

# Government data-driven decision-making (DDDM) framework implementation. Test case: crisis management

Deliverable 1.4: To-be situation report

**Technical Support Instrument**

*Supporting reforms in 27 Member States*



Funded by  
the European Union



REPUBLIC OF ESTONIA  
GOVERNMENT OFFICE

This document was produced with the financial assistance of the European Union. Its content is the sole responsibility of the author(s). The views expressed herein can in no way be taken to reflect the official opinion of the European Union.

The project is funded by the European Union via the Technical Support Instrument, managed by the European Commission Directorate-General for Structural Reform Support.

This report has been delivered in June 2022, under the EC Contract No. REFORM/SC2021/076. It has been delivered as part of the project "Government data-driven decision-making (DDDM) framework implementation. Test case: crisis management".

© European Union, 2024



The Commission's reuse policy is implemented by Commission Decision 2011/833/EU of 12 December 2011 on the reuse of Commission documents (OJ L 330, 14.12.2011, p. 39 – <https://eur-lex.europa.eu/eli/dec/2011/833/oj>).

Unless otherwise noted, the reuse of this document is authorised under the Creative Commons Attribution 4.0 International (CC BY 4.0) licence (<https://creativecommons.org/licenses/by/4.0/>). This means that reuse is allowed, provided that appropriate credit is given and any changes are indicated.

**Directorate-General for Structural Reform Support**  
REFORM@ec.europa.eu  
+32 2 299 11 11 (Commission switchboard)  
European Commission  
Rue de la Loi 170 / Wetstraat 170  
1049 Brussels, Belgium

# Executive summary

## Scope of the Project

This report has been developed within the Project carried out by PricewaterhouseCoopers EU Services EESV (hereinafter – PwC) on behalf of the DG REFORM, according to the specific contract No. REFORM/SC2021/076 (21EE02), signed on 14 October 2021. The report covers the items required in the Request for Service (RfS).

This report covers Outcome 1 of this Project – **Government data-driven decision-making**. Separate reports are issued for Outcome 2 and 3 which all combined comprise the complete package of deliverables.

## Purpose of the Project and Report

### Content of the Report

The report has been drafted for the purpose of describing the to-be version of the data-driven decision-making system and analysing the legal environment of the described system.

The report includes a complemented overview of the system functionalities (main and technical functionalities), as well as the target data model, new organisational and governance structures and business processes, DDDM system architecture, issues to be solved in the processing of different types of data, methodological guidelines, and legal analysis of DDDM system.

### Objective of the DDDM system

The DDDM system envisages the automation of the preparation of decision drafts in a way in which related data and information are gathered, pre-processed, and visualized to make it easier for users and decision-makers to better understand the content. The initial decision draft would be prepared without significant manual user intervention. The role of the user remains to review the offer by DDDM system and, if necessary, correct and supplement it.

The DDDM system makes it possible to speed up the decision-making process, reduce the unintentional or intentional subjectivity of the draft decision preparation, and increase the transparency of the decision-preparation process. At the same time, the automated solution enables the meetings and sessions of the government cabinet to receive data-based answers to questions that arise on an ongoing basis, the answer to which may require separate preparation time today and therefore postpone the decision to the future.

### Expected outcome

The user of the draft decision has at their disposal a system that significantly reduces the time spent searching for information and data related to the preparation of drafts. It supports the inclusion of all relevant data, makes usable information and data more transparent, and creates decision alternatives that are based on objective considerations. At the same time, when developing the tool, it is considered that all technological platforms used in policymaking and proceedings (including co-creation environment, session information system, etc.) must be connected to each other as much as possible.

## Key Findings and recommendations

- This report set out to describe the DDDM system comprehensively, but some parts of the system still need detailed analysis during its early implementation. At the time of the preparation of this report, it was known that the Beneficiary Government Office planned to start developing the system in stages, however the priority order of the development projects was not known. Therefore, when implementing development projects, we recommend carrying out a detailed preliminary analysis that would define specific business requirements and technical conditions regarding the relevant development.
- Develop and establish an open data format standard so that data made available through open data portals are published under agreed terms. It would simplify the start-up stage of the DDDM system and the rapid organisation for data collection. The standard is also necessary for all other data consumers that use open data.
- Need to develop guidelines and tools for data quality control. DDDM must provide users with data to analyse and eliminate significant deficiencies in data quality. Ensuring data quality is the data

holders' responsibility. Their work could be simplified by providing appropriate instructions and tools.

- We recommend considering the creation of aggregated data exchange interfaces with data sources to avoid personal data processing. A similar service exists today at the Statistical Office in Estonia.
- For faster and more effective deployment of data sources, publishing a data profile with metadata should be considered. In this case, the inclusion of data would be better planned, and the selection of datasets would be easier.
- Automatic data modelling is a direction that should be paid more attention to in the future, as modelling is usually done manually today. Modelling here refers to the automatic identification of data (tables) interrelationships, visualisation, and merging of the dataset into an analysis database.
- Generating metadata from an existing database is a technology that has little use today. However, the possibilities for carrying out such processes has already been created and should also be introduced in DDDM.
- Today, there is no widely used repository for analyses in Estonia, where it would be possible to get acquainted with the analyses previously performed on the data and the data source. In terms of analysis tools, there is a multiplicity of tools and, as a rule, the analysis model (or source code) is not transferable from one system to another. In DDDM, it is possible to remedy this situation.
- The use of synthetic data is still not widespread now and there are no registered data sets. To solve certain tasks, the use of synthetic data in DDDM should be considered to reduce the need to process personal data.
- As the data exchange speeds of networks increase, the use of data directly on the server of the data source (data federation) becomes more relevant. The purpose of DDDM is not to form a data warehouse from all available data, which is an expensive activity, but to perform analyses by moving only the results of analyses between servers.
- For every public sector information system, there must be a legal regulation, or a legal basis based on which the system processes the data. The legal status of DDDM also needs to be resolved both to create and maintain the system and to process the data with limited access.

# Lühikokkuvõte

## Aruande eesmärk ja ulatus

Aruanne on koostatud Euroopa Komisjoni struktuurireformide toe peadirektoriaadi (DG REFORM) tellimusel PricewaterhouseCoopers EU Services EESV (edaspidi PwC) poolt läbiviidud Projekti raames vastavalt 14. oktoobril 2021. aastal allkirjastatud lepingule nr REFORM/SC2021/076 (21EE02). Aruande koostamisel on lähtutud Projekti lähteülesandes esitatud nõuetest.

Aruandes kajastatakse ainult Projekti esimese tulemiga piiritletud teemasid – andmepõhise otsustusprotsessi edendamine. Eraldi aruanded väljastatakse Projekti teise ja kolmanda tulemi kohta, mis kokku moodustavad lepingus ettenähtud väljundid.

Käesolev aruanne on koostatud eesmärgiga anda ülevaade Vabariigi Valitsuse otsustusprotsessidesse integreeritavast tehnoloogilisest lahendusest (DDDM süsteem), et andmete kasutamine poliitikakujundamisel oleks läbipaistev, lihtne ja kiire.

## Aruande sisukirjeldus

Aruandes on esitatud kirjeldus DDDM süsteemi tehnilisest ülesehitusest (põhi- ja abifunktsioonid) ja arhitektuurne nägemus, pakutud välja vastutavad organisatsioonid ja rollid ning kujundatud perspektiivsed äriprotsessid. Täiendavalt on analüüsitud õigusraamistikku ning esitatud ettepanekud õigusruumi muudatusteks DDDM süsteemi rakendamisel. Aruandes on esitatud ka andmemudeli ettepanek ja meetodilised juhised, sh andmetötluse automatiseerimist võimaldavate väljakutsete lahendamise nimistu.

## DDDM süsteemi eesmärk

DDDM süsteem näeb ette otsuse-eelnõude ettevalmistamise automatiseerimist viisil, milles seonduvad andmed ja teave koondatakse, need eeltöödeldakse, visualiseeritakse kasutajatele ja otsustajatele lihtsamini hoomatavaks ning koostatakse esmane otsuse projekt ilma kasutaja olulise sekkumiseta. Kasutaja rolliks jääb pakutu ülevaatamine ning vajadusel korrigeerimine ja täiendamine.

DDDM süsteem võimaldab kiirendada otsustusprotsessi, vähendada otsuse-eelnõu ettevalmistamise tahtmatut või tahtlikku subjektiivsust ning tõsta otsuse ettevalmistamise protsessi läbipaistvust. Ühtlasi võimaldab automatiseeritud lahendus valitsuse kabineti nõupidamistel ja istungitel saada andmepõhiseid vastuseid ka jooksvalt tekkivatele küsimustele, millele vastamine võib täna vajada eraldi ettevalmistusaega ja seetõttu otsuse edasilükkamist tulevikku.

## Oodatav tulemus

Otsuse-eelnõude koostajate käsutuses on töövahend, mis oluliselt vähendab ajakulu eelnõude ettevalmistamisega seonduva info ja andmete otsimisel, toetab kõikide asjakohaste andmete kaasamist, muudab läbipaistvamaks teabe ja andmete kasutamise, ja tekitab otsuse alternatiivid, mis lähtuvad objektiivsetest kaalutlustest. Seejuures peetakse töövahendi arendamisel silmas, et kõik poliitikakujundamisel ja menetlemisel kasutatavad tehnoloogilised platvormid (sh koosloome keskkond, istungite infosüsteem jt) peavad olema võimalikult palju üksteisega seostatud.

## Tähelepanekud ja soovitus

- Käesolevas aruandes on püütud DDDM süsteemi kirjeldada küll terviklikult, ent mõningad süsteemi osised vajavad detailanalüüsi selle rakendamise käigus. Aruande koostamise ajal on teada, et Riigikantselei planeerib süsteemi hakata arendama hajusalt ning etapiti, kuid arendusprojektide prioriteetne järjekord ei ole täpselt teada. Seetõttu soovime arendusprojektide rakendamisel viia läbi täpsustav eelanalüüs, mis määratleks spetsiifilised ärinõuded ja tehnilised tingimused asjakohase arenduse asjus.
- Töötada välja ja kehtestada avaandmete formaadi standard, et avaandmete portaalide kaudu kättesaadavaks tehtud andmed oleks avaldatud kokkulepitud tingimustel. See lihtsustaks DDDM süsteemi käivitamist ja andmehõive kiiret korraldamist. Standard on vajalik ka kõikidele teistele andmetarbijatele, mis kasutavad avaandmeid.
- Töötada välja juhised ja töövahendid andmekvaliteedi kontrollimiseks. DDDM peab andma kasutajatele andmed analüüsimiseks ning soovib välistada olulisi puudusi andmete kvaliteedis.

Andmekvaliteedi tagamine on andmete vastutava töötaja kohustus, mistõttu saaks nende tööd lihtsustada asjakohaste juhiste ja töövahendite pakkumise kaudu.

- Soovitav on kaaluda andmeallikatega agregeeritud andmevahetuse liideste loomist vältimaks isikuandmete töötlemist. Sarnane teenus on täna olemas Statistikaametil.
- Andmeallikate kiiremaks ja efektiivsemaks kasutuselevõtuks tuleks kaaluda andmete profiili avaldamist koos metaandmetega. Sellisel juhul oleks andmete kaasamine paremini planeeritav ja andmestike valimine lihtsam.
- Andmete automaatmodelleerimine on suund, millele tasuks edaspidises suuremat tähelepanu pöörata, kuna täna tehakse modelleerimist reeglina käsitsi. Modelleerimise all on siin silmas peetud andmete (tabelite) omavaheliste seoste automaattuvastamist, visualiseerimist ja andmestiku ühendamist analüüsiandmebaasiks.
- Metaandmete genereerimine olemasoleva andmebaasi pealt on tehnoloogia, mida on täna vähe kasutatud. Reaalsed võimalused sellise protsessi teostamiseks on aga juba loodud ja tuleks ka DDDM-is kasutusele võtta.
- Eestis puudub täna laiemalt kasutatav analüüside repositoorium, kust oleks võimalik tutvuda varem andmeallika andmete osas tehtud analüüsidega. Analüüsivahendite osas valitseb tööriistade paljusus ja reeglina ei ole analüüsimudel (või lähtekood) ühest süsteemist teise ülekantav. DDDM-is on võimalik seda olukorda parandada.
- Sünteetiliste andmete kasutamine on hetkel veel vähe levinud ja puuduvad registreeritud andmestikud. Teatud ülesannete lahendamiseks tasuks sünteetiliste andmete kasutamist DDDM-is kaaluda, et vähendada isikuandmete töötlemise vajadust.
- Võrkude andmevahetuskiiruste kasvades muutub üha aktuaalsemaks andmete kasutamine otse andmeallika serveril (*data federation*). DDDM-i eesmärk ei ole moodustada kõikidest saadaval olevatest andmetest andmeladu, mis on väga kulukas tegevus, vaid teha analüüse võimalikult õhukeses lahenduses liigutades serverite vahel peamiselt ainult analüüsile tulemusi.
- Iga avaliku sektori infosüsteemi jaoks peab olemas olema õiguslik regulatsioon või õiguslik alus, mille alusel süsteem andmeid töötleb. Ka DDDM-il tuleb õiguslik staatus lahendada nii süsteemi loomiseks kui piiratud ligipääsuga andmete töötlemiseks.

# Table of Contents

- 1. Introduction..... 9**
  - 1.1 Scope of the Report.....9
  - 1.2 Project Timeline .....11
  
- 2. DDDM System Functionality ..... 12**
  - 2.1 Main Functionalities of the Technical Solution .....13
  - 2.2 Support Functionalities of the Technical Solution .....26
  
- 3. Governance and Business Processes ..... 28**
  - 3.1 Governance Structure.....28
  - 3.2 Business Processes .....30
  - 3.3 Legal Analysis.....32
  
- 4. Target Data Model..... 35**
  - 4.1 Data Structure Types.....35
  - 4.2 Issues to be solved in the Data Structure.....37
  - 4.3 Logical Data Model of the DDDM Central Part.....37
  - 4.4 Data Integration Data Model of the DDDM System.....38
  - 4.5 Analysis Data Model of the DDDM system.....38
  
- 5. DDDM System Architecture ..... 39**
  - 5.1 System Context .....39
  - 5.2 System Components .....39
  - 5.3 Data Search Component Model .....41
  - 5.4 Data source Interfaces.....43
  
- 6. Issues to be solved in the processing of different types of data ..... 44**
  - 6.1 Catalogue of Issues.....44
  - 6.2 Establishing aggregated Data API-s for Data Sources .....47
  - 6.3 Synthetic Data .....47
  
- 7. Methodological Guidelines ..... 48**
  - 7.1 Obtaining Data Access .....48
  - 7.2 Data Processing Methodology.....48
  - 7.3 Data Analysis Methodology Cornerstones .....49
  - 7.4 Data Usage of Data Sources Data .....49
  - 7.5 Sticking Visualisations into a Document.....49

<b>8. Appendices .....</b>	<b>50</b>
8.1 Extended Legal Analysis in Estonian .....	50
8.2 List of conducted Interviews .....	50



# 1. Introduction


## 1.1 Scope of the Report

### 1.1.1 Purpose of the Report

The report has been drafted for the purpose of describing the to-be version of the data-driven decision-making system and analysing the legal environment of the described system.

The report includes a complemented overview of the system functionalities described in Deliverable 1.3, as well as the target data model, governance, and business processes, DDDM system architecture, issues to be solved in the processing of different types of data, methodological guidelines, and legal analysis of DDDM system.

The approach and results of the topics are described in the respective paragraphs.



