TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Scharlett Hansson 192542YVEM

Metamodel development in data catalogue using immunisation report dataset

Master's thesis

Supervisor: Janek Metsallik
MSc
Aivi Saar MA

Tallinn 2024

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Scharlett Hansson 192542YVEM

Andmekataloogi metamudeli arendamine, immuniseerimisteatise andmekoosseisu näitel

Magistritöö

Juhendaja: Janek Metsallik
MSc
Aivi Saar MA

Tallinn 2024

Author’s declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Scharlett Hansson

02.01.2025

Abstract

This thesis focuses on the development of a meta-model of a data catalogue for the management of health data, using the example of an immunisation record. The work builds on the European Health Data Space (EHDS) regulation, which aims to improve the re-use of health data and facilitate data-driven decision making in research, policy making and innovation. The EHDS regulation emphasises the importance of harmonised metadata to ensure data discoverability, accessibility, interoperability and re-use.

The importance of this research stems from the need to improve the data management capacity of the Estonian health sector and to link it to international standards, thus supporting the objectives of European health data sharing. Using a design science research methodology, existing metadata management standards were analysed and a prototype metamodel was developed. The work evaluated the suitability of the healthDCAT-AP standard for the creation of health data catalogues in the Estonian context. This thesis is written in English and is 50 pages long, including 5 chapters, 6 figures and 1 table. All the sources used to write the research have been cited and 44 sources were used. The Google Scholar search engine was used to search for sources.

The main results show that although healthDCAT-AP allows for generic metadata descriptions, it does not fully cover the specific needs of Estonian health data, especially in terms of semantic coherence and technical integration. The study underlines the need for the development of a data repository to support a more efficient sharing and re-use of health data at both national and international level. For this thesis, design research is utilised.

Annotatsioon

Andmekataloogi metamudeli arendamine, immuniseerimisteatise andmekoosseisu näitel

Magistritöö „Andmekataloogi metamudeli arendamine, immuniseerimisteatise andmekoosseisu näitel“ keskendub terviseandmete haldamise ja taaskasutuse tõhustamisele Euroopa Terviseandmete Ruumi (EHDS) regulatsiooni kontekstis. Töö eesmärk on arendada ja hinnata andmekataloogi metamudelit, mis vastab Eesti tervishoiusektori vajadustele ning toetab terviseandmete jagamist rahvusvaheliselt tunnustatud standardite alusel.

Uurimistöö teoreetiline osa analüüsib metaandmete haldamise standardeid ja nende rakendatavust tervishoiuvaldkonnas, rõhutades healthDCAT-AP standardi potentsiaali. Praktilises osas kavandati ja hinnati standardil põhinev metamudeli prototüüp, mille sobivust testiti erinevate kasutusjuhtude alusel, sealhulgas teadlaste, poliitikakujundajate ja arendajate vajaduste kontekstis.

Töö tulemused näitavad, et kuigi healthDCAT-AP toetab üldist metaandmete haldust, vajab see täiendusi, et tagada andmete semantiline sidusus ja tehniline integreeritavus. Magistritöö rõhutab andmekataloogi arendamise tähtsust tervishoiuandmete tõhusaks haldamiseks ja taaskasutamiseks, pakkudes lahendusi, mis aitavad Eesti tervishoiusektoril täita EHDS regulatsiooni nõudeid ja parandada andmepõhiste otsuste kvaliteeti.

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti 50 leheküljel, 5 peatükki, 6 joonist, 1 tabel.

List of abbreviations and terms

AP	Application profile
Conceptual data model	An abstract representation of the real world [1]
Data catalogue	Respository or database to harvest or ingest metadata from different sources. [2], [3], [4], [5]
Data model	Data representation specifying properties, structure and relationships[1]
Data element	A unit of data, such as country code, local number[1]
DATS	Data Tag Suite
DCAT	Data catalogue vocabulary
DDI	Data Documentation Initiative
EU	European Union
EHDS	European health data space
FAIR	Findable, Accessible, Interoperable, Reusable
Metadata	Data about data, provides context to the actual data, not including personal or health data [2], [5]
Metamodel	Metamodel often related to data model. Metamodel can have different levels: a high-level conceptual model for describing relationships between systems. A lower-level model, described details of attributes, entitites, elements and processes. [2]
TEHIK	Health and Welfare Information Systems Center
TEHDAS	Joint Action towards European health data space
OBDM	Ontology-based data management
Physical model	Represents the view how data is stored in database [2]
RDF	Resource description framework

Table of contents

List of figures

List of tables

1 Introduction

The European Health Data Space (EHDS) regulation aims to improve healthcare efficiency, support research and innovation, and enhance public health initiatives. Therefore, to enhance evidence-based decision-making in regulation, policy, research and health technology assessment, and clinical practice, the use of real-world data and real-world evidence, including patient-reported outcomes, should be encouraged. To maximise their potential, datasets made available for secondary use should be as complete as possible.[6]. These aforementioned possibilities and drivers that Estonia is seeking to address, both the 2030 Development Plan from the Ministry of Economic Affairs and Communications, promoting data reusability, implementation of novel technologies and data quality [7], alongside with the 2030 Population Health Plan from the Ministry of Social Affairs emphasize data-driven decisions [8].

The need for reusable data was acknowledged in the Health Sense project. The project aimed to develop a universal data model and standards for health pathways, aligning with international standards. It focused on facilitating secondary use of health data and minimizing data duplication. The project team emphasized the need to streamline interoperability management methodologies to enhance the multiple uses of health data across different interest groups. [9].

To enhance the reusability of research data and minimize data duplication the EHDS regulation stipulates the creation of a cross-border infrastructure for secondary use of electronic health data by designated contact points in the Member States. Datasets accessible across borders must be accompanied by a metadata catalogue specifying their source, scope, key characteristics, nature and access conditions. The European Commission will develop an EU Datasets Catalogue (HealthData@EU) connecting the national catalogues of datasets established by health data access bodies and other authorized entities.[6], [10]. The HealthData@EU Pilot project, a two-year EU4Health initiative, launched in October 2022, aimed to build a pilot version of the EHDS infrastructure. This project, involved 17 partners, connected data platforms across the EU to facilitate cross-border health data sharing. Key priorities included metadata discovery,

standardized data access requests, and guidelines for data standards, quality, security and transfer. [11]. One of the deliverables from HealthData@EU Pilot project was the standard for facilitating seamless cross-border sharing of electronic health data across the EU, called healthDCAT-AP[12]. HealthDCAT-AP is a specialized version of the DCAT application profile, designed to share information about health-related datasets and data services across Europe. By developing a healthDCAT-AP standard, HealthData@EU aims to standardize health metadata within the EHDS. This will enhance interoperability, discoverability, and accessibility of electronic health data across the EU. [13].

Considering all the abovementioned aspects, the Health and Welfare Information Systems Center (TEHIK) faces a number of challenges ahead, including the implementation of data catalogue. In collaboration with the Ministry of Social Affairs, the Estonia towards EHDS (EST2EHDS) project is underway, which aims at the creation of national services and infrastructures, the creation of a functional Health Data Access Body (HDAB) and the development of digital business capabilities that will enable interaction between the HDAB, data holders and data users. One of the expected outcomes of EST2EHDS by 2026 is a data catalogue with at least 10 health-related published datasets.[14].

Data catalogue is a repository of metadata, that has been harvested from diverse sources [2], [5]. Metadata is essentially data about data and provides essential information about data, enabling organizations to understand, manage, and process it effectively. It ensures data quality and facilitates various data operations, including integration, security, and governance.[2]. Harmonizing data and metadata standards can significantly improve data integration and analysis, enabling more efficient research and better healthcare outcomes.[15], [16]. It accelerates data dissemination in a standardized format, increasing the likelihood of secondary analysis and citations. By exposing metadata in a standardized format, datasets become discoverable through search engines.[16], [17], [18], [19]. The lack of standardized metadata is a significant barrier to making research data FAIR (Findable, Accessible, Interoperable, and Reusable). Despite various efforts to address this issue, confusion persists due to the proliferation of best practices and recommendations. Even when metadata standards are adopted, metadata is often not exposed or reused effectively.[6], [16], [17], [20]. This thesis aims to design the metamodel for data catalogue, adapted to the needs of healthcare sector and validate the hypothesis: healthDCAT-AP standard is suitable for data catalogue as a data exchange standard agnostic metamodel.

2 Literature review

Thesis focuses on data catalogue, metadata management and ontology standards that can be used for metadata management. The usage of ontology standards improves the findability of information on the web.[16] This overview of latest research aims to widen the perspective on metadata management across domains and bring new perspective for metamodel development.

2.1 Existing data catalogues, its value and used ontology standards

Data has become incredibly valuable [18], [21], helping us solve problems like pandemics and climate change. To make the most of this data, we need to describe it clearly and organize it into catalogues. This is especially important for open data, which is shared freely by governments, researchers, and others. Metadata, which is like a digital label for data, helps us find, use, and understand data. It's also essential for reproducing experiments and making informed decisions. Not having good metadata can be very costly, with one study estimating a loss of billions of euros for the European economy.[21].

Public service metadata is like a digital label that tells us what a government service is, how it works, and what laws it follows. This label helps people find the right service and ensures that different services can work together smoothly, even across borders.[4]. While commercial options like Collibra Data Catalog, Accurity data catalogue, Zeenea data explorer and acryl.io exist or open-source tools like TrueDat, Datahub and Magda are also available. [22]

In Estonia, the government offers institutions a free data governance tool called RIHAKE. This tool allows institutions to describe their datasets, including their content, classifiers, lists, machine-readable services, and data and business glossaries. RIHAKE fully adheres to the national data description standard and enables data descriptions to be easily transferred to the upcoming data portal. RIHAKE operates on a 'one institution, one dataset' principle, providing institutions with an overview of their datasets. In contrast, the data portal generates an interagency perspective on datasets and their reusability.[22] Whilst RIHAKE adheres to national data description standard, which combines in itself Dublin Core, data vocabulary version 3 (DCAT-AP) and for glossary descriptions ISO

25964 “Thesauruses and their interoperability with others with other dictionaries” , it also has some properties, that are outside these named international standards, E.g. data glossary name, data origin, geographical coverage.[23]. The Estonian Open Data Portal is a centralized web platform that serves as a data catalogue, providing descriptions of datasets and facilitating their discovery and reuse. Its primary goal is to consolidate open data from various sources and make it easily accessible to both private and public sector users, thereby promoting data reuse. In addition to the data catalog, the portal also features use cases and articles related to open data, offering valuable insights and information for users. Although, it should harvest the metadata from RIHAKE, some of these datasets have not been updated for a while now. For example, dataset for immunisations against COVID-19 should been updated weekly, but last updated date is January 2023. There are many more that should be updated at least this year, but was updated 2022 or at the beginning of 2023. [24].

