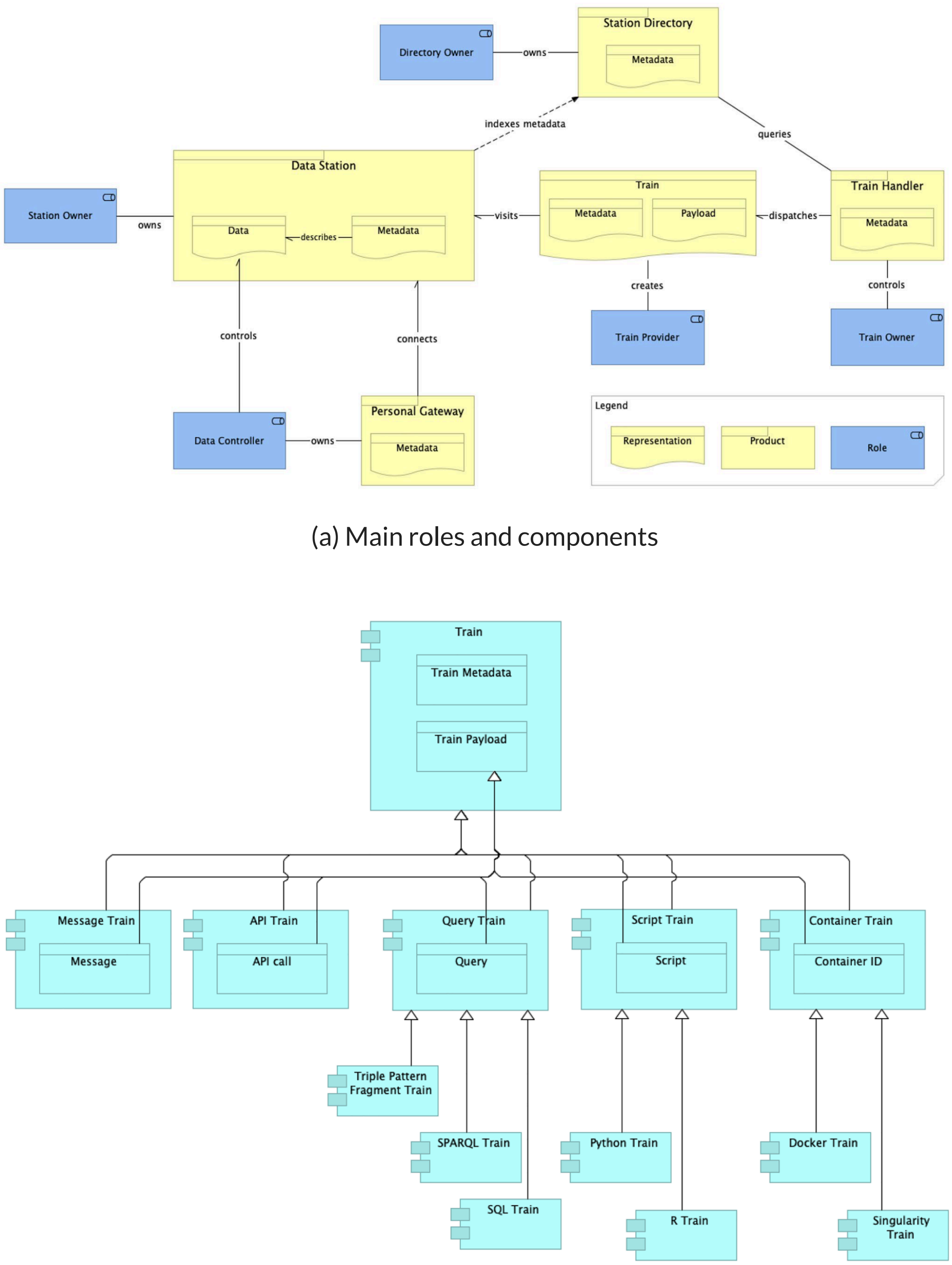
Towards a unified ontology of data stations and data hubs

1 The Personal Health Train architecture

We take the Personal Health Train (PHT) architecture as our starting point [1], wherein 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 Figuur 1a, while the various types of trains are shown in Figuur 1b. A glossary of the concept is provided in Paragraaf 5.



(a) Main roles and components

(b) Train types

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

In their paper Silva Santos LO Bonino da, Ferreira Pires L, Martinez V, Moreira J, Guizzardi R [1] continu to describe more details of the PHT architecture including i) the various functions, services, interface and internal components of the data station; ii) the data visiting process; and iii) the data staging concept in the case data access has been authorized, but the station is not capable of executing the train and needs to stage a capable station with enough resources to run the train. We will consider these details later.

As an aside, it is good to mention that the authors of the PHT architecture have initiated the development of two specifications after publication of this paper, namely:

- FAIR Data Point specification, which covers only the metadata and catalog part of the PHT architecture;
- the FAIR Data Train specification, which covers the full scope of the original paper but at the time of writing is still incomplete.

2 Mapping PHT to the DSSC Blueprint 2.0

To arrive at consistent conceptualization of data stations and trains, Tabel 1 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 that '… 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. We tentatively conclude that it is possible to have a consistent conceptual mapping between, at least at the high level, of the PHT architecture into DSB2. We will return to this matter, when more detailed functions and technical standards are considered for the Archimate specification.

Tabel 1: Mapping the key concepts from the PHT architecture [1] 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 realizes 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

3 The lakehouse architecture as the *de facto* standard for populating data stations

The PHT architecture does not specify *how* the data stations should be populated with data. Also the DSB2 only describes how the 'Data, Services and Offerings descriptions' building block should provide data providers the tools to describe a data product appropriately and completely, that is, tools for metadata creation and management.

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

4 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

Tabel 2 lists known examples of existing health data platform architectures along these two trade-offs.

Tabel 2: Broad categorization of health data platforms

	primary	secondary
centralized	openHIE [2], Digizorg, Nordics	kapseli, Mayo, …
decentralized	RSO Zuid Limburg, Twiin portaal, …	many federated analytics research networks such as x-omics programme and EUCAIM

5 Glossary

- *Data Controller* (Business Role) is the role of controlling rights over data.
- *Data Station* (Business Product) is a software application responsible for making data and their related metadata available to users under the accessibility conditions determined by applicable regulations and the related Data Controllers.
- *Directory Owner* (Business Role) is the role of being responsible for the operation of a Station Directory.
- *Personal Gateway* (Business Product) is a software application responsible for mediating the communication between Data Stations and Data Controllers. The Data Controllers are able to exercise their control over the data available in different Data Stations through the Personal Gateway.
- *Station Directory* (Business Product) is a software application responsible for indexing metadata from the reachable Data Stations, allowing users to search for data available in those stations.
- *Train* (Business Representation) 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.
- *Train Handler* (Business Representation) is a software application that interacts with the Stations Directory on behalf of a client to discover the availability and location of data and sends Trains to Data Stations.
- *Station Owner* (Business Role) is the role of being responsible for the operation of a Data Station.
- *Train Owner* (Business Role) is the role of using a Train Handler to send Trains to Data Stations.

- *Train Provider* (Business Role) is the role of being responsible for the creation of a specific Train, e.g. the developer of a specific analysis algorithm.

Bibliografie

1. Silva Santos LO Bonino da, Ferreira Pires L, Martinez V, Moreira J, Guizzardi R (2022) Personal Health Train Architecture with Dynamic Cloud Staging. SN Computer Science 4. https://doi.org/10.1007/s42979-022-01422-4
2. (2024) OpenHIE Framework v5.2-En. https://ohie.org/. Accessed 27 aug 2024