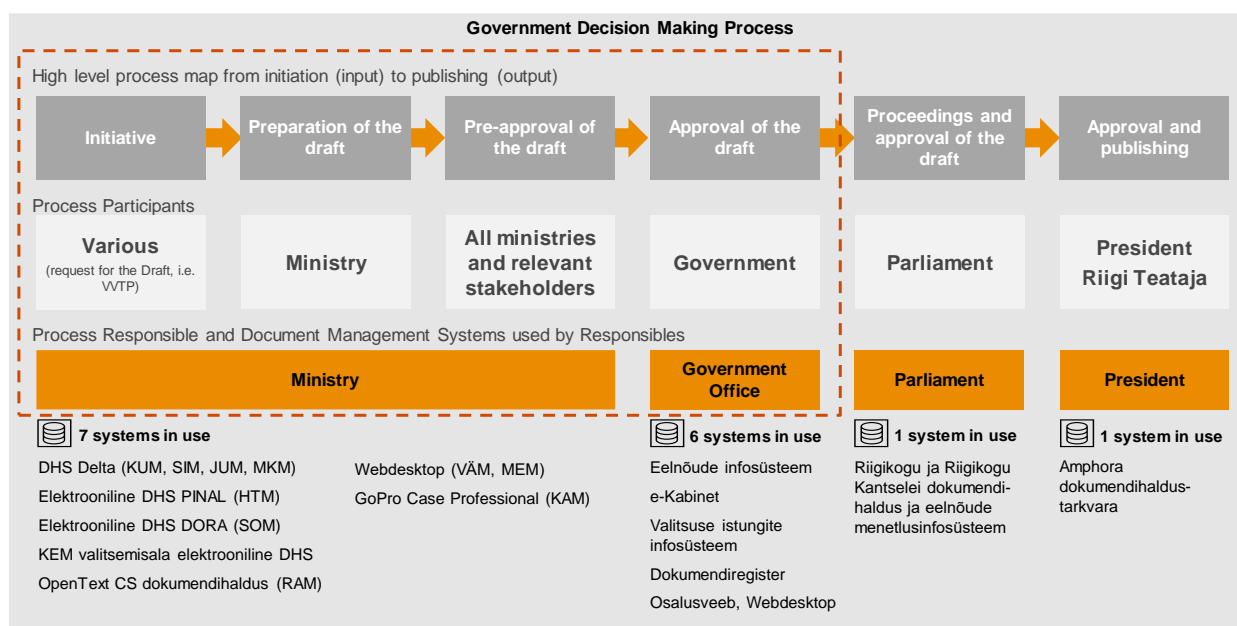
**This report covers only Outcome 1** – Government data-driven decision-making framework implementation. Separate reports are issued for Outcome 2 and 3 which all combined comprise the complete package of deliverables.

### 1.1.2 Scope of Outcome 1

The decision-making process in general involves several institutions (Ministry, Government Office, Parliament, President) as illustrated in Figure 1. As there are many different types of legal documents and decisions in Estonia (described in Deliverable 1.1.), the level, scope and course of the decision-making processes vary.

**The Project Scope** approved in Deliverable 1.1. covers the areas of responsibility of Ministries and Government Office as shown in Figure 1.

Figure 1. Scope of the Project by Institutions in Outcome 1



It was acknowledged that certain types of Documents are submitted to the Parliament for proceedings and approvals, and Legal Drafts are sent to the President for announcement and publication in the Riigi Teataja.

However, considering the purpose of the Project, **the working process and practices at the Parliament and the President are not covered.**

In summary, the Project Scope encompasses the following:

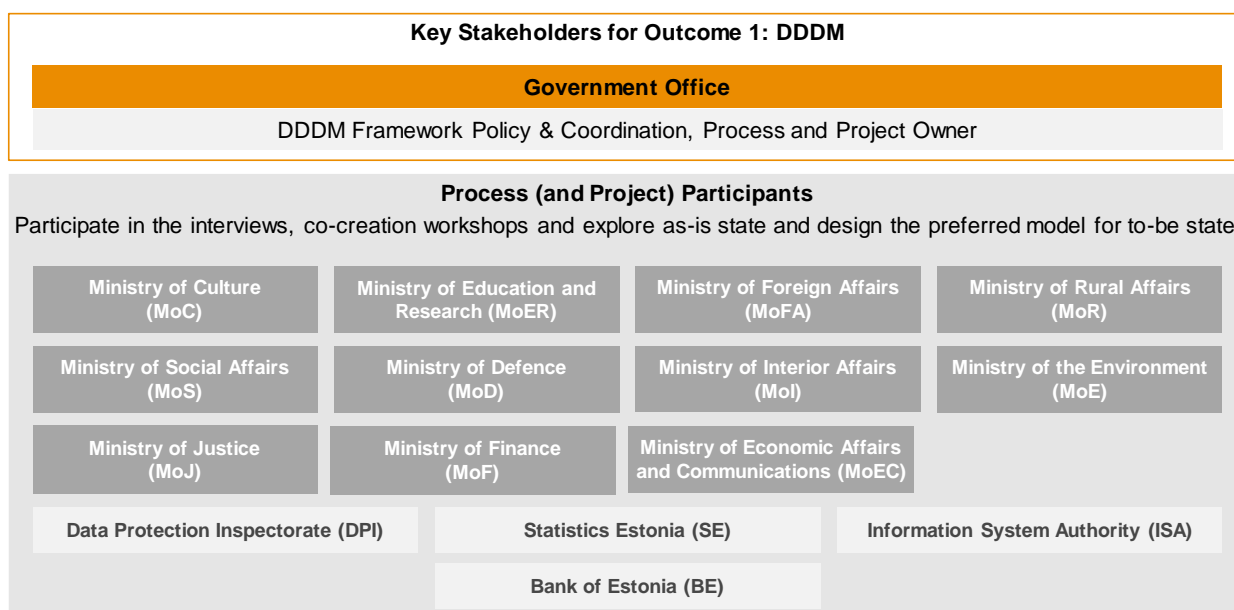
Table 1. Project Scope

Area	Description
<b>1. Institutions</b>	Process Responsible: <ul style="list-style-type: none"> <li>• Ministries</li> <li>• Government Office</li> </ul>
<b>2. Document Type</b>	<ul style="list-style-type: none"> <li>• Government Memorandum</li> </ul>
<b>3. Process</b>	<ul style="list-style-type: none"> <li>• End-to-End process of Government Memorandum</li> <li>• End-to-end describes a process that covers the process from beginning to end and provides a complete output for Government decision-making</li> </ul>
<b>4. Data and Technology</b>	<ul style="list-style-type: none"> <li>• Data and Technology used in the process of Government Memorandum</li> </ul>
<b>5. People</b>	<ul style="list-style-type: none"> <li>• Participants and decision-makers such as public servants and/or third parties (i.e., subject matter experts) involved in the process of Government Memorandum</li> </ul>

### 1.1.3 Project Stakeholders for Outcome 1

To conduct an effective stakeholder engagement, the following key stakeholders and process participants for the Outcome 1 (Figure 2) that are participating in the Project work have been identified.

Figure 2. Outcomes 1: Key Stakeholders and Project Participants

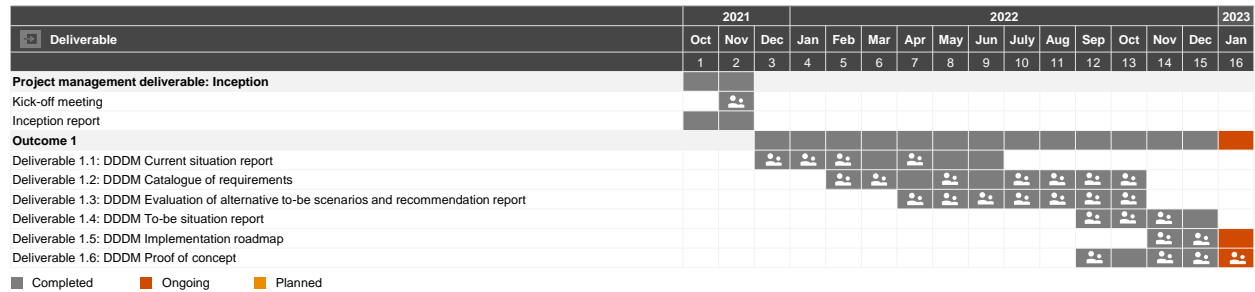


## 1.2 Project Timeline

### 1.2.1 Timeline

Figure 3 provides a high-level overview of the project activities and timeline. The activities of the fourth deliverable took place from September 2022 to December 2022.

Figure 3. Project Activities and Timeline



The draft deliverable was issued for review on November 18, 2022, and since then the deliverable was mostly modified and updated based on the comments and suggestions from key stakeholders such as the DG Reform, OECD, the Beneficiary and Statistics Estonia.

## 2. DDDM System Functionality

The DDDM system as a data-driven automated system for supporting the Estonian Government decisions must have both an automated backend system and a modern web-based user interface. It is reasonable to use an existing system as a user interface or extension of the DDDM system, such as KOOS<sup>1</sup>.

All activities related to data-driven decision-making must be automated to the greatest extent. Automation is possible in areas where there is enough data and digital knowledge. Digital knowledge is fuel for Artificial Intelligence (AI). The knowledge can advance through machine learning on **high-quality data**. The goal is to create a fully automated system where the government can use descriptive, predictive, and prescriptive analytics to accelerate high-quality decision-making processes.

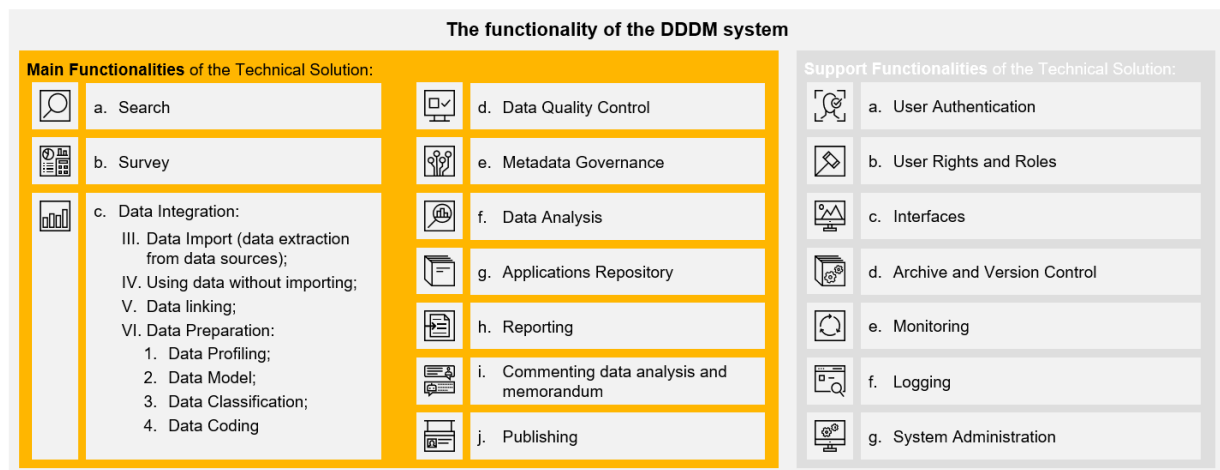


Figure 4. Government Memorandum 2.0 – DDDM System Architecture

As shown in Figure 4, the DDDM system functionality is divided into two parts:

1. Main functionality.
2. Support functionality.

Figure 5 details the main functionalities directly related to data analysis. The DDDM system has two sides. The first side is drafting the texts of memorandums, and the second side is the data analysis subsystem. The simplest way of data analysis is shown in the figure below.

<sup>11</sup> <https://www.just.ee/oigusloome-arendamine/riigi-koosloome-keskkond>

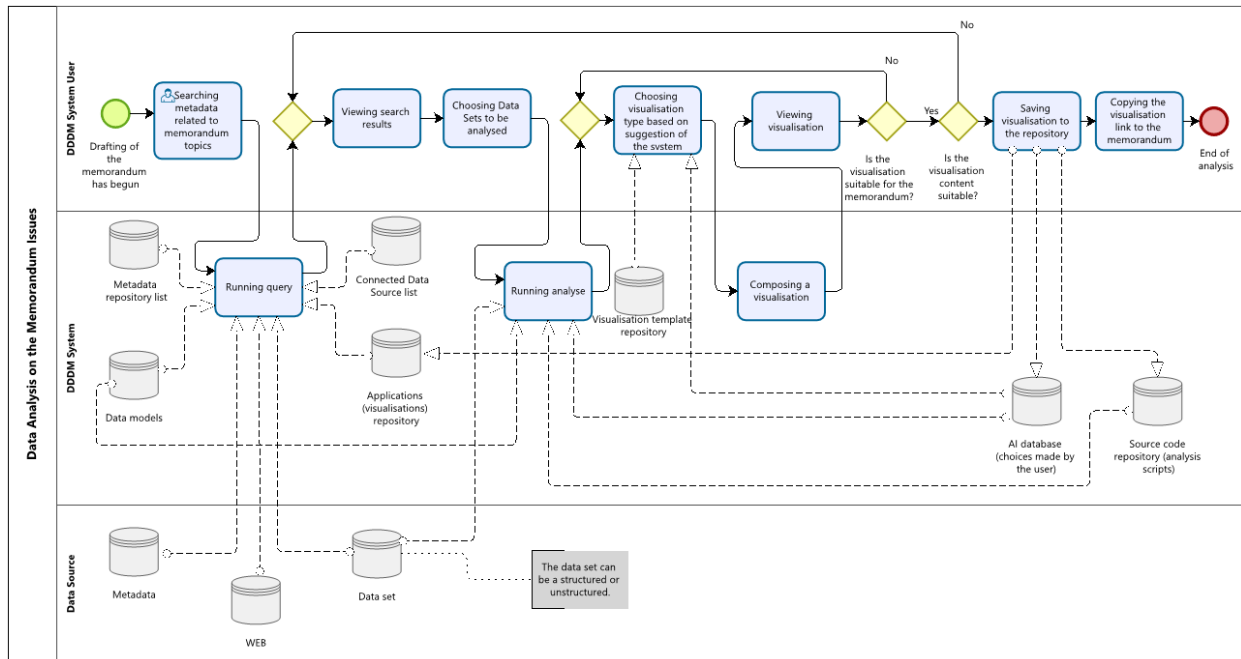


Figure 5. The Simplest Data Analysis Workflow in the DDDM system

There is a variety of different workflows in the system. One of the most important aspects of data analysis is data modelling, closely related to data integration. Data models are descriptions of how the data in a data source can be linked, and what main columns or pieces of the data should be used. Data sources can be any relevant metadata repository (e.g. Public Data Gateway, in estonian *andmete teabevärv*), web service and/or data set.

Additionally, a data quality check and data profiling workflow should be established to ensure quality output from the DDDM data analysis subsystem. The following chapters describe the functionalities that need to support the DDDM system analysis process in the future.

## 2.1 Main Functionalities of the Technical Solution

### 2.1.1.1 Data categories and features

The DDDM system must support data usage for different data categories.

Data can be divided into the following **main (first-level) categories**:

1. Data.
2. Metadata.

The second-level categories are:

1. Structured data.
2. Unstructured data.

Both the data and the metadata can be either structured or unstructured.

The third-level categories are:

1. Aggregated data.
2. Microdata<sup>2</sup>.

<sup>2</sup> EUROSTAT microdata definition: <https://ec.europa.eu/eurostat/web/microdata>

**Data features** are:

1. The data is provided with metadata. Additionally, many data sources can provide partial metadata or metadata that is not up to date.
2. A dataset can be accessed, retrieved, and used on the server of a data holder (e.g. SQL server), or the data can be accessed but must be downloaded, stored, and used locally (on-premises).
3. The data can be either cleansed or non-cleansed. The non-cleansed data can be referred to as raw data.
4. The data can be either classified or non-classified.
5. The data can be either coded or non-coded.
6. The data format can be either machine-readable or non-machine readable.
7. The data structure information can be machine-readable (e.g. a database DDL script with descriptions of data types, lengths, mandatory attributes, and relationships), if provided, or non-machine readable (e.g. the data is in a CSV file that has no structure definition).