The data catalogue for biomedical datasets has been focusing in the data that has been generated during funded healthcare or biomedical projects. Consequently, the dissemination, deposition, and interconnection of these project outputs are often inconsistent and challenging, leading to missed opportunities for data reuse and the risk of redundant efforts. Therefore, implemented in 2017 the Data Catalogue, as a centralized platform to improve the discoverability of datasets produced, curated and reused within different medicinal projects. The Data Catalogue’s initial model was limited. Designed for translational medicine and clinical trials, it focused on patient data and interventions, making it difficult to include other study types, such as animal or in vitro studies. To broaden the scope and comply with FAIR principles, the current Data Catalogue uses the Data Tag Suite (DATS) model. DATS was created for data discovery prototype and is used in several other projects. It offers a standardized way to represent metadata across diverse projects. DATS model is interoperable with other standards such as Data Catalogue Vocabulary (DCAT), schema.org schemas and other ontologies used for data catalogues. Although, the authors state, that through development and curation efforts, they significantly expanded the Data Catalogue, increasing the number of entries from 77 to 356. This includes a substantial rise in public datasets from 67 to 185 and an increase in the number of represented objects from 10 to 186. Welter, et al aims to further enhance the Data Catalogue by expanding ontology annotations within the metadata. These annotations will be used to improve the search functionality through semantic query

expansion. However, they face with one significant challenge beyond the Data Catalogue's direct control is the lack of consistent data deposition in many projects. While the Data Catalogue tracks study and dataset metadata, data not submitted to repositories risks being lost after project completion. The authors of this research state, that the Data Catalogue is a valuable resource that encompasses various data and experiment types offering a standardized and curated representation of project-and data-level metadata. It aligns with FAIR principles in its design and enhances the FAIRness of the metadata it contains. Also, they understand, that ongoing development and curation will strive to further solidify the Data Catalogue as a fully FAIR-compliant mature asset for the scientific community. [3].

Oliveira, et al (2024) state that semantic layering and data categorization provide context and control, making data meaningful for analysis. They propose that applying these concepts to reports and dashboards can streamline their development, creation, maintenance and management. This paper introduces Data Catalogue, a framework for documenting and connecting measures with their underlying context while preserving crucial restrictions for business insights The Data Catalogue consists of interconnected components, such as data profiling, data quality, alert and data security. These can be combined or separated as needed for discovering, analysing, monitoring and controlling incoming data. The system autonomously creates metadata based on predefined rules. The authors, explored Power BI as a use case, creating a graph to visualize key Power BI concepts, including DAX functions for metric creation.[25]. DAX is a specialized language designed for working with data models using formulas and expressions. It's employed in various Microsoft products, including Power BI, Analysis Services and Power Pivot for Excel. [26] . Oliveira, et al (2024) mapped these concepts to Data Catalogue base concepts to establish relationships between terms, data and concepts in each layer. The resulting knowledge graph not only represents data stored in the Data Catalogue but also assists users in validating reports and maintaining measure traceability for ongoing maintenance activities.

Esbai, et al (2023) presents a model-driven architecture (MDA) approach to streamlining the development of data warehouses. By leveraging UML class diagrams as conceptual models, they propose a transformation process to generate both a physical information model (PIM) representing a data warehouse repository and a physical system model (PSM) representing an online analytical processing (OLAP) cube. The primary benefit is

increased productivity and reduced implementation costs. To facilitate the transformation process, they developed query/ view/transformation (QVT) rules that convert UML class diagrams into relational data warehouse and OLAP cube models. Subsequently, acceleo query language was employed to generate SQL and XML code, enabling the automatic creation of multidimensional data warehouse structures and OLAP cubes. The key contribution of this work lies in providing well-defined metamodels and comprehensive transformation rules to bridge the gap between conceptual models and operational data warehouse and OLAP components. By automating the generation of relational data warehouse and OLAP cube models, including implementation code, thus the enhancement of efficient and reliable data is in the data warehouse. [27].

However, Tseung, et al (2023) argues that inconsistent use of standardized properties due to fragmented schemas application. Therefore, for biological repositories adopted schema.org dataset standard, to describe infectious diseases metadata. Tseung, et al (2023) developed a user-friendly and adaptable data schema. This schema leverages the established Schema.org framework, ensuring seamless integration with data cataloguing projects like Google Dataset Search and BioSchemas. Rather than proposing a new standard, they focused on connecting existing ones by extending relevant Schema.org classes with disease-specific properties. Furthermore, to promote data exchange across different platforms. [16]. On the other hand, schema.org states itself that it is designed for health and medical web content, targeting entity relationships rather than clinical markup or data exchange. While, it's broad scope, specific content may use only a subset of the schema. It complements existing medical vocabularies and allows for the annotation of entities with medical codes.[28].

Croce, et al (2024) approached to the diabetes mellitus dataset challenges, that they have collected from over 600 000 patients, so they adopted an ontology-based data management approach. An ontology-based data management involves using an ontology, a share conceptualization of the domain, and declarative mappings to link data sources to this ontology. Croce, et al (2024) examined the diabetes datasets to understand its semantics and the underlying domain. To comprehend the diabetes dataset, they analysed its metadata, including medical standards and proprietary encodings. For modelling the data and metadata of the diabetes domain. They used W3C Web Ontology Language (OWL), which includes a comprehensive set of relevant concepts, their interconnections and defining characteristics. While incomplete, the ontology presented in the diagram

provides valuable context for interpreting the database data. Based on the ontology, they designed a database schema that closely aligned with the conceptual model. The schema was divided into two parts: data and metadata. Data tables stored actual data, while metadata tables contained information about data types, value ranges, national codes and other relevant attributes. To conclude, the authors state, that the diabetes dataset presents unique challenges due to its heterogeneity and the need to support various research objectives. To streamline data analytics, a comprehensive data preparation process is essential. This involves conceptualizing and cleansing the data once creating a unifies dataset for multiple analyses. Additionally, ongoing data refinement can be driven by the discovery of new inconsistencies during exploration.[29].

To summarize, as many different people, as many different approaches to data cataloguing. Some use application profile and separate tooling, to visualize and publish metadata descriptions. Some use data catalogues for connecting and maintaining datasets produced by different projects, and on the other hand the data warehouse finds solutions within its domain and with schemas, to bring context to the data, with metadata descriptions.

2.2 Previous research related to immunisation report and its metadata requirements

The immunisation and COVID-19 infection dataset offer a rich source of data for various analytical approaches, including the examination of a single data feature or variable, while being valuable for examining multiple properties to understand complex relationships and interactions between them. Bourdas et al (2023) used for their research immunisation for COVID-19 infection dataset [30], being available from Dutch openData portal, where they have described following variables to used dataset: version as version number of the dataset, date of report – date and time on which data file was created, region level, date of statistics, population, coverage and links to other relevant datasets.[31].

Brunson, T, et al (2023) employed a data dictionary that outlines various entities and their attributes to facilitate effective data management and interoperability in health information systems, particularly in vaccine-induced disease datasets. The researchers utilized vaccine ontology to categorize diverse vaccine types, including live, inactive flu vaccines and yellow fever vaccines. Vaccine investigation ontology is used to classify

various variables, such as min and max age, gender, race, immunisation time, immunisation time unit and many more.[32].

To conclude, not many researchers describe used dataset metadata and usually datasets used in researches, are not described and catalogued for future references. Therefore, the analysis of open data portals and its consistent of immunisations datasets is looked into.

2.3 The application of DCAT-AP in health data management

The IDERHA project aims to implement the principles of EHDS2 and recommended standards for data discovery, using RDF vocabulary DCAT-AP. Although, the enabler for secondary use of data on semantic interoperability would be OMOP standard. For data exchange also enabled by DICOM and HL7 FHIR standards. The authors state, that key challenge lies in the ongoing refinement of certain standards, such as healthDCAT-AP, the extension of DCAT-AP. Furthermore, extensions to the OMOP Common data model (CDM) are necessary to represent patient-generated health data. Finally, effective mapping between standards like FHIR and OMOP is crucial for seamless data exchange.[18].

Life sciences research heavily depends on high-quality data generated throughout the entire research process, from biological material sourcing to data analysis and sharing. Therefore, Müller, et al (2023) faces the challenges new age big data continuously give, such as projects use different data formats, standards and terminologies, lack of metadata, that provide the context to the data, lack of data quality – ensuring the accuracy, reliability and completeness and much more. To tackle these challenges, the authors developed software to support researchers to publish their datasets and, in a FAIR, (findability, accessibility, interoperability and reusability) manner. The fair data point (FDP) acts as a software layer that enables data owners to expose data and metadata in a FAIR manner, while also allowing users to discover and access relevant data. The fair data point architecture includes an application programming interface (API) for software interactions, a Web client for human-oriented access, and key components like the metadata provider, access control, and metadata schemas. The metadata provider facilitates metadata access, while the access control manages permissions for adding, deleting or editing metadata. The fair data point supports various digital object types, such as repositories, catalogues, datasets, and utilizes metadata schemas to ensure consistency

and validation. The FDP leverages the W3C's Data Catalogue Vocabulary (DCAT) version 2 as its metadata foundation. DCAT provides a model for describing catalogues, facilitating interoperability between data catalogues published on the Web. It enables the standardization of dataset and data service descriptions, improving discoverability and aggregation of metadata from multiple catalogues. DCAT distinguishes between Datasets (collections of data) and DataServices (services that provide access to datasets). Catalogues, a subclass of Dataset, represent aggregations of metadata records. The FDP extends DCAT by introducing the FAIRDataPoint concept, a specific type of Data Service that serves metadata catalogues and records. FAIRDataPoint is a subclass of MetadataService, which is itself a subclass of DataService. To address the specific needs of data portals in Europe, the DCAT Application Profile (DCAT-AP) was developed. The DCAT-AP specification aims to improve data discoverability by enabling cross-data portal search. Several extensions, such as DCAT-AP-JRC and Health DCAT-AP, aim to standardize metadata for specific domains, aligning with FAIR principles and the needs of the European Health Data Space.[33].

Population Health Information Research Infrastructure (PHIRI) launched its Health Information Portal in May 2021, serving as a centralized platform for accessing population health, healthcare data, information, and expertise across Europe. The authors claim: *"this portal is a first-of-its-kind-resource",* to offer a comprehensive collection of 350+ population health data sources from national and international organizations, 170+ health information projects within countries and across Europe, 190+ training materials and 15 research networks. PHIRI maintains a network of 27 national nodes, continuously expanding. These nodes serve as national liasions, bringing together relevant stakeholders within their respective countries. The Health Information portal adhering to the FAIR principles (findable, accessible, interoperable and reusable). Established metadata catalogues on the health information portal for health data, population health studies, publications and capacity building alongside with training activities. Designed metadata templates for each catalogue, ensuring consistent information organization and description across Europe, followed existing metadata standards, such as DDI Lifecycle, DCAT-AP and Dublin Core with user-driven adjustments made in the second phase of the project. For the discoverability of the portal researchers implemented schema.org on additional health information portal pages, improving the portal's findability by search engines.[34].

3 Methodology

This thesis methodology basis on the Design Science Research (DSR) methodology proposed by Peffers, et al (2007). The Design Science Research is continuously evolving, although its core principles have remained the same. The process entails and promotes rigorous design, evaluation and communication of IT artifacts intended to solve identified organizational problems. These artifacts can range from traditional constructs and models to innovative social solutions or new properties of resources. As DSR continues to advance, it will undoubtedly play a crucial role in shaping the future of information systems and technology. [35]. Thesis writing process based on the DSR is shown in Figure 1.

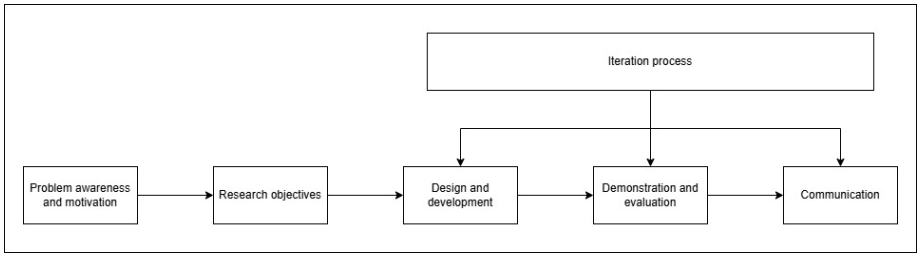


Figure 1 Thesis process based on the design science research [35].

The outcome of this thesis would be designed artifact developed metamodel for data catalogue by the deliverable of this research as described in chapter Results. The metamodel is evaluated by acceptance criterions in described use cases.

