergebruik.pdf

[28] S. Zhang, R. Cornet, and N. Benis, “Cross-Standard Health Data Harmonization Using Semantics of Data Elements,” *Scientific Data*, vol. 11, no. 1, p. 1407, Dec. 2024, doi: 10.1038/s41597-024-04168-1.

[29] “OpenHIE Framework v5.2-En.” Accessed: Aug. 27, 2024. [Online]. Available: https://ohie.org/