### 2.1.1.2 Data protection

A crucial issue of the DDDM system is data protection. Without data protection features, the system cannot exist. There are various levels of data security measures:

1. Legal level.
2. Organisational level.
3. Technical level.

The use of data must be grounded on the rule of law. The DDDM system must have law-based or agreement-based permission to use the data. An option for resolving this issue is the use of the research environment for data processing and analysis of Statistics Estonia. Statistics Estonia has already established a legal environment. The second option is to design and create legal and technical environment where the prevention of infringement of individual rights is addressed through different technical measures.

The DDDM system must have an IT organisation that has the capability to monitor data usage and a user authorisation system.

### 2.1.1.3 Machine learning on the user activity

The system should consider every functionality utilised by the user in order to make suggestions to the user based on previous user's analysis, e.g. recommending dimensions to be analysed based on previous user's analysis.

### 2.1.1.4 Search

The data search functionality is the first element required for the efficient support of the DDDM users.

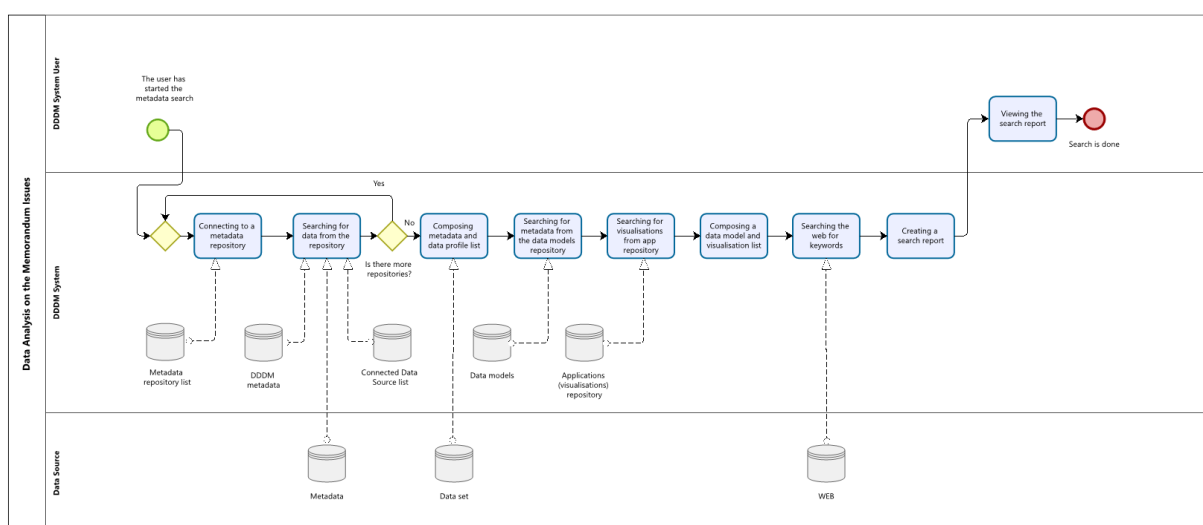


Figure 6. The Data Search Workflow in the DDDM system

The data can be found in numerous data sources. However, finding the proper one is impossible in most cases due to the lack of accurate metadata. If the data user is not familiar with the data content in the data source, the results of the analysis may potentially lead to misguided decisions in the government. The consequences of the decision will come quietly due to an imperceptible misunderstanding of the data profile.

In order to avoid such problems, it is important to use the following measures when searching for data:

1. Ensuring the availability and quality of metadata (data descriptions).
2. Visualising data profiles to explain the scope and content of the data.
3. Displaying examples of data visualisations generated by the DDDM system.
4. Displaying previously performed analyses of the same data.
5. In the case of text data, make context-sensitive search and data analysis available.
6. Conducting data quality checks and visualising quality indicators.
7. Labelling the data with keywords in the data source, if possible, and use label-based search as an additional data search capability.
8. Providing data steward contacts for advice on data content, semantics, and quality.

If the DDDM system can utilise the measures mentioned above, the results and outcomes of data-driven decisions could be substantially improved over those without the DDDM system.

#### 2.1.1.5 Survey

Digital data are not always available for analysis. In such cases, the data user must collect the data through a survey.

In order to arrange a survey, the system is required to have the following use cases:

1. Creating a survey questionnaire.
2. Defining the questions.
3. Defining answer options (lists).
4. Defining and generating samples. A sample is a set of respondents.
5. Searching for contact information of respondents, e.g. performing an X-road query to the Business Register to obtain contact information of respondents from a particular field of activity.
6. Conducting the survey.
7. Analysing and visualising survey results.
8. Downloading the survey responses as a file.

Survey data can be used as a standard data table or file. **The survey tool is a non-mandatory part of the DDDM system** due to complexity foreseen while designing and implementing the functionality.

The survey tool can be used by many software products, e.g. Microsoft Forms or Google Forms, etc., but it should be noted that **common tools do not have the sample generation features and interfaces to Estonian datasets**. There is also no capability to obtain contact details of the sample to automatically send questionnaires to specific respondents by e-mail. Such features are crucial to obtain additional data in the most time-efficient manner without incurring additional costs for survey organisation.

#### 2.1.1.6 Data Integration

The main issue of efficient utilisation of data today is its accessibility. There are different datasets in Estonia, intended for internal use in different agencies and authorities. Legal aspects of publishing data for other purposes are not the topic of data integration, however, they must be considered before starting data integration. **The data integration part of the DDDM system must resolve technical issues of data integration.** The integration issues are mainly related to:

1. Different data formats and types of data elements that are incompatible with each other.
2. Different semantics of technically similar data.
3. Binding (interfacing) of data consumer and data holder systems for data exchange.

**Binding, compatibility and semantics issues must be resolved in the integration process.** The integration process can be automatic if the data has a standardised API for data exchange, a standardised format and a well-defined machine-readable semantic description that the DDDM system can understand without manual intervention in the integration process.

**Input data can come from a variety of external or internal sources**, as well as from different collection instruments, including extracts from administrative and other non-statistical data sources. Administrative

data or other non-statistical data sources can substitute for all or some of the data collected directly during the survey.

Data integration is a process in which the data from different data sources required for data analysis are gathered and arranged for data preparation and modelling. Technically, data integration can be done in several ways:

1. Structured data.
  - Retrieving data from a relational database.
  - Retrieving data from a web service.
  - Uploading data from a file (xlsx or csv) to the DDDM system.
  - Retrieving data from statistical databases (e.g. Statistics Estonia databases<sup>3</sup>, Eurostat databases<sup>4</sup>).
  - Retrieving data from the research environment of Statistics Estonia.
  - Retrieving data from spatial databases (maps and spatial objects, e.g. Estonian Land Board databases<sup>5</sup>, Inspire databases<sup>6</sup>).
  - Using data in the original data source.
2. Unstructured Data:
  - Querying data from websites.
  - Retrieving data from text documents.
  - Using data in the original data source.

The data integration may include:

1. Combining data from multiple sources to produce integrated statistics.
2. Combining geospatial data and statistical or non-statistical data.
3. Matching or recording linkage routines to link data from different sources.
4. Data fusion - integration followed by reduction or replacement.
5. Prioritising - determining when two or more sources contain data for the same variable with potentially different values.
6. Linking data with metadata.

It is important that the data integration method depends on the legal status of the data user and the data source. Some users are authorised to use aggregated data only, and the use of data with limitations is prohibited.

#### 2.1.1.6.1 Data Import

Data import is a process by which the DDDM system extracts data from a data source into its own analytical database.

In order to use data from the data source, the following work must be performed during the development project:

1. **Determine whether the data is available.** Availability can be created either through the publication of open data in the open data portal or through a corresponding web service administered by the data holder.
2. **Determine whether the data has descriptions** and whether the content of the existing dataset matches the data descriptions.
3. **The description** of data in the database **must be machine-readable**. If it is not machine-readable, it must be transformed into machine-readable. Otherwise, it is not possible to use the database for data search.
4. Check whether there are legal grounds for using data from the data source. If not, establish a legal basis for the use of the data in cooperation with the database administrator.

---

<sup>3</sup> <https://andmed.stat.ee/et/stat>

<sup>4</sup> <https://ec.europa.eu/eurostat/data/database>

<sup>5</sup> <https://geoportaal.maaamet.ee/>

<sup>6</sup> <https://geoportaal.ee/>



**To interface a database**, the database must support the following use cases:

1. Publishing database data to an open data portal.

OR

1. Managing the rights to browse database data.
2. Issuing data from the database through the web service.
3. Automatic monitoring of the performance of the data issuing web service.
4. Real-time technical support of the web service performance to ensure availability.

#### 2.1.1.6.2 Using data without importing (Data Federation)

The data can also be used without importing. If the data holder can guarantee the sustained availability, system performance, including archival and retrieval, of the data source, it is possible to utilise the data without importing all the data into the DDDM system database. Then the visualisation will do real-time queries to the dataset. When visualisations are archived, the microdata used in the visualisation must be copied to the visualisation object (file).

Data usage without importing is reasonable in the case of large-scale data (approx. more than 10 GB) if the data holder's system has sufficient computing capacity. However, the computing capacity issue can be on the DDDM side, regardless of whether the network between the data holder's database and the DDDM system uses a high-speed connection. The data holder's system must have a database server or similar system capable of processing queries sent by the data consumer's system. Using data without importing is not possible if the data location is only a data store. In this case, an application server is required. It also means that the data holder system must have an API to receive queries from the data consumer. The application server makes queries and calculations and sends back an answer to the data consumer's query.

#### 2.1.1.6.3 Data linking

Data linking is an integral part of multi-source data processing and analysis. **Linking is a data processing task where it is necessary to find keys connecting the same objects or subjects in different data sources.** For example, the personal ID code is the key to link individuals across different data sources and systems. The DDDM system can link the data if the keys are defined and known in data sources. The key may differ from the technical key of the database. The contents of a key can consist of multiple columns in the database.

The data linking operations can be performed by a third party, e.g. Statistics Estonia, or a service provider who has people, skills, software, hardware and legal rights to perform the task.

In the long term, the possibility of creating a **public classifier** that would systematise the semantics of linking data elements should be considered. It means that there should be data type definitions for linking elements. If the data element has the same specific type as another element in another dataset, elements can be used as linking elements. Every data element must have a data type defined in the metadata. If the dataset does not have such a type of definition, the DDDM system cannot link the data automatically.

#### 2.1.1.7 Data Preparation

Data preparation is the process of transforming data into a format suitable for analysis and visualisation.

##### 2.1.1.7.1 Data Profiling

Data profiling is the process whereby the data user identifies the data structure, content and relationships between datasets or data tables. It is necessary for the creation of a data model that can be used as a basis for data visualisation.

Data profiling must incorporate the following use cases:

1. Detection of data structure from a data file or database.
2. Generation of statistical metadata based on data structure and microdata content. The metadata may have different views on the data, such as value range, value frequency, data type, data format and other characteristics of the data columns, to gain a better overview of the data.
3. Identify relationships between datasets and tables using data content.
4. Storing the data profile.
5. Displaying the data profile.
6. Linking the data profile to data.

A data profiler is a tool for data analysts and modellers, who need automated metadata generation capabilities to create better data models more efficiently than they can at present.

Data profiling can also be an automated process. The DDDM system must contain features for automatic data profiling.

#### 2.1.1.7.2 Data Model

Data modelling is necessary to obtain useful insights about the data to be analysed, making data consumers' work more efficient. In the data modelling process, the DDDM system or data analyst should address the following use cases:

1. Selecting data tables or datasets to be placed in the data model.
2. Autodetecting relationships between data tables in the data model based on the data in the tables.
3. Establishing relationships between data tables or datasets manually.
4. Establishing relationships between data and dimensions.
5. Uploading auxiliary data into the system and linking them to the data.
6. Setting filters on the data. Filtering out incorrect or unnecessary data from the data model.
7. Setting data display formats, e.g. date format, currency format or floating-point numbers format.
8. Describing data model metadata (table, column, and relationship descriptions).
9. Displaying data model as an ERD (entity-relationship diagram).
10. Generating ERD layout automatically.
11. Setting the ERD layout manually. It is necessary if the number of data tables exceeds 4-5 tables.
12. Exporting data model metadata as a document.
13. Publishing the data model in the data model repository of the DDDM system.

The result of the data modelling is a well-designed and described representation of relational data for data consumers (or data analysts). The modelled layout is a starting point of data visualisation. **The model must always be executed if there is more than one data table to be analysed together.**

**In order to automate the compilation of the data model,** the following activities must be automated:

1. Detecting primary keys in the dataset.
2. Detecting possible related columns.
3. Detecting relationships between data tables.
4. Relationship multiplicity detection.
5. Providing the user with model options.

The above listed activities are mostly introduced and already existing in commonly known analysis tools such as Microsoft Power BI, Tableau or similar, hence the automation of data models is doable provided that data quality is ensured.

#### 2.1.1.7.3 Data Classification

In most cases, the data must be classified before the system can proceed to analyse it. It means that the system should calculate a class code for each data record in the dataset based on other code columns or columns containing unstructured data, e.g. text data or images. Coding procedures may assign numerical codes to text values according to a predefined statistical classification to facilitate data collection and processing.

With a classification dictionary where the system can detect the compliance between text and code, the DDDM system can classify a data record based on the text data. Classifying images is more complex. In most cases, it is not possible to classify images without using machine learning before classifying the dataset.

#### 2.1.1.7.4 Data Coding

The data from different data sources can be classified using different classifications that have the same or similar classification values. In such cases, the DDDM system must perform coding based on the coding dictionary. Some data columns may have hard-coded value categories in the administrative data source.

Coding and classification are important procedures to prepare the dataset for analysis.

#### 2.1.1.8 Data Quality Control

In addition, data quality management must be inspected automatically, the use cases are as follows:

1. Data quality control.

2. Defining data checks.
3. Running data checks.
4. Preparing data quality report.
5. Improving data quality according to the identified problems.
6. Automatic correction.
7. Manual correction in the source database by the data holder.
8. Issuing a data quality assessment to the data consumer through the online service.
9. Service for receiving descriptions of data errors. Through this, the data holder can be informed about the problems detected in the database.

The system must have a data control solution that allows controlling the data regardless of its origin and characteristics. Errors can occur in the data both when entering it in the data source, when transferring the data to DDDM, and when transforming a data set inside DDDM. This means that data checks must be available at every stage of the data lifecycle.

Data quality control requires the creation of data checks. Methodology of the data quality check is introduced in section 7.2 on page 48. This can be done by a user who is a data analyst. Such a user can be an SA employee or a DDDM user with corresponding competence.

Data checks can be run automatically by the system or manually depending on the need. The result of the run shall be a data quality report that the user uses when interpreting the results of the data analysis.

#### 2.1.1.9 Metadata Governance

The system should use the Estonian metadata infrastructure RIHA(KE), regardless of whether it applies to the database. It is a metadata management system for every database owned by the Estonian public sector agency or authority. Not all databases and data elements are described in the RIHA. RIHA and/or RIHAKE should be linked to the DDDM system.

The user of the DDDM system must have the possibility and capability to use metadata to avoid errors in the interpretation of data analyses. Selecting data for analysis is very challenging if there is no description of the data in the data content, e.g. what are data objects in the database or how many variables the data contain, how the data is classified, what are the classifier elements, what are the meanings of the elements, etc.

Metadata usage options:

1. Querying metadata.
2. Reading the metadata.
3. Searching for keywords from the metadata to find relevant data columns (or variables) in the database.

**A repository of DDDM analysis applications must have its own metadata governance system.** It should be noted that the application metadata does not coincide with metadata on data. The memorandum user must be provided with the following use cases:

1. Searching for an analytical application based on the application description.
2. Commenting and describing analytical applications for further efficient use of the application.