Preliminary research began with literature overview following the methodology as qualitative descriptive research [36] The literature overview aimed to formulate the hypothesis for this thesis more accurately, dependent on the current approaches for the research problem. The literature overview provides an up-to-date view on recent progress in regards to discussed and used standards and their adoption. Due to timely and evolving topic only Google Scholar search engine was used and since 2023 articles, paper, opinion reviews, books were found.

The articles, papers, books were excluded by written language – used only English and Estonian written sources. Excluded sources not eligible for the topic nor the field of research. Therefore, the selection of papers was based on titles, keywords, abstracts, relevant topic to the specific research field, such as health sector or health related data management. Also, papers written in English or Estonian. The papers and research

published across the world were cited in this research, due to topic being new and not yet published a lot research from Europe. The forerunners in EHDS interoperability are the countries participated in EHDS pilot project, where the healthDCAT-AP extension developed. *The European Interoperability Framework was last updated in 2017, therefore used for theoretical part, to illustrate the impact and the level of architecture for interoperability standards.* The data management body of knowledge published in 2017, used for metadata management principles and explaining key definitions used in this research. Based on the preliminary research results the problem was formed more accurately and hypothesis, to solve, stated.

Research ethics are ensured by the monitoring of information security and research aspects within TEHIK, where research objectives and outputs are agreed with the management. Although, this research describes metadata and does not collect, analyse nor handle personal or sensitive raw data. Therefore, ethics committee approval is not necessary.

3.1 Problem formation

This thesis is strongly aligned with author's day-to-day work in the Health and Welfare Information Systems Center, with being the product owner and manager for Information Center, and its contributing to Estonia towards European Health Data Space (EST2EHDS) project. EST2EHDS aim is to strengthen the digital capabilities to share data cross-border, across Europe.[37] Appendix 3 illustrates the future HealthData@EU digital capabilities blueprint. The pink dots illustrate the hig-level processes that are mandated by the EHDS regulation and what TEHIK, alongside with the Ministry of Social Affairs, is obligated to fulfill.[6] Whilst, getting in depth to this topic, author's interest grew as well, therefore this research is great opportunity to showcase author's interest and passion for metadata management.

Amongst different relevant issues that EHDS addresses, this thesis focuses solely metadata and data management aspects. Therefore, the key articles concerning health data for secondary use, that mandate specific actions:

1. Article 33: Identifies types of health data that must be made available for secondary use.

2. Article 57: Tasks the European Commission to develop an EU dataset catalogue, integrating national catalogues into HealthData@EU.

3. Article 55: Mandates health data access bodies to provide metadata about available datasets, including source, scope, characteristics, nature, and access conditions. [6].

Aforementioned aspects and key articles also mandates a data quality and utility label for health data intended for reuse. This requires a data catalogue to describe and manage metadata to enable secondary use of health data in Estonia.[6], [9], [37]. Metamodel, a high-level model of how to describe the metadata, is foundational layer in data catalogue. The metamodel aims to unify logical, physical data with business metadata[2].

Research scope comes from the need to develop data catalogue, where to publish metadata. In collaboration with the Ministry of Social Affairs, the EST2EHDS project is underway, which aims at the creation of national services and infrastructures, here we have in mind the creation of a functional Health Data Access Building (HDAB) and the development of infrastructures that will enable the interaction between the Health Data Access Authority, data holders and data users. One of the expected outcomes of EST2EHDS by 2026 is a data catalogue with at least 10 datasets published. [6], [14]. To establish well-thought data catalogue basis on designed metamodel. Therefore, the development of metamodel is essential and part of data catalogue establishment [2].

3.2 Selection of standard

The European Interoperability Framework (EIF) provides guidelines for developing interoperable digital public services [38]. The EHDS development aligns with this framework, aiming to establish interoperable digital public services for research, innovation, regulation and policymaking across Europe. The TEHDAS project investigated metadata standard landscape with criterions: core interoperability principles, generic use and expectations, principles for cooperation and interoperability. Therefore, a sample of standards were evaluated based on these criterions, including ContSys and HL7 FHIR, identifying gaps and opportunities for improvement.[39]. Similar approach conducted by the National Institute of Child Health and Human Development (NICHD), to enhance data interoperability in health information systems, for secondary use of health

data. Although, they approached standards by their governance domains, for example DCAT-AP as dataset information domain, extensible access control markup language as controls and policy domain.[40]. Both, TEHDAS working group and NICHD found that no single standard fully meets the schema requirements across all domains, necessitating the use of multiple standards and potential development of new value sets.[39], [40].

While, TEHDAS working group brought out, that DCAT-AP ranked highest on the data discoverability principles and in interoperability aspects[39], it does not provide variable level, so DCAT-AP is domain agnostic and focusing on high-level knowledge on the datasets available.[39], [41].

Data catalogue vocabulary (DCAT) is an W3C RDF vocabulary used to create machine-readable descriptions of data catalogues, promoting interoperability. DCAT can be used to create standardized descriptions of datasets and data services, promoting their discoverability and enabling their aggregation. This supports decentralization, federated search, and digital preservation initiatives. While initially developed for government data portals like data.gov and data.gov.uk, DCAT's versatility has made it a popular vocabulary for publishing data catalogues in diverse contexts. [33], [41]. DCAT includes classes for datasets and distributions, representing the data itself and access points.[41] Figure 2Figure 2 shows DCAT's class structure, consisting of catalogues, datasets, data services, distributions, and catalogue records. DCAT provides a framework for organizing and describing metadata about data resources. These classes are interconnected, with datasets and data services being described through catalogue records.[41].

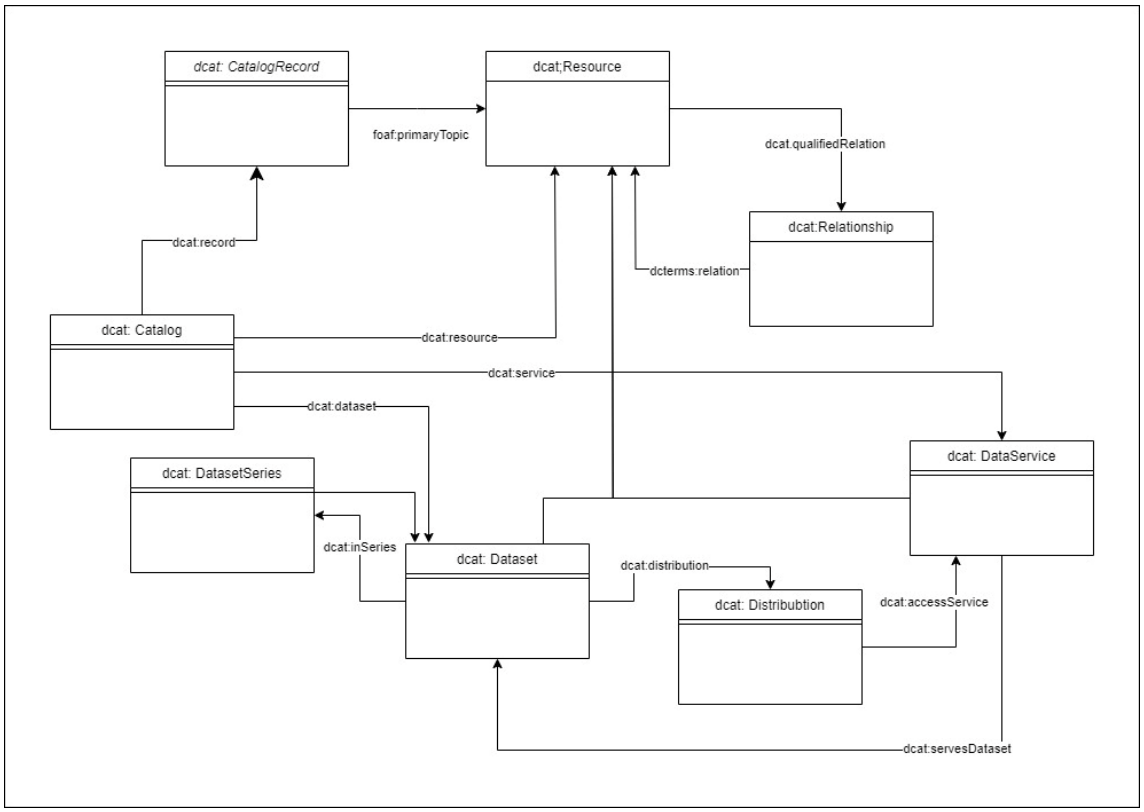


Figure 2. The author's simplification of overview of DCAT model, showing classes of resources and relationships between them. DCAT does not specify cardinality restrictions, except if indicated.

A standardized metadata model, like DCAT-AP, is essential for effective data sharing and reuse. Based on linked data principles, DCAT-AP provides a comprehensive framework for describing open data.[40]. DCAT is widely used standard for describing data catalogues, with DCAT-AP being a practical extension. The application profile is designed to facilitate data exchange without specifying technical requirements for communicating systems. The only requirement is the ability to handle RDF data. DCAT has been enhanced with clearer definitions, usage guidelines, and constraints to improve its usability and ensure consistent application. It also incorporates relevant classes and properties from other vocabularies. DCAT version 3 builds upon the previous version by expanding its scope to include additional use cases and requirements, also supports cataloguing not only datasets but also dataset series and introduces features for describing resource versioning. DCAT is flexible in handling diverse data formats, focusing on the abstract dataset rather than specific serializations. It supports describing data access services that provide various data options. For more detailed information, complementary vocabularies like VoiD can be used to express statistics about RDF-formatted datasets.[41].

HealthDCAT-AP is a specialized extension of the DCAT application profile (DCAT-AP), designed to facilitate the sharing of health-related datasets and data services within Europe. As a refined RDF vocabulary built upon DCAT-AP's foundation, healthDCAT-AP addresses the unique requirements of electronic health data[13], [18] . HealthDCAT-AP, built upon DCAT-AP, includes its essential classes: Catalogue, Catalogue Record, Dataset, Distribution, and Data Service. Utilizing RDF's flexibility, healthDCAT-AP introduces new metadata elements as triples, enriching the metadata model without affecting existing catalogue systems. The healthDCAT-AP application profile, published on December 22, 2023, by Work Package 6 of the EHDS2 pilot project, is a recommendation without official status or endorsement from any standards organization. Its structure and content are partially derived from the DCAT-AP v3.0 and DCAT-AP High Value Dataset specifications, maintaining consistency with other DCAT-AP specifications.[13].

Therefore, the development of healthDCAT-AP aims to standardize health metadata within the EHDS framework. This will enhance interoperability, findability, and accessibility of electronic health data across the European Union. The healthDCAT-AP aligns with the EHDS and EU4Health programme, supporting goals of the EHDS regulations, by focusing on privacy and security, healthDCAT-AP enables responsible and efficient sharing of sensitive health data.[6], [13]. Although, healthDCAT-AP is still in the development and not validated among member states, yet. During TEHDAS2 project, by the end of 2026, healthDCAT-AP will be validated on a larger scale.[13], [42]. The author has selected healthDCAT-AP for the metadata standard, to validate the standard in Estonian setting and provide feedback on the standard implementation to the healthDCAT-AP developers and EHDS community.

4 Results

The aim of this paragraph is to introduce all the findings that have occurred during research and development. This thesis aimed to design the metamodel for data catalogue, adapted to the needs of healthcare sector and validate the hypothesis: healthDCAT-AP standard is suitable for data catalogue as a data exchange standard agnostic metamodel.

4.1 Simplified high-level data catalogue architecture

Data catalogue will be situated into TEHIK infrastructure, keeping in mind all the information security aspects and the compatibility to E-ITS requirements. Although, data catalogue does not handle any raw data, rather than data about data. The simplified architecture model is based on the data management body of knowledge illustration about metadata architecture [2] and not showing the actual architecture for TEHIK data catalogue.