It should also be considered that metadata created in the DDDM system should be transferred back to the RIHA. RIHA is the primary data source of metadata, especially metadata on public registers and datasets belonging to the state information system. If RIHA could be a metadata management solution and online metadata repository for the DDDM, it would be the best solution from the state's perspective.

RIHAKE allows user to:

1. Read metadata about the data structure and descriptions from the existing database.
2. Manually update the metadata read from the database.
3. Publish metadata in RIHA.
4. In the future, it will be possible to publish metadata in the public data gateway (in estonian *andmete teabevärav*) as well.

RIHAKE is an existing metadata management system created by Estonian Information System Authority for every public sector agency in Estonia. RIHAKE can be used for metadata management on both

microdata and aggregated data. There can be information in the RIHAKE database on how an aggregate is calculated.

The use of RIHAKE is not mandatory. There can be more metadata management possibilities in the future.

### 2.1.1.10 Data Analysis

The analysis process should be automated to the greatest extent possible. The data source can have multiple dimensions and filtering options. In addition, many analysis scripts can be useful. Choosing the right ones may be a sophisticated process. If the user can select useful filters, dimensions and scripts, the result of the analysis can be more efficient without user intervention. It is important that first users make the right decisions to train the AI system in the DDDM. Such first users can be human trainers like machine learning experts or pilot users. It is aimed that algorithms learn from data; hence someone must feed a machine learning algorithm with data.

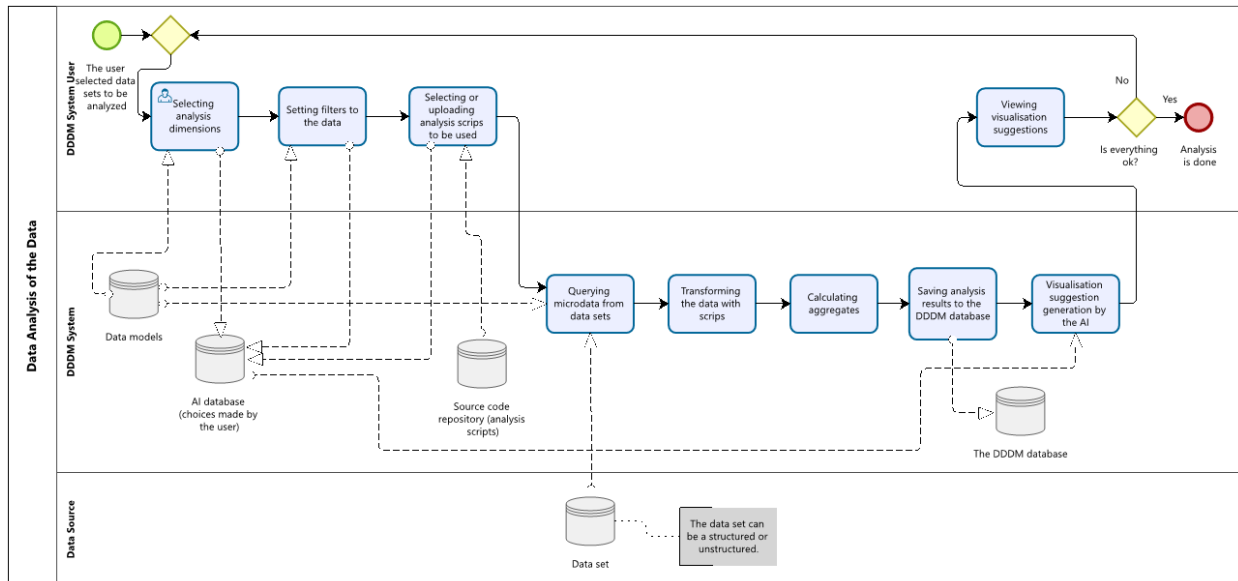


Figure 7. The Detailed Data Analysis Workflow in the DDDM system

The following use cases must be implemented in the data analysis subsystem:

1. Searching, sorting and filtering the data in the analysis database (prepared data model).
2. Selecting prepared data for analysis.
3. Selecting dimensions for grouping the data in visualisations.
4. Selecting units to be used in the visualisation.
5. Performing transformations on data columns, if necessary, e.g. setting up generalised dimensions that have values such as “greater than 1,000” and “less than 1,000”.
6. Calculating aggregates.
7. Displaying microdata.
8. Displaying aggregated data.
9. Selecting visualisation types to be used in visualisations.
10. Visualising analysed data:
  - Visualising geospatial data.
  - Visualising numerical data.
  - Visualising text data analysis.
11. WHAT-IF functionality with variable input parameters.
12. WHAT-IF with forecasting features.
13. Describing and commenting on visualisations – compilation of data records.
14. Publishing visualisations and explanations to visualisations.
15. Gathering feedback on the analysed data and visualisations.

#### 2.1.1.10.1 Geospatial data visualisation

Geospatial data sources are developed in Estonia. Many examples of visualisation and analysis can be observed on the map.



Processing and analysis of geospatial data requires special GIS software. The use of geospatial data can be divided into two general types:

- Spatial data analysis.
- Utilising the base map (see Figure 8) and additional map layers to enrich the visualisation (see Figure 9 and Figure 10).

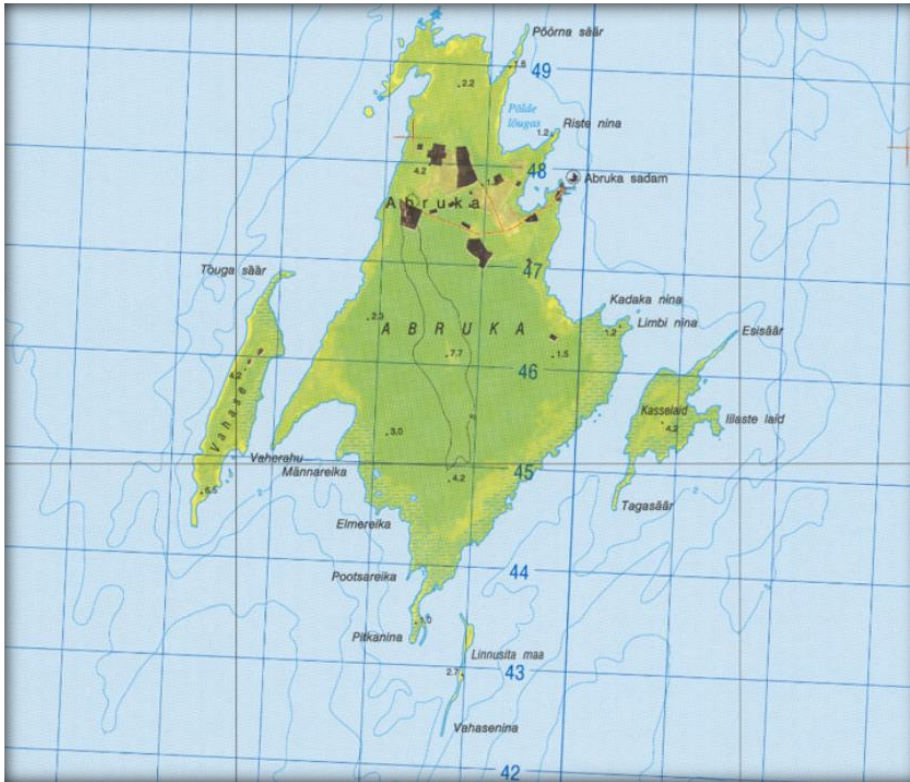


Figure 8. An example of a base map

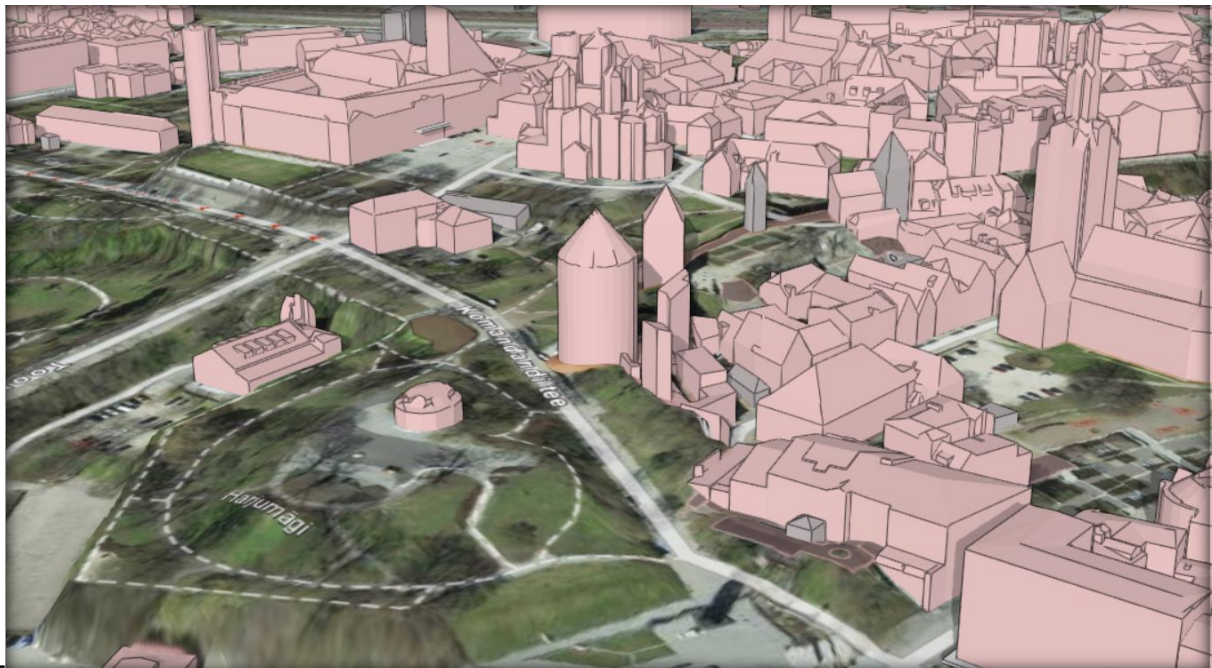


Figure 9. 3D map



Figure 10. Transparent map layers on the base map, ARIB agricultural parcel map layer

All these map types shown in the figures above can be used to visualise the data on the map. For example, agricultural production in figures can be visualised using the ARIB field map layer. The visualisation is dynamic, and the user of the visualisation can have an overview and a detailed representation at the field level in figures and as a map. Adding an additional layer, e.g. fertilizer usage or soil composition, creates a very comprehensive representation.

Functionalities of the DDDM system for analysis and visualisation are the following:

1. Analysing spatial data in spatial data formats on a GIS server, e.g. PostGIS<sup>7</sup>, ArcGIS<sup>8</sup> or similar.
2. Placing map layer in WMS<sup>9</sup> format for visualisation.
3. Placing map layer in WFS<sup>10</sup> format for visualisation.
4. Linking the map layer and a numerical dataset, aggregated or microdata.

Spatial data is any type of data that directly or indirectly refers to a specific geographic area or location. It is important to use spatial data in a standardised format, e.g. coordinates in WGS84 or L-EST97, or to record the name of the location correctly to use automatic georeferencing of the object to the map data, or other spatial dimensions such as roads or other shapes. The different map layers can be used in the same visualisation, regardless of whether the layers are in the same coordinate system or the DDDM system has the capability to convert different coordinate systems.

The satellite data can also be used for spatial analysis, e.g. to assess the environmental condition of water bodies. The EstHUB<sup>11</sup> is a source of satellite data on the Baltic Sea region and Estonian territory.

---

<sup>7</sup> <https://postgis.net/>

<sup>8</sup> <https://en.wikipedia.org/wiki/ArcGIS>

<sup>9</sup> <https://www.ogc.org/standards/wms>

<sup>10</sup> <https://www.ogc.org/standards/wfs>

<sup>11</sup> <https://geoportaal.maaamet.ee/est/Ruumiandmed/Riiklik-satelliidiandmete-keskus-ESTHub-p443.html>



It is possible to automate many geospatial analyses today. The technology for different data analysis modules on satellite data is available as an open-source products<sup>12</sup> and commercial products<sup>13</sup>. Open-source geospatial data analysis source code can be implemented in DDDM system if needed. Plenty of useful geospatial data analysis algorithms are already available in QGIS such as NDVI<sup>14,15</sup> or NDWI<sup>16</sup> indexes generation. These indexes are used widely in agriculture. Indexes can be used in both macro and micro level decision making process.

The Estonian Land Board has a wide range<sup>17</sup> of WMS/WFS data services for the user.

The DDDM system must automatically provide the user with different data services (map layers) that the Estonian Land Board has prepared for the user, depending on the scale of the numerical dataset linked to the geoinformation. The dataset can be georeferenced if it contains a geographical dimension in the form of geographical coordinates or addresses.

#### 2.1.1.10.2 Numerical data visualisation

Data analysis in the DDDM system must be automated to the greatest extent possible, especially when analysing numerical data that have traditional workflows.

For automated analysis, the system must have several autodetection features:

1. Detecting measures in a dataset.
2. Detecting dimensions in the dataset:
  - Detecting time dimensions.
  - Detecting geographic dimensions.
  - Detecting other dimensions.
3. Detecting unit columns using unit abbreviations classifiers.
4. Detecting which dimensions can be used for data analysis and which are most represented in the dataset records.
5. How does the distribution of dimension values in the dataset.
6. Which dimension values have the highest weight.
7. Interestingness index: what are the more interesting distributions in the dataset, the interestingness index must be calculated by the DDDM system.
8. Rating of the data quality in the column.

More sophisticated and accurate autodetection features can be:

1. Calculating dependency functions based on the distribution of dimension values.
2. Running third-party analysis (e.g. R scripts).

Recommended set of visualisation types for the numerical data analysis in the DDDM system may be, but not limited to:

1. Data table.
2. Matrix.
3. Numerical aggregate chart.
4. Value frequencies.
5. Line chart with time series of aggregates.
6. Bar chart.
7. Stacked chart.
8. Multidimensional bar chart.
9. Pie chart.
10. KPI chart.
11. Slicer for data filtering.

---

<sup>12</sup> <https://www.qgis.org/en/site/>

<sup>13</sup> <https://www.arcgis.com/index.html>

<sup>14</sup> <https://gisgeography.com/ndvi-normalized-difference-vegetation-index/>

<sup>15</sup> [https://en.wikipedia.org/wiki/Normalized\\_difference\\_vegetation\\_index](https://en.wikipedia.org/wiki/Normalized_difference_vegetation_index)

<sup>16</sup> [https://en.wikipedia.org/wiki/Normalized\\_difference\\_water\\_index](https://en.wikipedia.org/wiki/Normalized_difference_water_index)

<sup>17</sup> <https://geoportaal.maaamet.ee/est/Teenused/WMSWFS-teenused-p65.html>

12. Map visualisation supporting Estonian base chart.
13. Map visualisation with WFS support for using different map layers, e.g. administrative units.
14. Heatmap visualisation for graphical display of numerical aggregates on the map.

#### 2.1.1.10.3 Text data analysis visualisation

There are several types of text analytics of which some have been highlighted:

1. Word statistics and search.
2. Context-sensitive analysis and search.
3. Guided machine learning and text analysis.
4. Semantic analysis of the text.
5. Extracting numerical data from text.
6. Dialogue bot:
  - Based on standard questions and answers.
  - Based on guided machine learning questions and answer system.

The representation of the text analysis result can be a frequency table of words, sentences in the text, a list of sentences for which information was detected, word cloud, or a translation of the text.

#### 2.1.1.10.4 WHAT-IF functionality with changeable input

The DDDM system must have WHAT-IF features to help the government test various scenarios that the decision may entail.

There can be several ways to perform a WHAT-IF analysis. The most straightforward way to generate different results for the same analysis is to change the input parameter for analysis. The DDDM system should have dynamic visualisations with sliders that change the input parameters of the analysis. It is the way to provide the best possible user experience.

For example, if the output of the analysis is to show the impact of the value-added tax rate on the sector indicators, the tax rate should be presented in the visualisation as a slider field. Therefore, the user could easily adjust the value of the field and the content of the visualisation would change. If different input parameters are combined in this way, the user can get a powerful tool with a very simple user interface. The assumption is that the preliminary work, such as building the data model and creating the visualisation, has been done in advance before the end user can experiment with the tool.

#### 2.1.1.10.5 WHAT-IF with prediction features

WHAT-IF functionality can be implemented using prediction functions in the same visualisations. Thus, the user can get a time series to influence decision-making process.

#### 2.1.1.11 Applications Repository

If a DDDM system user creates visualisations, datasets or data models that may be useful for other users, the system should have a repository for storing the scripts they have created. Such repositories can be observed on different platforms, e.g. R script repositories<sup>18</sup>, Power BI<sup>19</sup> or Tableau<sup>20</sup> example galleries. The DDDM AI features can use script usage statistics to suggest scripts in particular situations.

If the DDDM system uses a data model and visualisations repository, it is necessary for the visualisation, or data model, or both to form an independent object containing the data and the model.

It means that the data and the data model are in the same file, and this file contains all the necessary components to render a visualisation or data model.

The repository should have the following use cases:

1. Repository user and storage management.
2. Uploading data models to the repository.
3. Uploading visualisations to the repository.

---

<sup>18</sup> [https://cran.r-project.org/web/packages/available\\_packages\\_by\\_date.html](https://cran.r-project.org/web/packages/available_packages_by_date.html)

<sup>19</sup> <https://community.powerbi.com/t5/Data-Stories-Gallery/bd-p/DataStoriesGallery>

<sup>20</sup> <https://www.tableau.com/data-insights/dashboard-showcase>



4. Checking the quality of the source code of visualisations and data models.
5. Searching repository objects in the repository.
6. Downloading repository objects by the user.

An important topic for data analysis is a code list and classification data repository. An example of such a repository can be found at Statistics Estonia: <https://klassifikaatorid.stat.ee/>.

#### **2.1.1.12 Reporting**

The user may use the same report (visualisation) multiple times to analyse different time periods in the past and future. Standard reports are part of the DDDM system repository. Reports are data visualisations. It is important to have the possibility to automatically update the data from the data source in the report in the DDDM system.

#### **2.1.1.13 Giving feedback on artefacts**

It is crucial to receive feedback from different specialists and experts on data analysis and visualisations. Each visualisation and memorandum in the DDDM system must have a feedback section.

#### **2.1.1.14 Publishing**

DDDM must have an environment for publishing the results of data analysis (visualisations), one for an authorised user and another for other public users. It is important to publish visualisation to the public user to get feedback before the government considers the results of the analysis. An essential part of the publishing environment is the feedback subsystem.

The publishing environment must contain the following functionalities:

1. Uploading visualisations containing micro- and aggregated data.
2. Displaying dynamic visualisations.
3. Running visualisation as a standalone application for integration into any web application, e.g. a government website or internal government systems.
4. Downloading visualisations for further development.
5. Deleting visualisations from the publishing environment.
6. Authorised download of microdata from the visualisation.

Integrating visualisation into the memorandum is also a publishing feature. The memorandum and visualisation must form an independent file that can be a static application or a source file for further development and updating.

## 2.2 Support Functionalities of the Technical Solution

Estonia has a well-developed IT infrastructure with interoperability between different systems. It is crucial to utilise common systems, e.g. authentication system of Estonian people in the DDDM. Therefore, this chapter describes a list of supporting functionalities. Without these functionalities, the system cannot operate, or the data security measures will be too weak, and the system will be an illegal platform for the user.

### 2.2.1.1 User Authentication

User authentication must be performed using TARA. TARA is a standard solution for user authentication in Estonia.<sup>21</sup>

### 2.2.1.2 User Rights and Roles

User management should be implemented in the central TARA authentication system. TARA contains identity information of every citizen of Estonia. The system does not require an independent user management system. TARA automatically manages user information and provides user authentication functionality as required.

### 2.2.1.3 Interfaces

The Estonian interfacing platform is X-road. It is a secure network connecting mainly public sector information systems. Many databases are interfaced with X-road through X-road web services. The most frequently used data is published on the X-road, but it is a small part of the data needed for data analysis for memorandums. It means that there must be more capacity to transport the data. The DDDM system must have interfacing functionality to any required data source either directly or via a standardised format in at least the following ways:

1. SOAP web services<sup>22</sup>.
2. REST web services<sup>23</sup>.
3. ODBC<sup>24</sup>.
4. JDBC<sup>25</sup>.
5. CSV files<sup>26</sup>.
6. XLSX files.
7. XML files<sup>27</sup>.
8. JSON files<sup>28</sup>.
9. PostgreSQL database.
10. Oracle database.
11. MS SQL Server database.
12. MS Access files.
13. OData web services<sup>29</sup>.
14. Denodo<sup>30</sup>.
15. Docx files.
16. Pdf files.

---

<sup>21</sup> <https://e-gov.github.io/TARA-Doku/>

<sup>22</sup> <https://en.wikipedia.org/wiki/SOAP>

<sup>23</sup> [https://en.wikipedia.org/wiki/Representational\\_state\\_transfer](https://en.wikipedia.org/wiki/Representational_state_transfer)

<sup>24</sup> [https://en.wikipedia.org/wiki/Open\\_Database\\_Connectivity](https://en.wikipedia.org/wiki/Open_Database_Connectivity)

<sup>25</sup> [https://en.wikipedia.org/wiki/Java\\_Database\\_Connectivity](https://en.wikipedia.org/wiki/Java_Database_Connectivity)

<sup>26</sup> [https://en.wikipedia.org/wiki/Comma-separated\\_values](https://en.wikipedia.org/wiki/Comma-separated_values)

<sup>27</sup> [https://www.w3schools.com/xml/xml\\_what\\_is.asp](https://www.w3schools.com/xml/xml_what_is.asp)

<sup>28</sup> <https://www.json.org/json-en.html>

<sup>29</sup> <https://www.odata.org/>

<sup>30</sup> <https://www.denodo.com/en/denodo-platform/overview>

#### **2.2.1.4 Archive and Version Control**

The DDDM system must have an archiving system. An archiving object is a file containing data models, visualisations, and documents (memorandums) with data visualisations, as well as the data necessary to restore the archived object file. The system administrator must be able to move files to the archive and restore them from the archive. The archive can be a file data storage in the cloud or on-premises.

Version control of documents and visualisations can be organised using SharePoint or similar products supported by MS Office 365 software commonly used in the public sector.

#### **2.2.1.5 Monitoring**

All parts of the DDDM high-availability system must be monitored using ZABBIX<sup>31</sup> or similar software. There must be an organisational unit responsible for configuring the monitoring system, and the unit's specialist must resolve any incidents involving loss of availability.

All server backend services that serve the user interface or external systems must be monitored and should be highly available.

#### **2.2.1.6 Logging**

The system must have an event logging system. Each user-initiated event involving sensitive data must be logged into log files that are separate and independent from the system, and in a common format, e.g. text format. The system must log the following elements of event data:

1. Which user acted.
2. What the user did.
3. When the user acted.
4. Which functionality was initiated by the user.
5. What was the computer the request came from.
6. What files or datasets have been utilised by the user.

A separate tool must be provided to analyse the logs. This tool can be a standard data analysis tool such as Power BI, Tableau, etc. Log analysis can be performed by the system administrator or a special officer who has the rights to analyse logs.

#### **2.2.1.7 System Administration**

The DDDM system administration must cover the following tasks:

1. Deploying system components.
2. Authorising users.
3. Configuring system monitoring.
4. Resolving system-related incidents.
5. System data backup.
6. Archiving visualisations, data models and documents.
7. Configuring interfaces with data sources.

In order to help administrators and users manage the data integration from different sources, a data steward role should be established. Data should also be organised and managed after integration to ensure data availability to users.

---

<sup>31</sup> <https://www.zabbix.com/>

# 3. Governance and Business Processes

## 3.1 Governance Structure

### 3.1.1 Organisation

The organisation of the DDDM system constitutes of Government Office and Statistics Estonia, which are the leading authorities, and Information Technology Centre for the Ministry of Finance (RMIT), Information System Authority and Data Holders, which are the supporting authorities, as shown in the figure below.

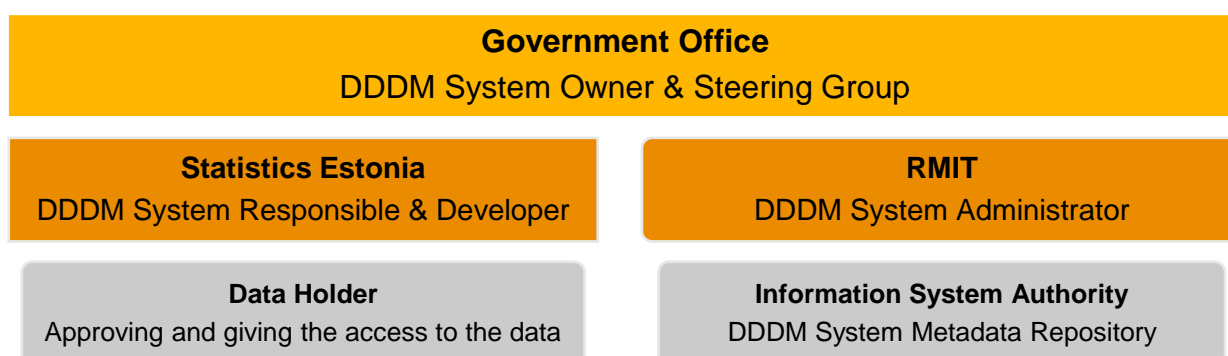


Figure 11. DDDM System Organisation

#### Leading Authorities

The key members of the system governance and management are the Government Office and Statistics Estonia.

**The Government Office** is the owner and governing body of the DDDM system, providing funding for the system; it is closely connected to the government, which will be the main beneficiary of the analytical outputs and decision-making proposals provided by the DDDM system and its users.

**Statistics Estonia** is the responsible body for the development and implementation of the DDDM system. Statistics Estonia performs the functions of administration, development and technical as well as procedural support of the DDDM system.

#### Supporting Authorities

**Information Technology Centre for the Ministry of Finance (RMIT)** is the centre that administers and develops the information systems of the Ministry of Finance and the Estonian Tax and Customs Board, including Statistics Estonia.

**Information System Authority** develops and manages the administrative environment and the catalogue for the state information system (RIHA), providing information on the following:

- which are the information systems that compose the state information system;
- which data are collected and processed and in which information systems (metadata);
- who are the owners, maintainers and contact persons of information systems;
- on what legal basis are the information systems operated and the data processed;
- the reusable components that ensure the interoperability of information systems (XML assets, classifications).

**Data Holders**<sup>32</sup> are organisations or individuals who, according to applicable laws or regulations, are authorised to decide on granting access to or sharing data under their control, regardless of whether such data are managed by the organisation or individual or by an agent on their behalf.

### 3.1.2 Roles and responsibilities of DDDM System

The DDDM does not automate the data gathering and analysis by itself. Automation requires a significant number of people across multiple roles. Automation can only be done manually. Machine learning could support this process; however, at the first stage, quality control and testing are mandatory for every new automation.

Table 2. Roles in DDDM user organisation

Role	Description
User	The DDDM user is a public servant who uses the DDDM for data analysis.
Administrator	Administrator sets up all DDDM system components, configures connections to data sources, monitors and resolves issues related to the availability of the DDDM system.
Data Steward	This role is the supervisory or data governance role within an organisation and holds responsibility for ensuring the quality and usability of the DDDM data assets, including metadata. A data steward supports users in understanding data semantics.
Data Source Representative	There must be a data source specialist and representative, a contact person from the data source side who can explain technical details of a particular data source to the DDDM administrator and user.
Data Analyst	Data analyst is a top-level DDDM user who processes data, performs transformations, models data, performs complex data analysis and provides support to other users. The data analyst should also be capable of establishing relationships between data.
Man of Law	A lawyer who drafts legislation for DDDM needs.
Data Warehouse Developer	Data warehouse developer assists the data analyst in conducting complex analyses and develops tools for data processing and analysis. Can be a contracting company.

<sup>32</sup> OECD, Recommendation of the Council on Enhancing Access to and Sharing of Data. <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0463>

## 3.2 Business Processes

### 3.2.1 General analysis process

The next step after data search is the analysis of open data (section 3.2.2) or personal and sensitive data (section 3.2.3). The business process related to open data is the fastest way to get preliminary results of data analysis, hence it is advised to initiate the work with open and available data (either micro- or aggregated data). In many cases, it is enough to facilitate effective decision-making. If the result is not sufficient, the second iteration is necessary. The subsequent iteration should be performed on microdata and may be subject to legal, technical and operational limitations.

As indicated in Figure 13 below in section 3.2.3, additional analysis can be performed in the DDDM or in Statistics Estonia, depending on the legal status of the DDDM regarding the processing of a certain dataset. If Statistics Estonia performs aggregation and/or linking of data, DDDM can obtain the required result of the analysis.

### 3.2.2 Open data interfacing related business process

The interfacing of the open datasets is relatively simpler than interfacing datasets containing personal data or other data requiring protection. Open data do not require a special right to use the data in analysis. A technical problem with the open data lies in the fact that there is no **open data format standard** in Estonia. It means that in many cases every dataset has its own format and data querying solution must be different. Additionally, open data metadata has no standardised format and may not be machine-readable. Due the lack of standardisation, smooth automation cannot be applied. There is no protocol to automatically determine the format of a particular dataset. It is necessary to manually analyse how to make the data available in the DDDM system. In order to solve the problem, all datasets must produce machine-readable metadata, and the Government Office as a DDDM owner must set such requirements and obligations for entrepreneurs and government authorities that publish data for the DDDM. Automatic metadata generation in the DDDM is possible to some extent in cases with a revealed third standard.<sup>33</sup>

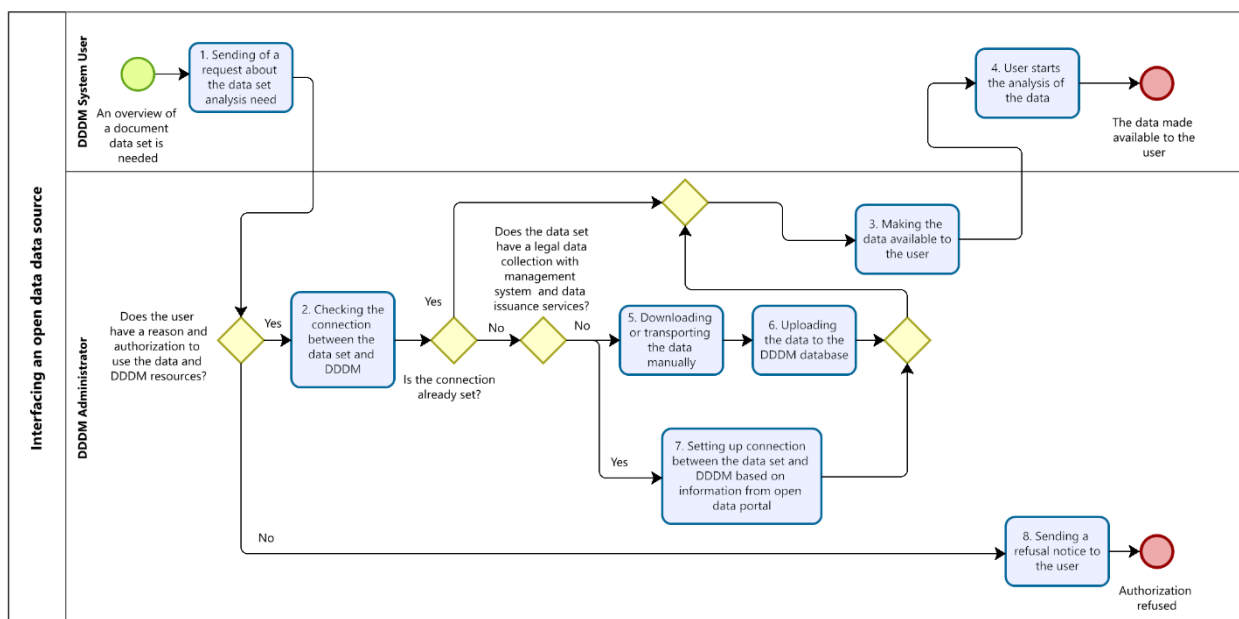


Figure 12. Data source interfacing process on open data.