The information portal is part of the Teabekeskus ecosystem, that publishes all relevant technical documentation, code systems and value sets used in Estonian health information systems, for data exchange and implementation of these services. Data catalogue would be the newest addition to this ecosystem, alongside with metadata management.[14]. The information portal would be the web interface, where data users, such as policy advisors, researchers, software developers, can search and explore published metadata.

Although, the core of metadata management would be data catalogue, where the metadata would be described according to metamodel. The data catalogue would serve described datasets to the web interface, and would be acting as back-end module, shown in Figure 3. The data catalogue will retrieve and harvest the metadata from data warehouse, service databases, modelling tools, ETL tools and data quality tools, to describe and manage organisation’s metadata, thus having the potential to unify the Ministry of Social Affairs organisations.

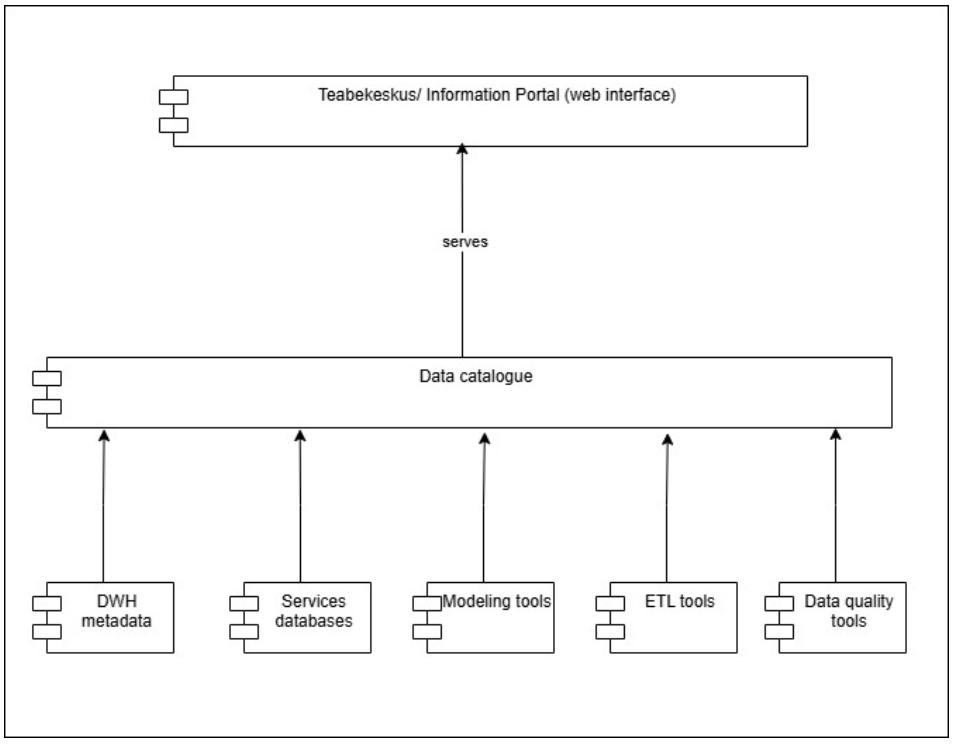


Figure 3. Authors' simplified architecture, used basis from the data management body of knowledge [2]

4.2 Use cases for metadata management

The EHDS regulation categorises metadata management to the use of secondary health data [6], therefore the use cases illustrate the main actors and their need for metadata.

The identified actors are: data manager, software developer, researcher and policy advisor.

Use case 1	
Title	As a researcher, I need clear and detailed data descriptions to assess the relevance and usefulness of datasets, so that I can identify the datasets needed for my research and use a broader sample in my study.
Actor	researcher

Objectives	Enable researchers to understand the scope, structure and content of the available datasets, enabling them to effectively identify relevant data and engage diverse populations in research.
Preconditions	1. Data descriptions are publicly available and published. 2. Descriptions are available for each dataset. 3. The researcher knows the objectives of his/her study and data requirements (e.g. population demographics, immunisation coverage).
Main scenario	1. Find relevant datasets 2. Access data descriptions 1. dataset title 2. main variables 3. population coverage 4. time period 3. Assesses the suitability of the dataset for your research work. 4. Combines several datasets, if necessary. 5. Requests access to data on the basis of data descriptions 6. Uses the requested datasets in his research work
Acceptance criterions	1. Metadata are detailed, standardised and easy to understand. 2. Researchers are able to identify datasets relevant to their research objectives. 3. Population size and coverage details are explicitly stated, allowing for an assessment of inclusion and adequacy.

	4. The researcher is able to combine datasets for a larger or more diverse population, if necessary 5. Details of access and any limitations in the metadata are clearly documented

Use case 2	
Title	As a policy advisor, I need clear and accessible data descriptions for immunisation data so that I can make data-driven policy decisions.
Actor	policy advisor
Objectives	Accessibility and knowledge of the meaning of immunisation data descriptions (vaccination coverage, vaccine types, demographics) to make evidence-based policy decisions.
Preconditions	1. The policy advisor has access to the immunisation data analytics 2. The data descriptions are up-to-date and organised so that the analysis has a clear context. 3. Policy advisor has an understanding of policy needs (public health, vaccination coverage).
Main scenario	1. Data collected during immunisation are stored in a database. 2.Immunisation metadata are described according to the meta-model. 3. Publish the metadata publicly. 4. Policy advisor have access to immunisation data descriptions for research purposes.

Acceptance criterions	1. Policy advisor can easily understand data descriptions 2. Data descriptions are clear and easy to interpret with definitions and implications. 3. Analysis is based on data descriptions to provide context. 4. The policy maker can use the data for data-driven decision making, in particular for immunisation policy and action planning.

Use case 3	
Title	As a data manager, I need to ensure that immunisation data is semantically interoperable so that it can be consistently understood, shared and reused across systems and stakeholders.
Actor	data manager
Objectives	Ensure semantic interoperability of metadata using standardised vocabularies, ontologies and formats, enabling seamless integration of data, understanding between different systems and stakeholders.
Preconditions	1. The data manager has access to guidelines or tools for implementing semantic interoperability. 2. Metadata standards and controlled vocabularies (SNOMED CT, LOINC, RDF) are defined and available. 3. The data manager is familiar with the basic principles of semantic interoperability.
Main scenario	1. Log in to the system 2. Review the dataset and the data description fields.

	3. Add descriptions, based on standardised vocabularies. 4.Check the completeness and accuracy of the descriptions 5.Ensure machine readability of metadata 1. Structure metadata in formats that support semantic interoperability. 6. Version, log changes downstream 7. Save and validate data descriptions 8. Test interoperability in systems 1. Share metadata with other systems or key actors, to verify interoperability. 2. Identify and correct inaccuracies and ambiguities. 9. Publish 10. Log out of the system
Acceptance criterions	1. Metadata are described by recognised dictionaries or ontologies. 2. The metadata structure supports semantic interoperability through machine-readable forms. 3. Documentation of standards and conformance is clear and accessible. 4. Metadata can be successfully integrated and interpreted by external systems.

Use case 4	
Title	As a software developer, I need detailed data descriptions and database structures to identify up-to-date data elements so that I can design and develop quality healthcare services.
Actor	software developer
Objectives	Allows the developer to easily assess the structure of the dataset to integrate accurate and relevant data elements into new healthcare services, ensuring quality and efficiency in development.
Preconditions	1. Datasets and their descriptions are available in a repository and/or data catalogue. 2. Metadata include information on the content, structure and timeliness of the data model. 3. The developer has a clear understanding of the requirements of the healthcare service.
Main scenario	1. Entering the data directory 2. Search for relevant datasets using search filters. 3. Accessing metadata, including dataset title, elements of the dataset, relationships between elements, data types and constraints, data formatting and compatibility, 4. Assess the timeliness and structure of the dataset, including how easily the structure of the dataset can be integrated into the planned service. 5. Plan database integration 6. Develop the data interchange within the service and the alignment of the dataset - conceptual systems with the catalogue.

	7. Iterate and optimise
Acceptance criterions	1. The developer will have easy access to and understanding of the data specifications, including the database structure and key elements to be generated. 2. Relevant datasets are identified and evaluated for use in the health service. 3. The service database is being designed or updated to harmonise the conceptual systems and descriptions. 4. Documentation ensures reproducibility and future maintenance.

The alternative scenario for all described use cases is to continue metadata management in excel sheets, that some of the organisations still use, but excel sheets does not help to unify business glossary terms nor automatically generate database structures.

4.3 Simplified data flow, integrated with metadata management

Based on the gathered knowledge within preliminary research and the knowledge from HealthData@EU, the metadata management process could enhance and unify the organisations knowledge of its data.[2]. Whilst, currently all the subprocesses align quite linear way, from agreeing on the data composition, to data analytics. Figure 4. shows that within metadata management the each subprocess adds or receives added value from the metadata management, such as better alignment within terms and understanding glossary terms for business owners, data quality rules, automatically generated database structure from the basis of metamodel and agreed data composition. Really important to highlight, is the potential to merge different data structures, mapping to groups and keep aligned for data warehouse, to keep the context of the data through time, because research can be done with data within 10years or 50years and using same dataset.

For the context and cohesiveness, the abovementioned use cases are highlighted in the data flow, to show which steps the actors need data descriptions and on what purposes.

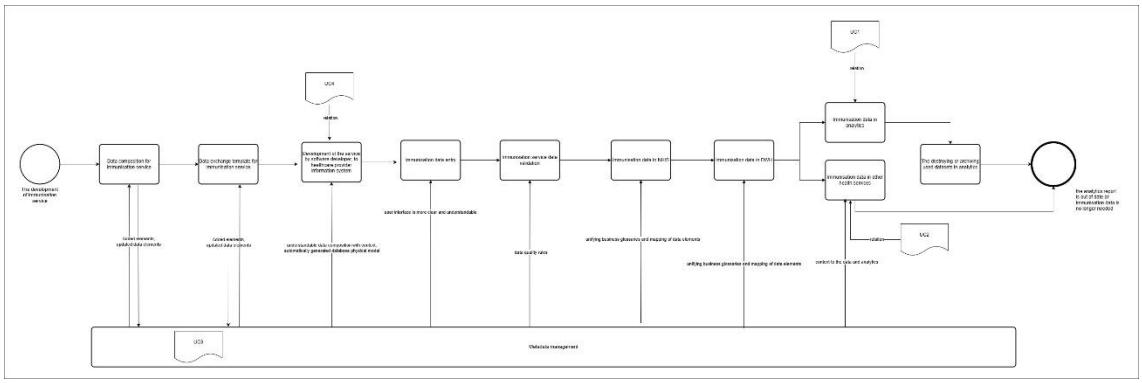


Figure 4. Authors' vision for simplified data flow, symbiosis with metadata management.

4.4 Proposal for data catalogue metamodel

The artifact of this thesis is the proposal of the metamodel for data catalogue, to maintain data lifecycle and trace the data usage in services, analytics or reearches.

Taking into consideration the data flow, how TEHIK processes are aligned and where metadata emerges into play and how the unifying of the terms really come into play, the conceptual metamodel takes into consideration the points or processes, where metadata emerges and each process affects the next. The conceptual metamodel is modelled, using the data management body of knowledge example[2] as inspiration and fitted according to data flows and the needs of the metadata users. Also, important to notice, that the aim of the data catalogue is also the power to trace data lineage, where data is used in different services or analytics, to attribute level.