Connecting DDDM to a data source on an open data portal may be organisationally simpler than connecting to a data source containing personalised data. It depends on how and in what format the data is shared.

<sup>33</sup> [Third normal form - Wikipedia](#)

### 3.2.3 Data source interfacing related business process

Every data source to be interfaced with the DDDM system should have a dataset to be used and the infrastructure for storing the data. The DDDM must be connected to the data source infrastructure using web services or other transfer protocol. As the goal is a fully automated system, alternative data transfer protocols, such as the hard disk drive, can only be utilised if the data is extremely valuable and does not lose significant value over time. The modern solution is that the data source establishes a very fast internet connection and then uploads the data to a cloud server or a local data sharing server that can also be used by the data user. The process of updating the data must also be resolved on the cloud server.

The process diagram below visualises a business process of interfacing a data source with the DDDM system.

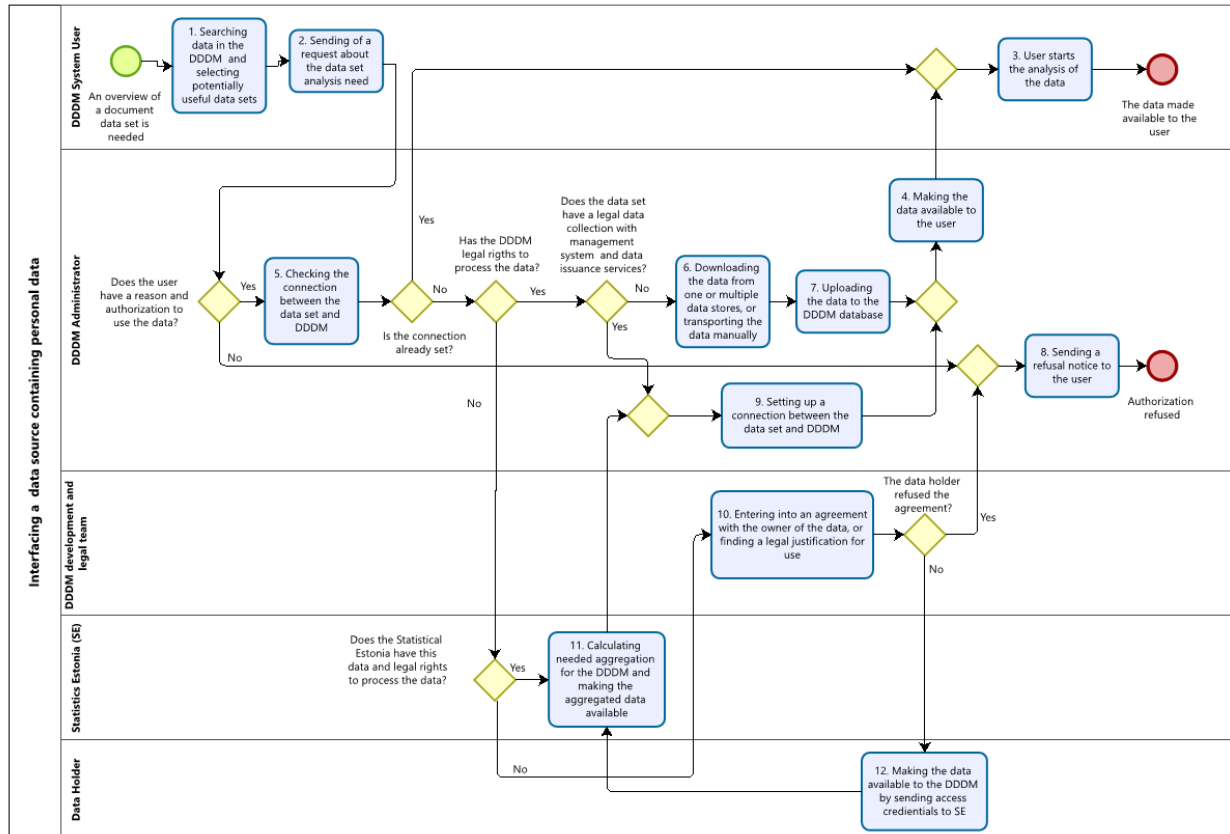


Figure 13. Data source interfacing process

The system **displays both connected and not yet connected data**. The data list contains data descriptions based on metadata available to DDDM. The metadata repositories must be connected to DDDM before the user starts searching and selecting data. The process of preparing the DDDM system is explained on Figure 5 on page nr 13.

Setting up a data source interface may be a complicated process depending on how standardised the data source is. Much depends on which regulation underpins the use of the database and whether the dataset has web services. If there is no web service for querying the data, the data gathering process may be manual. It means the user must manually establish connections and manually copy data.

### 3.3 Legal Analysis

The basis, functions and goals described in the project require the DDDM to be regulated at the legislative level, depending on the technical organisation, either as a separate database or by prescribing equivalent data processing rules. In all three scenarios, it is necessary to establish the legal basis for data processing and define the purposes of data processing.

#### Need for a legal framework

During the activities of the state, including the activities related to various administrative procedures, contractual relationships in the private sector, as well as open data, many data are generated, the processing of which is connected to the provisions of Section 26 of the Constitution of the Republic of Estonia (the inviolability of privacy) and the European Union law (GDPR) and other legislation regulating data management in the EU. The DDDM is also related to the Estonian national legislation (Public Information Act, Personal Data Protection Act, etc.) and the work of the existing governmental institutions (e.g. the Government of the Republic, ministries, Statistics Estonia, etc.). **The interim results also emphasise the need to establish a legal framework for the functioning of the system.** The technical data operations specified in the interim results require the corresponding legal basis (scope of data processing, purpose and other general principles of personal data protection). In addition, it is assumed that the **exchange of data is accomplished via the X-tee, the datasets described in the RIHA are used and other RIHA-related requirements are fulfilled.**

#### Ensure the fundamental rights of individuals

The most sensitive data are personal data (especially sensitive personal data) and data related to trade secrets, the processing of which will require a legal solution in the context of the planned DDDM. Pursuant to the Constitution of the Republic of Estonia and the European Union law, data processing must have a legal basis. It is proposed to implement the legal basis provided in Section 6, Clause D of the General Data Protection Regulation, which addresses the fulfilment of public law function and ensures the lawfulness of data processing.

In addition to ensuring the fundamental rights of individuals, it is necessary to provide for measures to protect trade secrets, for instance, when solving problems related to the economy, finance or a specific industry (e.g. food industry). It means that the DDDM regulations should provide for the right to data processing for all industries.

#### Define the new public services at the legislative level

In public law, the principle is that **what is not permitted is prohibited**. In the context of the DDDM, this means that **new public services provided through the DDDM must be defined at the legislative level**, especially those public services that affect the fundamental rights of individuals. In order to fulfil the public law function, the **state is obliged to apply the principle of cross-use of data in any case** (e.g. X-tee, RIHA and other public services, e.g. TARA). The private sector has not adopted such a broad approach and generally requires the consent of the recipient, which may be withdrawn at any time. A new public law function can be, for instance, to improve policy-making and decision-making or to perform impact assessment at a more general level. Further legal analysis may reveal that the legal basis for data processing already exists within existing tasks (e.g. statistical work or data sharing service). If it is a matter of performing a broader task, the DDDM function must be described in a separate law.

#### Provide clear arguments on how extensive data processing is justified

At the same time, special attention must be paid to extending the DDDM functionality and the existing rights to DDDM (e.g. statistical work). However, **it must be recognised that such extensive data processing must be convincingly justified** during the DDDM development phase. If until now the processing of data has been considered permissible for the implementation of a specific administrative procedure, e.g. for the payment of subsidies, **the preparation of memorandums of the Government of the Republic for policy formation requires clearer arguments as to how extensive data processing is justified.**

During the development phase of the DDDM, **it must be clearly defined what data are to be processed** for the development of the impact assessment of the memorandums of the Government of the Republic or for conducting surveys. Such rationale reveals the public law function performed by the system.



**Data processing for any purpose is not permitted.** When establishing the DDDM system, it is not necessary to introduce large-scale and fundamental changes to the legal system of Estonia. However, **there is a need to create a legal framework for the DDDM or to amend the existing data processing framework** (e.g. the National Statistics Act). The legal analysis has revealed that **Statistics Estonia has the best substantive and legal prerequisites** for achieving the goals of the current project. However, in this case, it is necessary to amend the **National Statistics Act**, since Statistics Estonia does not yet have the task of conducting impact assessments or any similar task, and **it should be analysed whether the functionality of the DDDM would cover the tasks of statistical work or data sharing services**. This issue must be addressed in all three scenarios.

### Create a separate regulation

**The first preference would be to create a separate regulation (a dataset) for the DDDM** or merge it with some other regulation, e.g., a) to provide for a new public law function (impact assessment); b) to formulate the goals of data processing; c) to define, if possible, the scope and volume of the data processed and to provide for other measures to ensure the fundamental rights, e.g. the data retention period; d) to make other technical corrections in the same legal act, as well as, if necessary, to amend other legal acts, the data of which are currently unavailable or the access to which is required by a special law (e.g. the Taxation Act); f) to define the work organisation of Statistics Estonia and the ministries, as well as data processing issues. In addition, **it is necessary to ensure appropriate measures for the protection of commercial secrets and the prevention of abuse of the system**.

### Resolve the legality of data processing

Other possible regulatory mediations such as a dataset or a metadata information gateway were also analysed during the project, but given the goals of the project, these solutions do not meet the substantive needs of the DDDM. Since the KOOS, the successor of the current information system for draft legislation (EIS), is under development, it has never processed personal data in such a manner. In case such functionality is added to the KOOS, **the issue of the lawfulness of data processing**, including the above-mentioned needs for changes, will still need to be addressed. If the DDDM and its database are merged with any existing dataset, the legal regulation of the dataset will need to be changed.

According to the three scenarios, there are several ways to build a system, but **good organisational arrangements do not negate legal issues**. It is also probable that there will be a need to amend the statutes of data providers as the restrictions on data transmission are described in the statutes of datasets and partially in the law. Other EU-funded IT projects are also regulated by law, e.g. the processing of spatial data deriving from the Inspire directive, the exchange of social security data between Member States, or the information exchange of medical prescriptions between Estonia and Finland.

### Gateway to other datasets

**If the DDDM does not process the data independently** (no substantive services are transferred to it), then **it should remain only as a metadata exchange layer**. This means that it is a **gateway to other datasets** (self-service dataset environment), enabling all relevant information to be processed in one place. In this case, the DDDM itself must not have a dataset.

In this case, the DDDM must also not be able to profile individuals or establish the connection between data and a specific individual. In this case, all databases and algorithms must remain in the current datasets (also from a legal perspective, they would remain independent datasets). However, this solution does not comply with the requirements described in the interim results of the project, because it is known that a database for the DDDM will be created under any circumstances (even temporarily). It also appears that the engine of the technical solution of the system (the application) is central, and a separate dataset connected to each data source will not be developed to fulfil the goals of the project (too expensive and clumsy). Although data may be stored in a distributed manner, their processing is still scheduled in the DDDM application.

### More complexity would require more risk management

Such a solution also presumes that the DDDM does not process personal data in any way and does not interface with the X-tee to process it. However, this principal contradicts the requirements described in the interim results of the project. The **most complicated aspect of such a solution is the permanent and fast access to the national data** recognised as internal information, which is an important input for

preparing the memorandums of the Government of the Estonian Republic. **Well-informed decisions cannot be made only based on open or non-personalised data.**

In the case of the described legal solution, the processing of personal or sensitive data will generally take place based on consent, not for the fulfilment of public law function. The state generally does not process personal or sensitive data based on consent as its business logic is described at the legislative level, which ensures its right to process data. **However, the more complex the underlying business logic of the DDDM becomes, the more protection it needs at the legislative level.**

#### **Provide reason and purpose for processing data**

When creating such a large system, the public interest must be considered from the outset in the governance of data processing. This means that **people must have an idea of who is processing the data and for what purpose.** The state has collected data to fulfil its public law function, and it has been given a fiduciary guarantee that cannot be abused. This means that people have provided their data to fulfil the obligations arising from the legislation, and specific rules must be established for any other use of the data. This, however, does not prevent the creation of new e-services and innovations. The principle must be balanced with other rights and obligations that the state has with respect to the data that is recognised as inside information (e.g. trade secrets).

# 4. Target Data Model

## 4.1 Data Structure Types

The data has different types as mentioned in Deliverable 1.3. Regardless of the data type, statistical methods of the data analysis must be used to provide understandable and comparable results for users and policy makers. The result of the data analysis is not a table of microdata records which is an unfathomable amount of data. The result should always be aggregated data or visualisation of the aggregated data.

When using a text dataset or other slightly structured datasets, the result of the analysis is also an aggregated dataset in the form of a table or visualisation. **Aggregated data** on the situation in a particular area are necessary to prepare **a memorandum or other policy-making document**. For the most part, microdata is unnecessary.

All this means that the DDDM system needs **numerical data as data facts** and **alphanumerical data for dimensions** according to which the data must be analysed. Numerical data can be used in aggregated form as required in the DDDM. Dimension data may contain both numbers and letters. It is an important rule for the dimensional data model, that is an integral part of almost every data analysis.

### 4.1.1 An example of aggregated data as facts and dimensions

**For example**, if the user of analysis results needs to answer the question of what is the age structure of the population in terms of employed and unemployed, the DDDM system needs all population data (approximately 1.3 million records in Estonia) with two dimensions – age range and employment status. In this case, the DDDM system requires a record for each person with at least two columns:

1. Employment status as dimension 1.
2. Age range as dimension 2.
3. Value as a fact.

A three-column data table is the input dataset for the DDDM system.

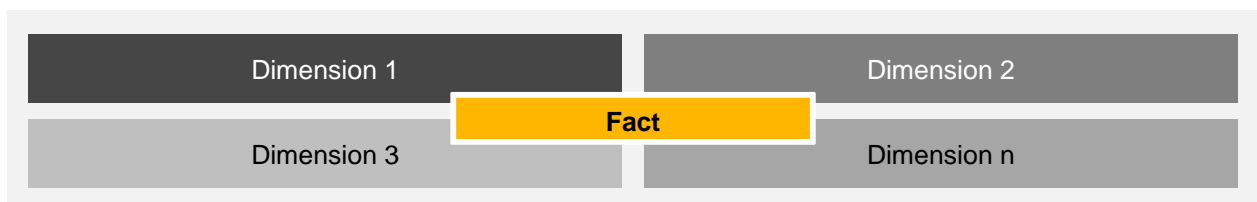


Figure 14. Facts and dimensions

A dataset can have from 1 to n dimensions. Facts can be micro-level facts or aggregated facts. If the fact is an aggregate, it can be a sum, an average or any calculation based on the initial microdata facts.

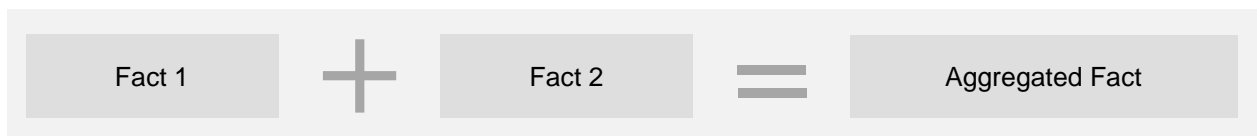


Figure 15. Aggregation of facts

**Importantly, when aggregating data, the possibility of data personalisation disappears.** It is essential for the DDDM system. Policy-making and decision-making require mostly non-personalised data.

In the example, the result of the analysis will be a three-column data table:

1. Age range value.
2. The number of employed persons.
3. The number of unemployed persons.

As can be seen, the structure of the input data does not match the structure of the output data. **The structure of the output data of the analysis** is always the same in the register and contains:

1. Aggregate value(s).
2. Dimension value(s).

A record can have more than one aggregate and more than one dimension. All visualisations can be generated using a data structure. The data structure is generally the same for all types of visualisations, regardless of what type of visualisation is needed. The difference lies in the dimensions that the visualisation uses.

**The input data for analysis** can be different in any data source, but they must be transformed into the same structure for the DDDM system analysis process:

1. Numerical fact(s).
2. Dimension(s).

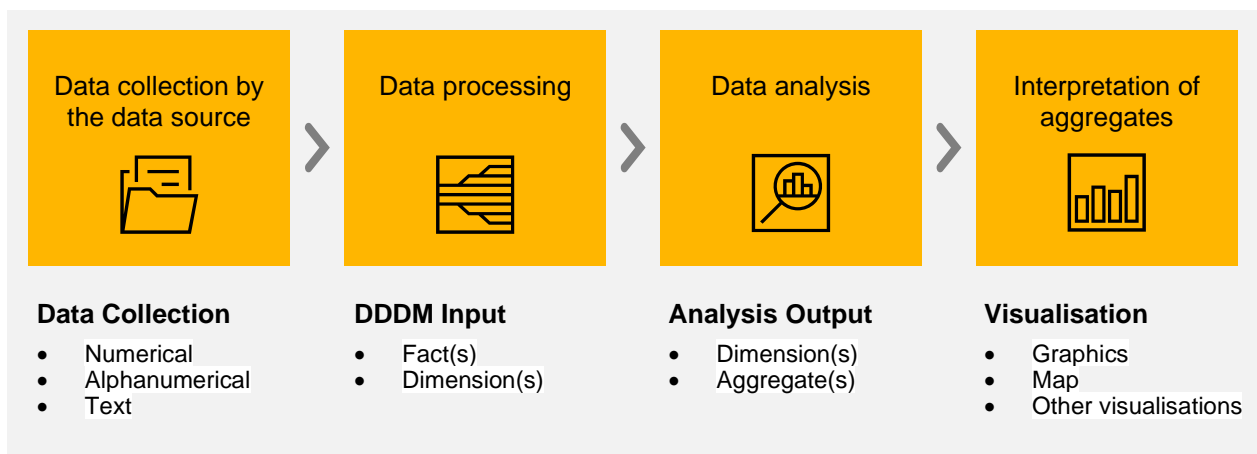


Figure 16. Data structure evolution

A dataset can have additional data columns, but the analysis process cannot use additional descriptive data in the input data for analysis. The user cannot process all microdata manually due to the large number of records. Most datasets are not human-readable in terms of getting an overview of the data. Transformations and aggregations are necessary to obtain an overview.

#### 4.1.2 Data Standardisation

If the purpose is to automate data analysis, the input data structure for the DDDM described earlier should be always the same. If there are multiple data sources, the output of each data source should be in the standardised format. The contents of the data are different in each case, but the format and structure can be and should be the same. Therefore, hundreds of data sources can be used in the DDDM system without the need to develop software for interfacing each data source.

Unstructured data and text data must also be transformed to the standardised fact-dimension structure. If a data source can produce aggregated data with the dimensions that the DDDM user needs, there will be no issues with personalised data and the user can obtain the overview they need. The aggregated data does not contain personal data and there is no need to protect the data in terms of GDPR in the DDDM system. It means the DDDM system can use the data as public open data.

#### 4.1.3 Prerequisites for the data standardisation

The DDDM system must have capabilities to:

1. Receive the standardised data.
2. **Require data sources to submit data in a standardised form.**

A data source must be able to:

1. Transform the data source's raw data into standardised DDDM input.
2. Require resources and funding to implement the standardised data source output format.

3. Have a data specialist developer who is proficient in the databases and can create a data transformation system.

## 4.2 Issues to be solved in the Data Structure

There are data issues that need to be addressed when developing the DDDM system:

1. If the data output of a data source contains only anonymised (aggregated) data, how can the DDDM system link the data from different data sources? An aggregate indicator does not have a linking value, e.g. a person ID code or similar identifier.
2. The aggregated data structure in the DDDM system input format does not match the original data structure of the data source – the transformation may be expensive and time-consuming for the data holder. The data holder can do the transformation, but the DDDM cannot do the transformation due to legal restrictions on access to the personalised data.
3. Data dimensions are not the same across data sources, e.g. address data, geographic coordinates or age ranges of persons are incompatible. The data requires additional transformation when dimensions are not compatible. If dimensions for the same subject are not comparable, it may not be possible to analyse data on that dimension.
4. All data in different places will never be in a standardised format, especially in external data sources from outside the public sector. It means that the DDDM system itself must also have data transformation and aggregation functionality.
5. Cross-linking between data from different data sources cannot be established when aggregating data. The problem could be solved in the systems of Statistics Estonia (SE). SE has approximately 100 microdata sources and processing of personal data is allowed by law. SE also has a dedicated technical environment where the DDDM user can perform data analyses and link data sources.

### 4.2.1 Cross-linking multiple data source data

The DDDM system has no rights to process some data; however, in some cases, it is necessary to link two or more datasets. **The linking can be done by Statistics Estonia**, which has access to approximately 100 datasets and is the preferred option. This is limited only by the data processing resources allocated to Statistics Estonia.

The second option – the right to process data will be assigned to the DDDM system. Linking is not possible without data processing rights and if the dataset does not contain specific fields that could be used as linking data or keys (i.e. personal code, date, business registry code, etc.).

Finding suitable keys is not possible without analysing both databases to be linked. The analysis is enabled by assigning the right to process the data.

### 4.2.2 Data source data transformation

Data transformation is necessary if the data source does not contain aggregated data the DDDM user requires and the data contains personal data or trade secret, and the DDDM user does not have sufficient rights to process such data.

There are the following solutions for the transformation of the data source data:

1. A data source generates an automated API of the aggregated data.
2. Data transformation in the data source is performed on a per-order basis. In this case, the data source must have a corresponding specialist.
3. The necessary data processing rights are obtained, and the transformation is performed in the DDDM system.

## 4.3 Logical Data Model of the DDDM Central Part

The DDDM system must have a database containing mandatory management data of the DDDM system. Data objects in this database can be:

1. Users.
2. User data stores.

3. User documents (memorandums, etc).
4. User projects.
5. Logs.
6. Other objects to be determined.

## 4.4 Data Integration Data Model of the DDDM System

Depending on the data types, the system should have a data integration data store that the system can use after the data integration process is completed.

### 4.4.1 Microdata model

Microdata must be stored in a relational database where each data source can have its own data model.

### 4.4.2 Aggregated data model

Aggregated data should be stored in the dimensional database. Data object there are facts and dimensions. These data come from the API of the aggregated data source if such APIs are installed. Similar data are generated daily in the data warehouses of authorities and ministries, but there is no standardised API yet.

### 4.4.3 Text data model

The text data should be stored in a document database with text query and analysis features. There is no state information system for documents, e.g. scientific publications.

### 4.4.4 Unstructured data store

There must be a repository for the data with an unknown structure. These data will be gathered from another data source with no metadata, as well as from csv files on the web. An unstructured data store has no restrictions on what data can be stored. It is feasible to establish data storage rules over time.

## 4.5 Analysis Data Model of the DDDM system

It is possible to have an analytical database with a specialised data model for smooth use of the data in analyses and visualisations.

It is possible to have two types of analysis data models:

1. Dimensional analysis data model.
2. The relational data model is in the second<sup>34</sup> or third normal form.

---

<sup>34</sup> [Second normal form - Wikipedia](#)

# 5. DDDM System Architecture

## 5.1 System Context

The DDDM system is a part of a larger system. It is imperative to establish interfaces between data sources to enable the user to view system output from different views, such as the KOOS user interface and documents generated by the system. If dynamic visualisation is required in system-generated documents, the dynamic content of the document should be an independent part of the system. It means that it is necessary to consider the possibility of using a distributed system architecture.

## 5.2 System Components

The system component is a part of the system, e.g. database server, data transformation engine, data source adapter, etc. The component implements one or more system functions. System components can be run on a computer. This chapter describes only system components. The DDDM system is a set of interconnected components that assist the DDDM user in drafting a Government Memorandum or other document that must contain data visualisations or the data itself.

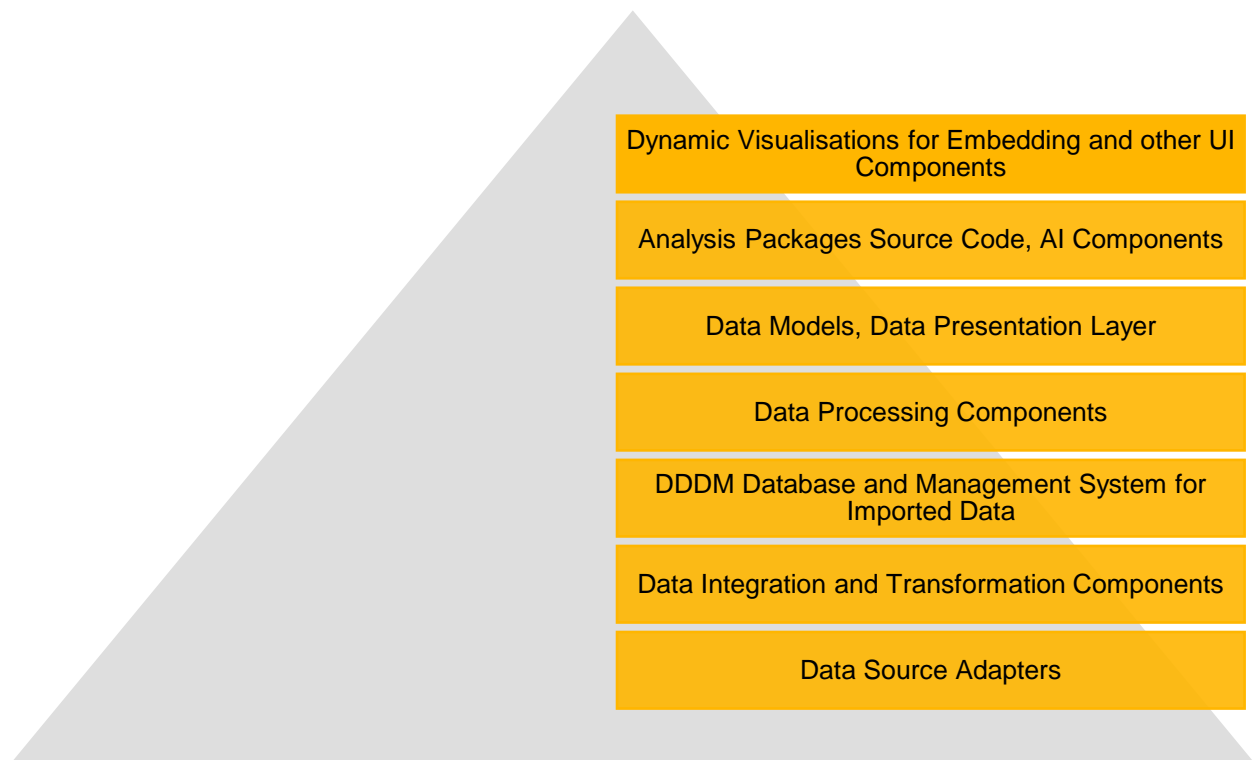


Figure 17. The DDDM system component layers

All DDDM system functions described above must belong to or be associated with a specific component. The implementation of the function may require more than one component, e.g. the data quality control function may require the database to store quality information in the DDDM system and a data source adapter to read data from a data source.

The level of system data security depends on the components and the components configuration of the system. In order to achieve the required security level, a specific set of components must be used to guarantee the functionality of the level. On the other hand, the set of components must meet the needs of the user. It means that the system is a sophisticated set of different components. The set must comply with mandatory user requirements and Estonian and European legislation. The system must prevent the user from performing unauthorised operations with data and assist the user in performing **mandatory and allowed** tasks.

This analysis follows the philosophy described above and should explain which components are needed by the user and which are required by law in the system.

### **5.2.1 Dynamic Visualisations for Embedding and other UI components**

For using analysis results in a document, e.g. in Government Memorandums, the analysis result or visualisation must be embedded in a web page (html) or Word document. The embedding result must be a dynamic visualisation in the document or on the web page. At least the visualisation image must be linked in the document, and the visualisation must exist on the server as a dynamic web application.

### **5.2.2 Analysis Packages Source Code, AI components**

In order to perform data analyses using already composed analysis code without programming the package from scratch, the DDDM system must have a repository of analysis source code. Reuse of analyses source code may be the most important capability of the DDDM system. The analysis package can be an AI component.

### **5.2.3 Data Models, Data Presentation Layer**

Without the data model, it is very difficult to perform sophisticated data analyses. Information about data structure is an important part of technical metadata. The DDDM system must have a system-aided data modelling component.

### **5.2.4 Data Processing Components**

The DDDM system must have data processing capabilities, e.g. to calculate aggregates based on the data source data or to classify and code the text data.

### **5.2.5 DDDM Database and Management System for Imported Data**

The DDDM must have a database for the transformed data. In addition, there is a set of management data for the DDDM system and data for the user choice suggestions.

### **5.2.6 Data Integration and Transformation Components**

Depending on the analysis needs, the data from a data source must be imported into the DDDM system or can be used in the data source itself. In both cases, the data should be transformed into a usable format. In most cases, it is not necessary to import all the data from the data source and the data must be transformed. For this purpose, the DDDM system must have data transformation components.

### **5.2.7 Data Source Adapters**

Each data source must be interconnected with the DDDM system if the dataset from the data source is to be used. The interfacing may be a trivial task or require software development. Development needs are related to the technology used by the data source system. If the data source system has a standardised database system, e.g. SQL Server or PostgreSQL, the interfacing tasks can be accomplished using ODBC or JDBC connection configuration. In a more sophisticated case, a development project may be required to create a custom interface. How complicated it can be, depends on the architecture of the data source.

There are legal restrictions on data processing, and it is probable that the data source should establish a data exchange application between the DDDM system and the data source. The application must aggregate the data and anonymise the data if the aggregation is insufficient to establish an acceptable level of data security. The aggregation and anonymisation cannot be applied in cases where the DDDM system needs personalised data to link them with data from other data sources. If data linking is necessary, the DDDM must have a legal right to process the data based on a legal act, or the data must be processed by Statistics Estonia, and they make the data available as an aggregated dataset.



## 5.3 Data Search Component Model

This chapter describes the data search component model. It is one specific view of the DDDM component model.