Therefore, Figure 5 illustrates the data composition for immunisation service as clinical level, with information model modelled and talked through with clinicians. The data exchange template represents the same step in data flow as shown in Figure 4. Data entrys will be retrieved from the developed service database, as the tables are harvested to data catalogue, for describing. Also, the connections with technical architecture level are being made through database representations. If data exchange template attributes have being structured with agreed upon terminology usage and used value sets, to identify the structured area, then also terminology bindings are considered as metadata.

The business metadata is populated with a glossary of terms and the terms that comprise it, the meaning of the terms is intended to express the meaning of these terms according

to the terms used in the health sector or the agreed meanings of business terms used in the organisation.

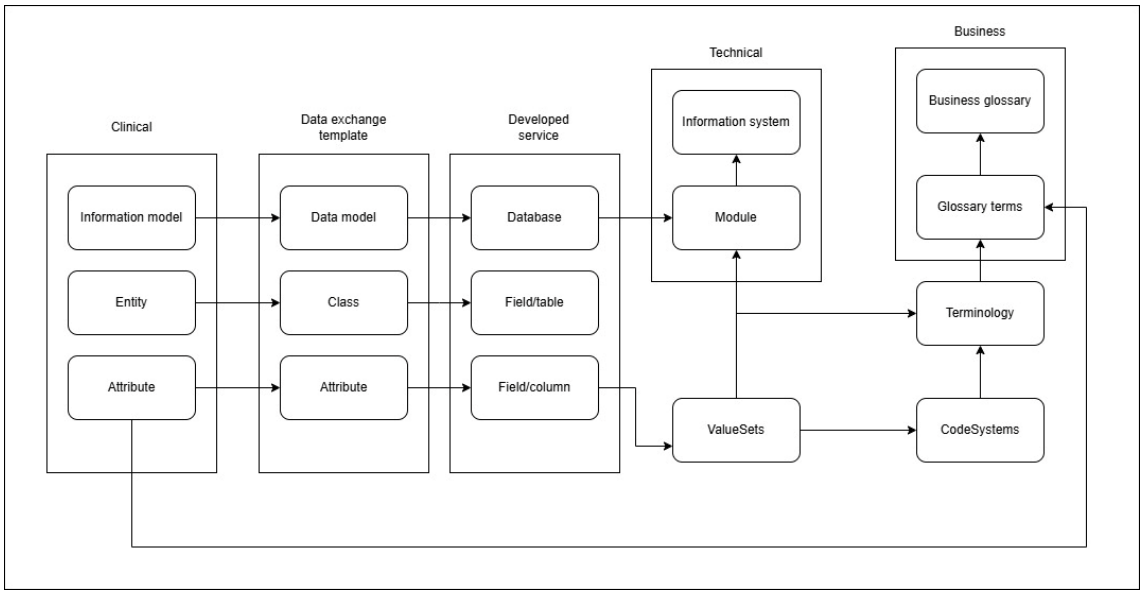


Figure 5. Author's proposal for conceptual metamodel.

Every model needs standard because the standardized structure streamlines data preparation through ETL processes, enabling consistent analysis technique across data usage. A standardized metadata ensures consistent labelling of interventions and outcomes, which is essential for machine learning and metadata management within clinical domain.[17].

This thesis aimed to design the metamodel for data catalogue, adapted to the needs of healthcare sector and validate the hypothesis: healthDCAT-AP standard is suitable for data catalogue as a data exchange standard agnostic metamodel. For the best use of the standard, all mandatory classes and the mandatory fields within them were identified. Annex 4 outlines the elements of a complete healthDCAT-AP and the corresponding immunization record data descriptions. From the data descriptions and the elements of the standard, it is clear that the standard does not go to the level of an attribute element.

In the healthDCAT-AP standard, we describe the data at a very conceptual level, with a statement of the authority publishing the dataset, free text descriptions for each item, and additional specifications that characterize the element rather than connecting at the attribute level.

To further illustrate, the Figure 6 expresses the immunization service data set description in the healthDCAT-AP standard, with most of the mandatory elements highlighted.

The healthDCAT-AP standard goes hand in hand with DCAT-AP and other RDF vocabulary terms, so it is important to highlight prefixed dictionaries at the beginning of the code, which significantly shorten the program code. The machine-readable form describes the dataset, including name, description in free form, but it is permissible to highlight the main attributes, although no association with the actual attribute is made in the background. The accessibility of the dataset, the publisher, and the quality label are also described, as required by the EHDS regulation. To highlight, it is example form for machine-readable metadata, not all links or identifiers are real.

```
@prefix dcat: <http://www.w3.org/ns/dcat#>.
@prefix dct: <http://purl.org/dc/terms/>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix foaf: <http://xmlns.com/foaf/0.1/>.
@prefix vcard: <http://www.w3.org/2006/vcard/ns#>.
@prefix dqv: <http://www.w3.org/ns/dqv#>.
@prefix oa: <http://www.w3.org/ns/oa#>.

<http://example.com/dataset> a dcat:Dataset ;
    dct:title ""Immunisation service"@en " ;
    dct:description "The EST2EHDS project needs to describe immunisation
dataset,
    in order to ensure structured data collection for immunisation as a
procedure and monitoring for data quality purposes.
    This dataset includes metadata, immunisation procedure (along with
adverse effects, manufacturer, dose quantity).";
    dct:identifier "test-dataset-id-0" ;
    dct:issued "2024-05-27T15:00:00+02:00"^^xsd:dateTime ;
    dcat:theme <https://www.wikidata.org/entity/Q58624061> ;
    dct:creator [
        a foaf:Agent ;
        foaf:name "Tervise ja Heaolu Infosüsteemide Keskus";
        dct:identifier "test-creator-id-000" ;
    ] ;
    dct:rights [
        a dct:RightsStatement;
        rdfs:label "Publicly available"@en ;
        ];
    dcat:mediaType [ <http://www.iana.org/assignments/media-types/text/tab-
separated-values>;
    ] ;
    dct:publisher [
        a foaf:Organization;
        foaf:name "Tervise ja Heaolu Infosüsteemide Keskus";
        foaf:mbox <mailto:info@tehik.ee>;
        foaf:homepage <https://www.tehik.ee/ >
    ] ;
    healthdcatap:healthTheme
<https://www.wikidata.org/entity/Q58624061>,<https://www.wikidata.org/entity/
Q7907952> ;
    dqv:hasQualityAnnotation [
        a dqv:QualityCertificate ;
        oa:hasTarget
<https://akk.tehik.ee/public_data_structure/immuniseerimise-teatis/1.0;
        oa:hasBody <https://acertificateserver.eu/mycertificate> ;
        oa:motivatedBy dqv:qualityAssessment.
        ];

      dct:license <https://opensource.org/license/mit>;
      dct:modified "2024-07-11T11:48:00.923Z"^^xsd:dateTime.
```

Figure 6. Machine-readable form of immunisation service, utilised in healthDCAT-AP standard

4.5 The acceptance criterions vs metamodel standard suitability

This chapter aims to validate the healthDCAT-AP suitability as metamodel standard. The proposed metamodel for the data catalogue is modelled according to the use cases and data flows described in the previous chapters. The suitability of the standard is validated against use case acceptance criteria, where the most relevant aspects for each metadata use case are highlighted.

The Table 1 shows that the researcher's needs are resolved by the healthDCAT-AP standard for data descriptions with 5 out of 5 needs as the machine-readable shape is also human-readable to be disclosed to further increase findability and reusability.

The needs of the policy maker are also fully addressed by the healthDCAT-AP standard, 4 out of 4.

However, the complication goes to the needs of the data manager and the software developer, where the Table 1 shows that the proposed standard solves 0.5 out of 4 and 0 out of 4 needs, respectively.

UC	Actor	Acceptance criterion	healthDCAT-AP
1	researcher	1. Metadata are detailed, standardised and easy to understand.	Yes
		2. Researchers are able to identify datasets relevant to their research objectives.	If published, then yes
		3. Population size and coverage details are explicitly stated, allowing for an assessment of inclusion and adequacy.	Yes, healthDCAT enables the descriptions for population size and coverage
		4. The researcher is able to combine datasets for a larger	Yes, the overview of datasets belonging to

		or more diverse population, if necessary	one catalogue is stated under the Catalogue property.
		5. Details of access and any limitations in the metadata are clearly documented	Yes, license and rights statement enables for researcher to understand the limitations within dataset
2	policy avisor	1. Policy advisor can easily understand data descriptions	Yes
		2. Data descriptions are clear and easy to interpret with definitions and implications.	Yes
		3. Analysis is based on data descriptions to provide context.	Yes
		4. The policy maker can use the data for data-driven decision making, in particular for immunisation policy and action planning.	Yes
3	data manager	1. Metadata are described by recognised dictionaries or ontologies.	Partly, RDF vocabularies are meant to categorise and publish metadata, but not store metadata understandable way

		2. The metadata structure supports semantic interoperability through machine-readable forms.	No, healthDCAT-AP structure does not support semantic interoperability fully.
		3. Documentation of standards and conformance is clear and accessible.	n/a
		4. Metadata can be successfully integrated and interpreted by external systems.	No, metadata described in healthDCAT-AP allows publish metadata understandable form, but not to integrate other systems
4	software developer	1. The developer will have easy access to and understanding of the data specifications, including the database structure and key elements to be generated.	No, healthDCAT-AP is too high-level to include database structure and key elements in a machine-readable form
		2. Relevant datasets are identified and evaluated for use in the health service.	No, relevant datasets are in free text.
		3. The service database is being designed or updated to harmonise the conceptual systems and descriptions.	No, healthDCAT-AP does not enable that.

		4. Documentation ensures reproducibility and future maintenance	n/a

Table 1. healthDCAT-AP validated against use case acceptance criterions.

4.6 Communication about the metamodel

This research will be presented to the chair's committee, which will give an overview of the aim of the research, its results and future plans.

The thesis will also undergo approval by the Head of Department and the Head of Information Security in TEHIK, which will provide an opportunity to apply the findings of the thesis in a real-life environment, which in turn will allow the author to plan the next iteration of the thesis and further refine his/her research.

On the topic of metadata and metamodel for data catalogue, the author presented his research proposal at the E-Helse i Norge 2024 conference in Norway[43].

The author of the thesis has also published the machine-readable sample data developed in the course of the research, together with documentation updates, on the public GitHub repository (https://github.com/Scharlett/thesis_metamodel).

5 Discussion

Thesis aimed to design the metamodel for data catalogue, adapted to the needs of healthcare sector and validate the hypothesis: healthDCAT-AP standard is suitable for data catalogue as a data exchange standard agnostic metamodel and this chapter provides a comprehensive analysis of the study's findings and their broader implications. Given the novelty of the given research topic and in the light of the approval of the EHDS regulation, the analysis is enriched with examples and practical perspective from author's day-to-day work. The aim is thoroughly addressing the research objectives initially posed at the beginning of the study.

It is for certain that the EHDS regulation is hitting fast forward [6] and EU wide communities forming for the utilisation and implementation of HealthData@EU [6], [42]. The European Health Data Space aims to enhance the availability and reusability of health data, to empower researchers, policy advisors and all data users. European Commission believes and have the uttermost confident that HealthData@EU strengthens the adoption of data-driven solutions, artificial intelligence being one of them. Also, better health outcomes and preventive medicine taking more impactful steps alongside with metadata management. [6].

Greenberg, et al (2023) views metadata as an important tool for organisations and companies looking to implement artificial intelligence. Also, they provide definition to metadata. Metadata provides context and structure for data, facilitating its discovery, access and reuse. Growing interest in metadata. Fueled by digital information, open access and FAIR data principles, has accelerated its adoption in research data management.[19]. This author has to agree to these authors, metadata has immense power to reshape today's data management practices, giving data owners a better insight into the data they have and enabling a move towards more informed data-driven decisions. More informed decisions are a driving force for saving money and enabling the integration of newer technologies into today's outdated processes.

To point out that both the Population Health Development Plan, prepared by the Ministry of Social Affairs [8], and the Digital Society Development Plan [7] foresee data-driven decisions by 2030, why not take the opportunity for TEHIK to implement metadata management principles in its own work?

Amadi, et al (2024) alongside with Welter, at al (2023) believes that metadata management in machine readable form and making available to data users is beneficial to all citizens [3], [17]. Although, they have rather different approaches to achieving machine readable data descriptions and publishing them. One used DATS model for the data catalogue and handles semantic definitions via JSON-LD forms. [3]. Amadi, et al (2024) approached on the other hand, relatively, more known way, as they used OMOP model for data alignment and mappings and for FAIRifying metadata standards for example DCAT, schema.org, Dublin Core, DDI and lots more. Still after the research, they state that in order to more FAIRifing results the adoption of schema.org and JSON-LD still needs to be considered.[17].

This author agrees with the researchers abovementioned, that the standard used in data catalogue needs to accommodate health data and possibly research datasets as well and these need to be mapped altogether. As Ross, et al (2023) pointed out that Estonia needed a tool that would make better use of the data sets that have been used, either in analytics or in research [9]. However, the results of this research have shown that healthDCAT-AP is not able to keep datasets semantically coherent at this level, which means that TEHIK probably needs to rethink its choice of metamodel standard. Although the hypothesis was disproved, it provides an opportunity to design a second iteration of this research and also to think through a selectable standard for a meta-model that would link datasets at specific levels and maintain the context that accompanies the data over a longer time span.

The importance of the study

The regulation of EHDS is waiting for approval from the European Parliament, due in December 2024.[6] The EU commission is actively empowering member states to participate in funding opportunities and develop their digital capacity. The Health and Welfare Information Systems Center being one of the participants in EHDS direct grant and developing its metamodel for metadata management being one of the corner stones in data cataloguing.[14] The member states are looking towards countries who have been

participated in the EHDS pilot project [11], waiting and contributing in TEHDAS2 [42], to be the forerunners in the implementation of EHDS regulation on a national level.

This study can be an example for other member states across Europe and best practice how to develop metamodel for data catalogue or how not to do, because not many countries have developed its metamodel for health data. Across Europe, member states are looking for answers in regards of data cataloguing. The results of this research are relevant to other data scientists or member states, who face the challenges in data catalogue implementation and the implementation of the EDHS regulation.

Lastly, the research is direct input for TEHIK data catalogue and waiting for metamodel to implement, test its suitability for national datasets and setting best practices how others have solved the problem of interoperability in big data setting and the data to be findable, accessible, interoperable and reusable.

Limitations

The study is constrained by limitations inherent in the dataset, specifically immunisation report dataset being the same, that was concluded in 2016 [44] and within the regulation it is soon to be possibly revised. For the metamodel, there are many standards that are enhancing the FAIR principles, thus chosen for this thesis only healthDCAT-AP, for this to be continuously empowered to implement, by the European Commission [6], [13].

In addition, the limited size of the multidisciplinary team restricts the generalizability of findings. Future research should involve more stakeholders to gather broader feedback and validate the metadata model.

Therefore, future work needs to look into combining the metamodel with data warehouse metadata, if needed then implement standard for secondary use in data warehouse. Future research needs to further develop the metamodel for clinical information models and combination of integration models, whether the metamodel can combine all logical data models, being standard agnostic and publish metadata for consumers and automatically extract database data models, for future developments of new services.

6 Summary

The thesis explores ways to develop a meta-model for a data catalogue for managing and sharing health data. The background is that the European Health Data Space (EHDS) regulation aims to create a standardised and secure data sharing infrastructure to support research and policy making. The paper includes a comprehensive literature review of existing data catalogues, metadata standards and their applicability in the health sector.

In the practical part, a draft health data catalogue was created using the healthDCAT-AP standard, and its suitability was assessed based on different use cases, including the needs of researchers, policy advisors and software developers. The results showed that healthDCAT-AP is usable for the creation of a generic data catalogue, but its limitations in terms of semantic coherence of data and technical integrability require further development.

The study emphasises the need to link metamodel to database structures and to provide metadata management solutions that allow easier access and greater re-use by data users. The resulting proposals are directly applicable to the future development of the Estonian Health and Welfare Information Systems Centre (TEHIK) and will contribute to the implementation of EHDS regulations in Estonia.

References

[1] the International Organization for Standardization and the International Electrotechnical Commission, “ISO/IEC 11179-1:2023,” https://www.iso.org/obp/ui/#iso:std:iso-iec:11179:-1:ed-4:v1:en.

[2] DAMA International, *The Data Management Body of Knowledge (DAMA-DMBOK)*, 2nd ed. Bradley Beach, NJ: Technics Publications, LLC, 2017.

[3] D. Welter *et al.*, “The Translational Data Catalog - discoverable biomedical datasets,” *Scientific Data 2023 10:1*, vol. 10, no. 1, pp. 1–9, Jul. 2023, doi: 10.1038/s41597-023-02258-0.

[4] E. Papadopoulou and E. Sakkopoulos, “Semantic cataloging of public services using basic government vocabularies and the data catalog vocabulary for a unified European digital market,” in *14th International Conference on Information*, 2023. doi: 10.1109/IISA59645.2023.10345951.

[5] O. Olesen-Bagneux, *The Enterprise Data Catalog*. O’Reilly, 2023. Accessed: Dec. 10, 2024. [Online]. Available: https://books.google.ee/books?hl=en&lr=&id=9SWuEAAAQBAJ&oi=fnd&pg=PP1&dq=metadata+data+catalogue&ots=LBhsS7mBTK&sig=v-Xe4Y0jssPTCvuvHKnBU6PIV9k&redir_esc=y#v=onepage&q=metadata%20data%20catalogue&f=false

[6] “CORRIGENDUM. REGULATION (EU) 2024 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of on the European Health Data Space and amending Directive 2011/24/EU and Regulation (EU) 2024/2847,” Nov. 27, 2024, *THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION, Brussels*. Accessed: Dec. 09, 2024. [Online]. Available: https://www.europarl.europa.eu/meetdocs/2024_2029/plmrep/COMMITTEES/ENVI/DV/2024/12-04/2022_0140COR01_EN.pdf

[7] Majandus- ja Kommunikatsiooniministeerium, “Eesti Digiühiskond 2030,” 2021, Accessed: Oct. 25, 2024. [Online]. Available: https://www.mkm.ee/en/e-state-and-connectivity/digital-agenda-2030

[8] Sotsiaalministeerium, “Rahvastiku tervise arengukava 2020-2030,” 2021.

[9] P. Ross, J. Metsallik, K. J. I. Kankainen, I. Bossenko, C. Mäe, and M. Maasik, “Health Sense: universaalse andmemudeli ja raviteekondade järjepidevuse standardi väljatöötamine lähtudes rahvusvahelistest uue põlvkonna terviseinfosüsteemide standarditest,” 2023. Accessed: Oct. 25, 2024. [Online]. Available: https://digikogu.taltech.ee/et/item/161b7131-f2f4-4022-9d6e-c50fde38b224

[10] M. Hildebrandt, “Ground-Truthing in the European Health Data Space,” in *The 16th International Joint Conference on Biomedical Engineering Systems and Technologies*, Science and Technology Publications, 2023, pp. 15–22. doi: 10.5220/0011955900003414.

[11] “HealthData@EU Pilot project .” Accessed: Dec. 09, 2024. [Online]. Available: https://ehds2pilot.eu/

[12] P. Derycke, C. A. Vande Catsyne, T. Korsgaard, and H. A. Huru, “Technical working group on the transition from existing metadata templates to HealthDCAT-AP,” 2024. Accessed: Dec. 09, 2024. [Online]. Available: https://ehds2pilot.eu/wp-content/uploads/2024/04/HealthData@EU-Pilot_MS6.2_Technical-working-group-on-the-transition-from-existing-metadata-templates-to-HealthDCAT_FIN-1.pdf

[13] P. Derycke, “HealthDCAT-AP. Unofficial draft,” Dec. 2023. Accessed: Oct. 08, 2024. [Online]. Available: https://healthdcat-ap.github.io/

[14] TEHIK, “EST2EHDS.” Accessed: Dec. 11, 2024. [Online]. Available: https://teabekeskus.tehik.ee/et/projekt/est2ehds

[15] E. Tacconelli *et al.*, “Challenges of data sharing in European Covid-19 projects: A learning opportunity for advancing pandemic preparedness and response,” *The Lancet Regional Health - Europe*, vol. 21, p. 100467, 2022, doi: 10.1016/j.

[16] G. Tsueng *et al.*, “Developing a standardized but extendable framework to increase the findability of infectious disease datasets,” *Scientific Data 2023 10:1*, vol. 10, no. 1, pp. 1–13, Feb. 2023, doi: 10.1038/s41597-023-01968-9.

[17] D. Amadi *et al.*, “Making Metadata Machine-Readable as the First Step to Providing Findable, Accessible, Interoperable, and Reusable Population Health Data: Framework Development and Implementation Study,” *Online J Public Health Inform*, vol. 16, 2024, doi: 10.2196/56237.

[18] R. Hussein *et al.*, “Getting ready for the European Health Data Space (EHDS): IDERHA’s plan to align with the latest EHDS requirements for the secondary use of health data,” *Open Research Europe*, vol. 4, 2024, doi: 10.12688/OPENRESEUROPE.18179.1.

[19] J. Greenberg, M. F. Wu, and W. Liu, “Metadata as Data Intelligence,” *Metadata as Data Intelligence. Data Intelligence*, vol. 5, 2023, doi: 10.1162/dint_e_00212.

[20] Z. Yang *et al.*, “Defining health data elements under the HL7 development framework for metadata management,” *J Biomed Semantics*, vol. 13, no. 1, pp. 1–15, Dec. 2022, doi: 10.1186/S13326-022-00265-5/FIGURES/6.

[21] R. Albertoni, D. Browning, S. Cox, A. N. Gonzalez-Beltran, A. Perego, and P. Winstanley, “The W3C Data Catalog Vocabulary, Version 2: Rationale, Design Principles, and Uptake,” *Data Intell*, vol. 6, no. 2, pp. 457–487, Mar. 2024, doi: 10.1162/DINT_A_00241/2197651/DINT_A_00241.PDF.

[22] Majandus- ja Kommunikatsiooniministeerium, “Andmehalduse tööriistad | Kratid,” 2024. Accessed: Oct. 07, 2024. [Online]. Available: https://www.kratid.ee/tooriistad

[23] Statistikaamet and Majandus- ja Kommunikatsiooniministeerium, “Andmekirjelduse juhis,” 2024. Accessed: Oct. 07, 2024. [Online]. Available: https://www.kratid.ee/_files/ugd/980182_87343e9f929143b1a566d1c96791ec09.pdf

[24] “Eesti avaandmete teabevärav,” Riigi Infosüsteemide Asutus. Accessed: Oct. 07, 2024. [Online]. Available: https://avaandmed.eesti.ee/

[25] B. Oliveira, A. Duarte, and Ó. Oliveira, “Towards a Data Catalog for Data Analytics,” *Procedia Comput Sci*, vol. 237, pp. 691–700, Jan. 2024, doi: 10.1016/J.PROCS.2024.05.155.

[26] “DAX Guide,” SQLBI. Accessed: Oct. 07, 2024. [Online]. Available: https://dax.guide/

[27] R. Esbai, S. Hakkou, and M. A. Habri, “Modeling and automatic generation of data warehouse using model-driven transformation in business intelligence process,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 30, no. 3, pp. 1866–1874, 2023, doi: 10.11591/ijeecs.v30.i3.pp1866-1874.

[28] “Schema.org documentation,” V28.1. Accessed: Sep. 24, 2024. [Online]. Available: https://schema.org/docs/gs.html#schemaorg

[29] F. Croce, R. Valentini, M. Maranghi, G. Grani, M. Lenzerini, and R. Rosati, “Ontology-Based Data Preparation in Healthcare: The Case of the AMD-STITCH Project,” *SN Comput Sci*, vol. 5, no. 4, pp. 1–12, Apr. 2024, doi: 10.1007/S42979-024-02757-W/FIGURES/3.

[30] D. I. Bourdas, P. Bakirtzoglou, A. K. Travlos, V. Andrianopoulos, and E. Zacharakis, “Analysis of a comprehensive dataset: Influence of vaccination profile, types, and severe acute respiratory syndrome coronavirus 2 re-infections on changes in sports-related physical activity one month after infection,” *Data Brief*, vol. 51, p. 109723, Dec. 2023, doi: 10.1016/J.DIB.2023.109723.

[31] “RIVM data.” Accessed: Oct. 17, 2024. [Online]. Available: https://data.rivm.nl/meta/srv/dut/catalog.search#/metadata/205d0bf4-b645-4e5b-84bc-f8ec482fd3f3

[32] T. Brunson *et al.*, “VIGET: A web portal for study of vaccine-induced host responses based on Reactome pathways and ImmPort data,” *Front Immunol*, vol. 14, p. 1141030, Mar. 2023, doi: 10.3389/FIMMU.2023.1141030/BIBTEX.

[33] H. Müller *et al.*, “BIBBOX, a FAIR toolbox and App Store for life science research,” *N Biotechnol*, vol. 77, pp. 12–19, Nov. 2023, doi: 10.1016/J.NBT.2023.06.001.

[34] M. Saso, N. Schutte, and B. Moreau De Lizoreux, “Final report”, Accessed: Oct. 17, 2024. [Online]. Available: www.phiri.eu

[35] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, “A design science research methodology for information systems research,” *Journal of Management Information Systems*, vol. 24, pp. 45–77, Dec. 2007, doi: 10.2753/MIS0742-1222240302.

[36] B. van Wee and D. Banister, “Literature review papers: the search and selection process,” *J Decis Syst*, 2023, doi: 10.1080/12460125.2023.2197703/ASSET/56C5165A-2905-4C94-BC5E-4CA58AC89101/ASSETS/IMAGES/TJDS_A_2197703_F0001_B.GIF.

[37] TEHIK, “Teabekeskus | TEHIK,” https://www.tehik.ee/teabekeskus. Accessed: Oct. 28, 2024. [Online]. Available: https://www.tehik.ee/teabekeskus

[38] European Union, “European Interoperability Framework,” 2017. doi: 10.2799/78681.

[39] F. Estupiñán-Romero *et al.*, “TEHDAS assesses data interoperability standards - Tehdas,” Dec. 2022. Accessed: Oct. 25, 2024. [Online]. Available: https://tehdas.eu/tehdas1/results/tehdas-assesses-data-interoperability-standards/

[40] E. K. Shriver, “Governance Metadata Standards: Landscape and Gap Analysis Services in Support of Standardizing Governance Metadata for Pediatric COVID-19 Data Linkage,” 2024, Accessed: Oct. 28, 2024. [Online]. Available: https://www.nichd.nih.gov/about/org/od/odss

[41] “Data Catalog Vocabulary (DCAT) - Version 3,” Aug. 22, 2024, *W3C*: 3. Accessed: Sep. 15, 2024. [Online]. Available: https://www.w3.org/TR/vocab-dcat/

[42] Finnish Innovation Fund Sitra, “TEHDAS2 joint action.” Accessed: Oct. 28, 2024. [Online]. Available: https://tehdas.eu/project/

[43] “Program - EHIN 2024,” https://ehin.no/2024/program/.

[44] *Sotsiaalministri 17. septembri 2008. a. määruse nr 53 „Tervise infosüsteemi edastatavate dokumentide andmekoosseisud ning nende säilitamise tingimused ja kord“ ja sotsiaalministri 18. septembri 2008. a. määruse nr 56 „Tervishoiuteenuse osutamise dokumenteerimise ning nende dokumentide säilitamise tingimused ja kord“ muutmine*. 2005. Accessed: Oct. 17, 2024. [Online]. Available: https://www.riigiteataja.ee/akt/122062016020

Appendix 1 – Non-exclusive licence for reproduction and publication of a graduation thesis[1]

I Scharlett Hansson

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis, supervised by Janek Metsallik, Aivi Saar and Kerli Linna
   1.1. to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;
   1.2. to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.
2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.
3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

02.01.2025

1 The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.

Appendix 2 – The dataset of clinical document immunisation report

Immuniseerimise teatise andmekoosseis

1. Meditsiinidokumendi andmed

    1.1. Dokumendi number

    1.2. Dokumendi konfidentsiaalsus

    1.3. Dokumendi kinnitamise aeg

2. Dokumendi koostaja andmed

    2.1. Tervishoiutöötaja ees- ja perekonnanimi

    2.2. Tervishoiutöötaja registreerimistõendi number

    2.3. Tervishoiutöötaja eriala

    2.4. Tervishoiutöötaja kontaktandmed

    2.5. Tervishoiuasutuse nimi

    2.6. Tervishoiuasutuse äriregistri kood

    2.7. Tervishoiuasutuse tegevusloa number

    2.8. Tervishoiuasutuse kontaktandmed

    2.9. Tervishoiuasutuse aadress või konkreetse korpuse (praksise) tegevuskoha aadress

3. Patsiendi andmed

    3.1. Isikukood või tundmatu isiku kood

    3.2. Ees- ja perekonnanimi

    3.3. Sugu

3.4. Sünniaeg

3.5. Tegelik elukoht

3.6. Sünnikoht

3.7. Kontaktandmed

3.8. Töökoha andmed

3.8.1. Asutuse äriregistri kood

3.8.2. Asutuse nimi

3.8.3. Amet

3.9. Õppeasutuse andmed

3.9.1. Õppeasutuse äriregistri kood

3.9.2. Õppeasutuse nimi

3.9.3. Õppeasutuse aadress

3.10. Patsiendi perearsti andmed

3.10.1. Perearsti ees- ja perekonnanimi

3.10.2. Tervishoiutöötaja registreerimistõendi number Terviseameti registri järgi

3.10.3. Tervishoiuasutuse äriregistri kood

3.10.4. Tervishoiuasutuse kontaktandmed

3.10.5. Tervishoiuasutuse aadress

3.11. Muude osaliste (eeskostja, lapsevanem) andmed

3.11.1. Isikukood

3.11.2. Ees- ja perekonnanimi

3.11.3. Seos patsiendiga

3.12. Patsiendi kontaktisiku(-te) andmed

3.12.1. Isikukood

3.12.2. Ees- ja perekonnanimi

3.12.3. Seos patsiendiga

3.12.4. Kontaktandmed

4. Anamnees

4.1. Anamnees

4.2. Kaebused

4.3. Sotsiaalsed olud

4.4. Terviseriskid

5. Immuniseerimise andmed

5.1. Immuniseerimise kuupäev

5.2. Mille vastu immuniseeriti

5.3. Immuniseerimisel manustatud annus ja preparaat

5.3.1. Immunpreparaadi ATC kood ja toimeaine(te) nimetus(ed)

5.3.2. Immuunpreparaadi nimetus

5.3.3. Partii number

5.3.4. Manustatud annus

5.3.5. Manustamise kordsus

5.4. Immuniseerimise kuuri andmed

5.4.1. Järgmise immuniseerimise kuupäev

5.4.2. Märge immuniseerimise lõpetamise kohta

5.5. Immuniseerija andmed

5.5.1. Tervishoiutöötaja ees- ja perekonnanimi

5.5.2. Tervishoiutöötaja registreerimistõendi number

5.5.3. Tervishoiutöötaja eriala

5.6. Kõrvalnähtude andmed

5.6.1. Kõrvalnähu diagnoos RHK-10 järgi

5.6.2. Kõrvalnähtude ilmnemise kuupäev

5.6.3. Kõrvalnähu diagnoosija andmed

5.6.3.1. Tervishoiutöötaja ees- ja perekonnanimi

5.6.3.2. Tervishoiutöötaja registreerimistõendi number

5.6.3.3. Tervishoiutöötaja eriala

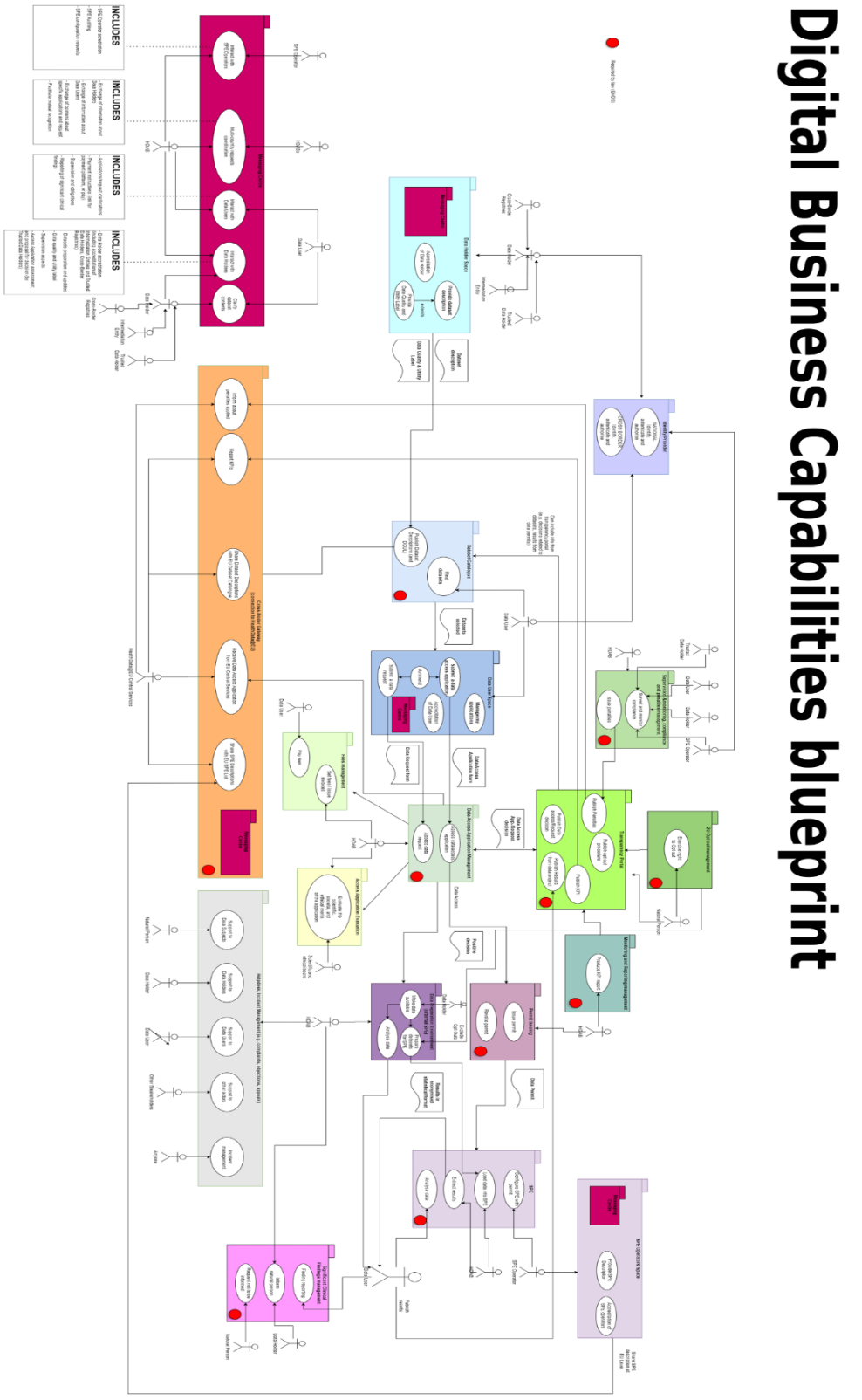
6. Tuberkuliin/Mantoux test

6.1. Testi kuupäev

6.2. Testi tulemus

7. Märkused

Appendix 3 – The proposed blueprint for HealthData@EU by European Commission



Appendix 4 – The described immunisation service dataset in healthDCAT-AP standard

Property	Card.	Definition	Immunisation data exchange template
Agent			
Name	1..*	A name of the agent.	dct:publisher [a foaf:Organization; foaf:name "Tervise ja Heaolu Infosüsteemide Keskus"; foaf:mbox <mailto:info@tehik.ee>; foaf:homepage <https://www.tehik.ee/ >];
URL	1..1	A webpage that either allows to make contact (i.e. a webform) or the information contains how to get into contact.	
email	1..1	A email address via which contact can be made. This property SHOULD be used to provide the email address of the Agent, specified using fully qualified mailto: URI scheme [RFC6068]. The email SHOULD be used to establish a communication channel to the agent.	
Catalogue			
applicable	1..*	The legislation that mandates the creation or	dcatap:applicableLegislation <http://data.europa.eu/eli/reg/2022/868/oj>;

legislation		management of the Catalog.	EHDS; Riigi infosüsteemi haldussüsteem; Tervise infosüsteemi andmekoosseisud ja nende esitamise tingimused
description	1..*	A free-text account of the Catalogue	dct:description "TEHIKu andmekataloogis kirjeldatakse keskse tervise infosüsteemi kogutavad teenuste kirjeldused ning publitseeritakse teabekeskuses ja andmete teabeväravas."@et;
publisher	1	An entity (organisation) responsible for making the Catalogue available.	dct:publisher [a foaf:Organization; foaf:name "Tervise ja Heaolu Infosüsteemide Keskus"; foaf:mbox <mailto:info@tehik.ee>; foaf:homepage <https://www.tehik.ee/ >];
title	1..*	A name given to the Catalogue.	dct:title "TEHIKu andmekataloog@et
Catalogue Record			
modification date	1	The most recent date on which the Catalogue entry was changed or modified.	06 Dec 2024 dct:modified "2024-12-06T18:47:54Z"^^<http://www.w3.org/2001/XMLSchema#dateTime>;
primary topic	1	A link to the Dataset, Data service or Catalog described in the record.	Health related immunisation dataset Immuniseerimise teatis (1.0) - Teabekeskus

Concept			
preferred label	1..*	A preferred label of the concept.	skos:prefLabel "immuniseerimisteatis" @et; skos:prefLabel "immunisation report" @en-gb
Dataset			
access rights	1..1	Information that indicates whether the Dataset is publicly accessible, has access restrictions or is not public.RDF example: dct:accessRights OpenData Protected Data, Sensitive Data	Open Data
applicable legislation	1..*	The legislation that mandates the creation or management of the Dataset. RDF example: dcatap:applicableLegislation	Riigi infosüsteemi haldussüsteem; Tervise infosüsteemi andmekoosseisud ja nende esitamise tingimused; avaliku teabe seadus
dataset distribution	1..*	An available Distribution for the Dataset. RDF example: dcat:distribution	dcat:distribution [a dcat:Distribution ; dcatap:applicableLegislation <http://data.europa.eu/eli/reg/2022/868/oj>; dcat:accessURL <https://teabekeskus.tehik.ee/et/vormingud/immuniseerimise-teatis/1.0> ; dct:format <http://publications.europa.eu/resource/authority/file-type/XML>; dcat:byteSize "800000"; dct:rights [a dct:RightsStatement; rdfs:label "Publicly available"@en];

			dcat:mediaType <http://www.iana.org/assignments/media-types/text/tab-separated-values>];
description	1..*	A free-text account of the Dataset. RDF example: dct:description	The EST2EHDS project needs to describe immunisation dataset, in order to ensure structured data collection for immunisation as a procedure and monitoring for data quality purposes. This dataset includes metadata, immunisation procedure (along with adverse effects, manufacturer, dose quantity).
geographical coverage	1..*	A geographic region that is covered by the Dataset. RDF example: dct:spatial	dcterms:spatial < N 59° 0′ 0" ;E 26° 0′ 0"> ; Estonia
health category	1..*	The health category to which this dataset belongs as described in the Commission Regulation on the European Health Data Space laying down a list of categories of electronic data for secondary use, Art.33. RDF example:	healthdcatap:healthCategory <http://healthdata.ec.europa.eu/EHR>;

		healthdcatap:healthCategory	
health data access body	1..1	Health Data Access Body supporting access to data in the Member State. RDF example: healthdcatap:hdab	healthdcatap:hdab [a foaf:Organization; locn:address [a locn:Address; locn:adminUnitL1 "EST"; locn:postCode "11317"; locn:postName "Pärnu mnt 132"]; foaf:mbox <mailto:info@tehik.ee>; foaf:homepage <https://www.tehik.ee/>; foaf:name "Health and Welfare Information Systems Center];
identifier	1..*	The main identifier for the Dataset, e.g. the URI or other unique identifier in the context of the Catalogue. RDF example: dct:identifier	https://akk.tehik.ee/public_data_structure/immuniseerimise-teatis/1.0
minimum typical age	1..1	Minimum typical age of the population within the dataset RDF example: healthdcatap:minTypicalAge	healthdcatap:minTypicalAge: "0";
maximum typical age	1..1	Maximum typical age of the population within the dataset RDF example:	healthdcatap:maxTypicalAge: "100";

		healthdcatap:maxTypicalAge	
number of records	1..1	Size of the dataset in terms of the number of records. RDF example: healthdcatap:numberOfRecords	healthdcatap:numberOfRecords: "124866488";
Number of records for unique individuals.	1..1	Number of records for unique individuals. RDF example: healthdcatap:numberOfUniqueIndividuals	healthdcatap:numberOfUniqueIndividuals: "8914722";
modification date	0..1	The most recent date on which the Dataset was changed or modified. RDF example: dct:modified	25 Nov 2024 dct:modified "2024-11-25T18:47:54Z"^^<http://www.w3.org/2001/XMLSchema#dateTime>;
population coverage	1..*	A definition of the population within the dataset RDF example: healthdcatap:populationcoverage	healthdcatap:populationCoverage "The population targeted by the immunisation service considers in general Estonian population "@en;
provenance	1..*	A statement about the lineage of a Dataset. RDF example: dct:provenance	dct:provenance [a dct:ProvenanceStatement; rdfs:label "The data for the immunisation service has been collected from Estonian healthcare providers,

			validated against data quality rules and schema rules, stored in central data warehouse"@en];
publisher	1..1	An entity (organisation) responsible for making the Dataset available. RDF example: dct:publisher	<rdf:Description rdf:about="http://publications.europa.eu/resource/authority/corporate-body/COUN_ASS_EC_EST"> <skos:inScheme rdf:resource="http://publications.europa.eu/resource/authority/corporate-body"/>
publisher type	0..*	A type of organisation that makes the Dataset available. RDF example: healthdcatap:publishertype	Hospital or Healthcare System Government Health Department
purpose	1..*	A free text statement of the purpose of the processing of personal data. RDF example: dpv:hasPurpose	dpv:hasPurpose [a dpv:Purpose; dct:description "The primary aim for this is to ensure semantic interoperability across Estonian healthcare ecosystem, therefore established immunisation service"@en];
quality annotation	1..*	Dataset, including rating, quality certificate, feedback that can be associated to the dataset. RDF example:	dqv:hasQualityAnnotation [a dqv:QualityCertificate ; oa:hasTarget <https://akk.tehik.ee/public_data_structure/immuniseerimise-teatis/1.0; oa:hasBody <https://acertificateserver.eu/mycert

		dqv:hasQualityAnnotation	ification> ; oa:motivatedBy dqv:qualityAssessment];
sample	1..*	A sample distribution of the dataset. RDF example: adms:sample	adms:sample [a dcat:Distribution ; dct:description "Structural data composition using xml format"@en; dcat:downloadURL <https://akk.tehik.ee/public_data_structure/immuniseerimise-teatis/1.0>; dcat:mediaType "http://publications.europa.eu/resource/authority/file-type/XML"];
theme	1..*	A category of the Dataset. RDF example: dcat:theme	healthdcatap:healthTheme <https://www.wikidata.org/entity/Q58624061>,<https://www.wikidata.org/entity/Q7907952> ;
title	1..*	A name given to the Dataset. RDF example: dct:title	dcat:Dataset, [a dcat:Dataset, dcterms:title "Immunisation service"@en ; dcterms:title "Immuniseerimisteatus"@et] ;
type	1..1	A type of the Dataset. RDF example: dct:type	dct:type <http://publications.europa.eu/resource/authority/dataset-type/OPEN_DATA>
Distribution			
access URL	1..*	A URL that gives access to a Distribution of the Dataset.	dcat:accessURL <https://teabekeskus.tehik.ee/et/vormingud/immuniseerimise-teatis/v/1.0>;

applicable legislation	1..*	The legislation that mandates the creation or management of the Distribution.	Riigi infosüsteemi haldussüsteem; Tervise infosüsteemi andmekoosseisud ja nende esitamise tingimused; avaliku teabe seadus
linked schemas	0..*	An established schema to which the described Distribution conforms.	dcterms:conformsTo <"HL7 versioon: V3-2005-NORMATIVE (HL7 Clinical Document Architecture, Release 2.0)">;
title	0..*	A name given to the Distribution	dcat:Distribution,[a dcat:Distribution: dcterms:title "Immunisation service"@en ; dcterms:title "Immuniseerimisteatus"@et] ;
Relationship			
had role	1..*	A function of an entity or agent with respect to another entity or resource.	dcat:hadRole <>
relation	1..*	A resource related to the source resource.	dcterms:relation <Immunization (IPS) - International Patient Summary Implementation Guide v2.0.0>, <https://opendata.digilugu.ee/docs/#/et/opendata/covid19/vaccination/readme>;
Rights statement			

conditions for access and usage	1..1	This property MUST be used to indicate the conditions if any contracts, licences and/or are applied for the use of the dataset. The conditions are declared on an aggregated level: whether a free and unrestricted use is possible, a contract has to be concluded and/or a licence has to be agreed on to use a dataset.	dct:accessRights <http://publications.europa.eu/resource/authority/access-right/NORMAL>;
Checksum			
algorithm	1	The algorithm used to produce the subject Checksum.	
checksum value	1	A lower case hexadecimal encoded digest value produced using a specific algorithm.	
Data Service			
endpoint URL	1..*	The root location or primary endpoint of the service (an IRI).	dcat:endpointDescription <https://akk.tehik.ee/public_data_structure/immuniseerimise-teatis/1.0>
title	1..*	A name given to the Data Service.	dcat:DataService, [a dcat:DataService:

			dcterms:title "Immunisation service"@en ; dcterms:title "Immuniseerimisteatus"@et];
Dataset Series			
applicable legislation	1..*	The legislation that mandates the creation or management of the Dataset Series.	
description	1..*	A free-text account of the Dataset Series.	dcterms:description <"">
title	1..*	A name given to the Dataset Series.	dcat:DatasetSeries, [a dcat:DatasetSeries: dcterms:title "Immunisation service"@en ; dcterms:title "Immuniseerimisteatus"@et];
Licence document			
type	0..*	A type of licence, e.g. indicating 'public domain' or 'royalties required'.	