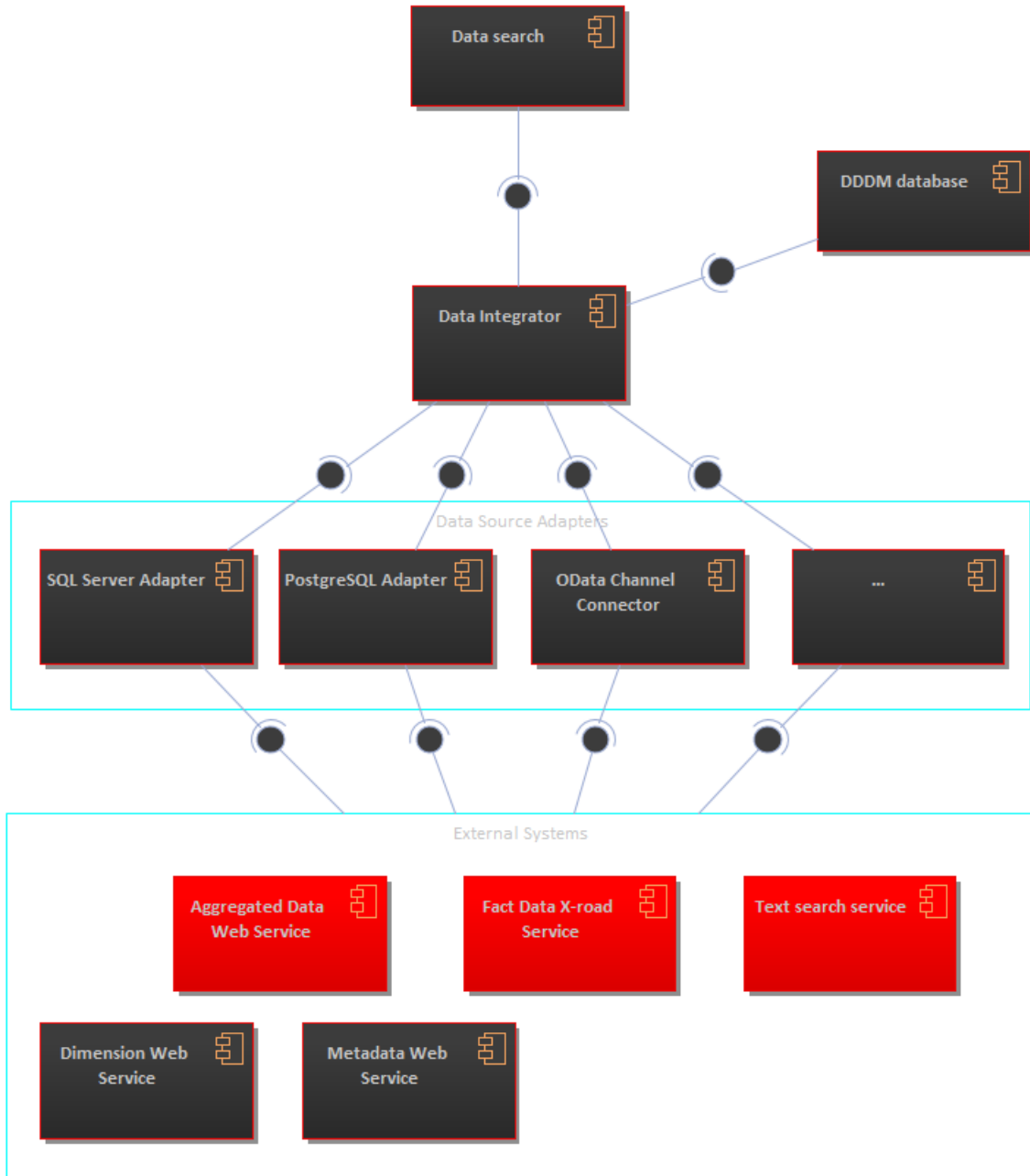


Figure 18. DDDM data search components.

### 5.3.1 Data Search

The data search is a core component of the search component package. This component uses the data integrator to obtain input data. Data search component package implements the functionality described in section 2.1.1.4.

### 5.3.2 Data Integrator

Data Integrator is a component for using and importing data from external data sources through data connectors. **An interface must be configured for each data source.** The data integrator uses the data from the data source and mediates the data flow. Data transformation also takes place in the data integrator, if it is necessary in a particular case. Transformation is necessary if the data from the data source has a non-standardised or unsuitable format for the DDDM.

Data integration is not a process where all data will be collected in a data warehouse. **There is no data warehouse in the DDDM system.** Only those data that cannot be used directly from the source are entered into the DDDM database. An automatic data update procedure should not be set up, as the data use is random without any specific period. A data warehouse with an ETL (extract-transform-load) process is necessary in reporting systems where the data use is continuous. The data use in the DDDM system is more based on ad-hoc analysis needs and is discontinuous. A periodic ETL process is not required, but an on-demand ETL is necessary.

### 5.3.3 DDDM Database

The DDDM database is an independent data store for DDDM system data. The data is intended to manage the system's operation. The system stores user choices and other relevant data in the database and will use the data for further suggestions.

### 5.3.4 OData Channel Connector

OData Channel Connector is an adapter for using data from the OData web service (<https://www.odata.org/>).

### 5.3.5 PostgreSQL Adapter

PostgreSQL Adapter is a component for reading data from a PostgreSQL server.

### 5.3.6 SQL Server Adapter

SQL Server Adapter is a component for reading data from MS SQL Server.

### 5.3.7 Additional Data Source Connectors

There may be more connectors for connecting data sources in the system based on real needs for connecting data sources. What data sources should be connected is currently unknown. There is no data source analysis to be done in the project.

### 5.3.8 Aggregated Data Web Service

It is essential to know which data sources should be in the DDDM system. Otherwise, it is impossible to establish a data connection. If it is necessary to establish several technologically different connections, it is a huge amount of work and the objectives of the DDDM cannot be achieved. The content of the connections should be standardised to the greatest extent possible.

Standardised data exchange services must be established to simplify the complexity of a data usage and connection setup. All data sources should have a data API that outputs aggregated (or anonymised) data to the DDDM. It is mandatory for the DDDM and could be used for other users of aggregated data. Statistics Estonia already has a REST API<sup>35</sup> for issuing aggregated data to data users.

### 5.3.9 Fact Data X-road Service

If data aggregation is not possible on the data source side or if microdata is required for analysis, a (standardised) fact data service is needed. It should be an X-road service due to the security requirements.

---

<sup>35</sup> <https://www.stat.ee/sites/default/files/2021-02/API-juhend.pdf>

The difference between a fact data service and an aggregated data service is that the fact service outputs microdata, while an aggregate data service collects all necessary facts before outputting the data.

### 5.3.10 Text Search Service

The text search service performs the search over text corpora by word or sentence. Search results can be a list of documents that contain the word or sentence, and the strength of relationship between the problem description and the text content. The result could also include the list of paragraphs in which the relationship is stronger. A simplified output must contain only statistics (number of documents), the relationship between which is above a certain level.

**It is crucial that the search is intelligent. It must consider words and sentences with the same meaning.** This means that the search is not just a word search.

A text search service must be on the data source side or on the DDDM side if the data is transferred to the DDDM. The DDDM system only uses the service and retrieves the search results.

### 5.3.11 Dimension Web Service

The dimension web service is a standardised dimension query service that outputs the list of dimensions in a dataset and the contents of the dimensions used in the dataset. Queries can be made one dimension at a time or all dimensions together in one response.

### 5.3.12 Metadata Web Service

The metadata web service is designed to query metadata on the aggregates and facts. It outputs a list of aggregates or facts, and their semantic descriptions.

## 5.4 Data source Interfaces

There are mainly two types of interfaces to be used in the DDDM:

1. Web service interface.
2. File or database interface.

Web service as an API is technically the best choice for data exchange between two systems.

File and database interfaces can be used if the data source does not have web services and does not have the capability to build such an API.

Web service, such as an API, is technically the best choice for data exchange between two systems. File and database interfaces can be used if the data source does not have web services and does not have the capability to build such an API.

The API can be a microdata API or aggregate data API. It depends on what information is needed for analysis and what is the legal basis of the required data.

# 6. Issues to be solved in the processing of different types of data

## 6.1 Catalogue of Issues

A dataset can have several attributes or features to be determined and considered before processing the data. Depending on the availability of a feature, different measures must be taken to ensure access to the data and resolve issues related to the feature. It is very important that the data containing personal data and/or trade secrets cannot be processed in the same manner as the data without personal data and trade secrets. If the data has these features and the DDDM does not have processing rights, the micro-level processing is illegal in the DDDM system, and the owner of the data source will never provide access to the data to the DDDM or the DDDM user. It means that an aggregation or anonymisation solution must be in place to retrieve data from the data source. All other features may also impose limitations on the use of data or restrict the data processing altogether.

These data features that can cause issues are described in the table below.

Table 3. Issues to be considered in processing different datasets

Issue code	Issue	How to consider?
I-001-UR	Unregistered data source. The dataset is unregistered and has no description in RIHA <sup>36</sup> .	If the dataset is not registered, the data gathering may be illegal, or the data is intended to be used by certain authorities to solve tasks arising from the law. Such data cannot have X-road services. The data source must be registered if X-road services are required.
I-002-PD	The dataset contains personal data.	If the data contains personal data, the data user must have a specific task arising from the law that requires this data. Without special permission from each person appearing in the data or legal right to process the data. One way to make data processing lawful is to hide personal data from the data through aggregation or anonymisation. The second way is to process the data at Statistics Estonia and the third way is obtain the right to process the data in DDDM.
I-003-TS	The dataset contains trade secrets.	If the data contains trade secret, the data user must have a specific task arising from the law that requires this data. Without special permission from each legal entity appearing in the data or legal right to process the data. One way to make data processing lawful is to hide trade secret data from the data through aggregation or anonymisation. The second way is to process the data at Statistics Estonia and the third way is obtain the right to process the data in DDDM.

<sup>36</sup> <https://www.riha.ee/Avaleht>

Issue code	Issue	How to consider?
I-004-IU	The data is for internal use only (for a public sector authority).	Data cannot be processed by the data user without permission from the authority holding the data.
I-005-NS	No system. The data is not gathered in a unified digital information system.	It is impossible to use data if there are no suggestions on how and where to find the data in real time.
I-006-NXRS	No X-road services for secure data exchange?	Data can be processed through a database or a link to a file. Obtaining authorisation for such activities can be difficult due to insecurity. The data user must have access to the network where the database server is located. It is not a secure way of sharing data from viewpoint of the data holder. The X-road is a legal and secure network for accessing data. This issue is not relevant for public or open datasets.
I-007-NMD	No metadata. Does the data lack descriptions and metadata at the data attribute or column level?	If the data lacks metadata, the workload of the data user can be significantly increased due to metadata creating work by the user, and conclusions from the data may lead to misleading decisions.
I-008-LDQ	Data quality of the dataset is not analysed.	If the dataset lacks quality control, significant errors can occur when drawing conclusions from the data. It must be considered, and the data user must take steps to verify data quality. Verification must be an automated solution. Manual quality check is not a sufficient solution and does not allow to find errors quickly.
I-009-TEXT	The dataset consists of text corpora – no well-defined data structure in the data content.	<p>The data in the text corpus cannot be analysed as a relational database. Statical analysis and aggregation are not possible without special processing of sentences in the content.</p> <p>Despite special content processing, the result of statistical analysis of the text data gives unreliable results in terms of the counting phenomenon. The result depends on a style of text, the number of text errors and the number of available documents. The only benefit of text data is to obtain additional information on how to describe and interpret the phenomenon mentioned in the text in addition to what the author of the text has already analysed and done on the topic. Statistical analysis of multiple text corpora often provides unstructured aggregated overviews in certain areas or fields of activity, as the data being analysed is unstructured, and some facts may be not detected. Text analysis can help in obtaining information about trends if the amount of text (documents) is sufficient and representative. A prerequisite of statistical text analysis is the availability of an appropriate text analysis tool that can classify and code texts.</p> <p>There are several code repositories in Estonia containing such code examples, as well as several commercial products; however, there is no free product for DDDM integration.</p>

Issue code	Issue	How to consider?
I-010-UN	Unstructured data.	The dataset structure is unknown, or the structure changes over time. It means that the structure can be of any kind. Detecting the structure is an additional task for extracting information from such datasets. Aggregates based on unstructured data are not reliable. Unstructured data can be used for discovering insights about what happened in the past. Clear statistics cannot be calculated. Unstructured data usage is beyond the scope of DDDM.
I-011-ST	State secret. The dataset contains state secrets.	If the dataset contains state secrets, the data are not subject to analysis in the DDDM system both at the level of aggregated data and microdata.
I-012-PSD	Private sector data. The data is owned by a private company.	If the data is owned by a private company, the DDDM cannot use the data unless there is a legal requirement to provide the data to the state. A company can agree to provide the data to the DDDM but is not required to. The once-only principle is an e-government concept that aims to ensure that citizens, institutions and companies must provide certain standard information to the authorities and administrations once. It means that the DDDM should detect whether the company has already submitted the data to the government. In Estonia, there is an "Arvandlus 3.0" initiative which provides for automation and simplification of data submission to the government. The initiative is still under development, and it is possible that the DDDM will be able to retrieve metadata on the submitted data (reports) from a special Estonian data collection that will be created in the near future. Currently, there are 421 periodic reports with 57,973 data fields in Estonia <sup>37</sup> . This is only a part of the data that companies have to offer.
I-013-NADA	No aggregated API data	The data source does not have an API for aggregated data. Those can be created using data warehouses that most of public entities have.

37

<https://app.powerbi.com/view?r=eyJrIjojOWRmMTIyYmMtOTE4Yi00ZWZkLWlxZjMtYmI4MTU4NDgxMDIzliwidCI6IjRlOWM2MDRhLTUwNDMtNDQ2YS1iYzk4LTgxNzdmNmVlYTliNyIsImMiOiI9&pageName=ReportSection976866578e-ea6520d5b4>

## 6.2 Establishing aggregated Data API-s for Data Sources

Building an API of aggregated (statistical) data for each data source is an option for solving the issues of personal data and trade secrets processing.

Building an API of aggregated (statistical) data for each data source is an option for solving the issues of personal data and trade security data processing. If a data source creates an API to issue aggregated data for a data user, there will be no issues with protecting personal and trade secret data.

The API must have:

1. Data aggregation engine or data warehouse on the data source side.
2. Web service layer for data users. The API must have X-road REST or REST services. The X-road is mandatory if the data cannot be published as an open dataset.
3. Data security checker that controls all data in terms of security sent to the user. The web service output must contain aggregates. If the response to a query contains an aggregate corresponding to only one microdata record, the system must not issue the data.

The input of the web service must have:

1. The code of the aggregate what the query must contain.
2. Dimension by which the aggregate must be calculated.
3. Aggregate's start and end date.
4. Aggregation period for the time series of aggregates (e.g. day, week, month, etc.).

The output of the service must have:

1. Requested aggregate values.
2. Time period identifiers.
3. Unit of measure of the aggregate (e.g. piece, euro, kg, km, etc.).
4. Dimension by which the aggregate was calculated.
5. Metadata. Description of the aggregate.

It is important to agree between the DDDM and a data source on which aggregates and dimensions the data source system can issue to the DDDM. This agreement should be a legal act (or a written agreement) with metadata on facts, dimensions, and aggregates to be exchanged.

**This solution solves issues arising from I002-PD, I003-TS and I006-NXRS.**

## 6.3 Synthetic Data

The synthetic data<sup>38</sup> solution should also be tested. A synthetic dataset is an AI-generated microdata dataset. The input is microdata with personal data or other non-public data, and the output is microdata without data about real persons.

Synthetic data could solve the issue of personal data processing; however, the issue of data linkage will arise if there is a need to cross-link multiple data sources.

Synthetic data generation is not a well-known technology; however, it may help to solve the issue of limiting the processing of personal data. This technology should be implemented in the data source system. The exchange of data is similar to the exchange of aggregated data.

---

<sup>38</sup> [https://en.wikipedia.org/wiki/Synthetic\\_data](https://en.wikipedia.org/wiki/Synthetic_data)

# 7. Methodological Guidelines

## 7.1 Obtaining Data Access

To gain access to the data, the user must commit to the correct use of the data. It may be complicated depending on the legal basis of the dataset. Estonian legislation and European law must be considered. Data usage rules may be set in a special legal act that is established by the owner of the dataset. There is no general legal basis for authorising the DDDM system to access these datasets.

Processing rights for DDDM can be set in various ways as described in chapter 3.3 Legal Analysis. The preferred way is to publish aggregated data or use other measures such as synthetic data, data anonymisation or shared computing for DDDM.

The process of authorising the user to access a specific dataset is shown in the following figure.

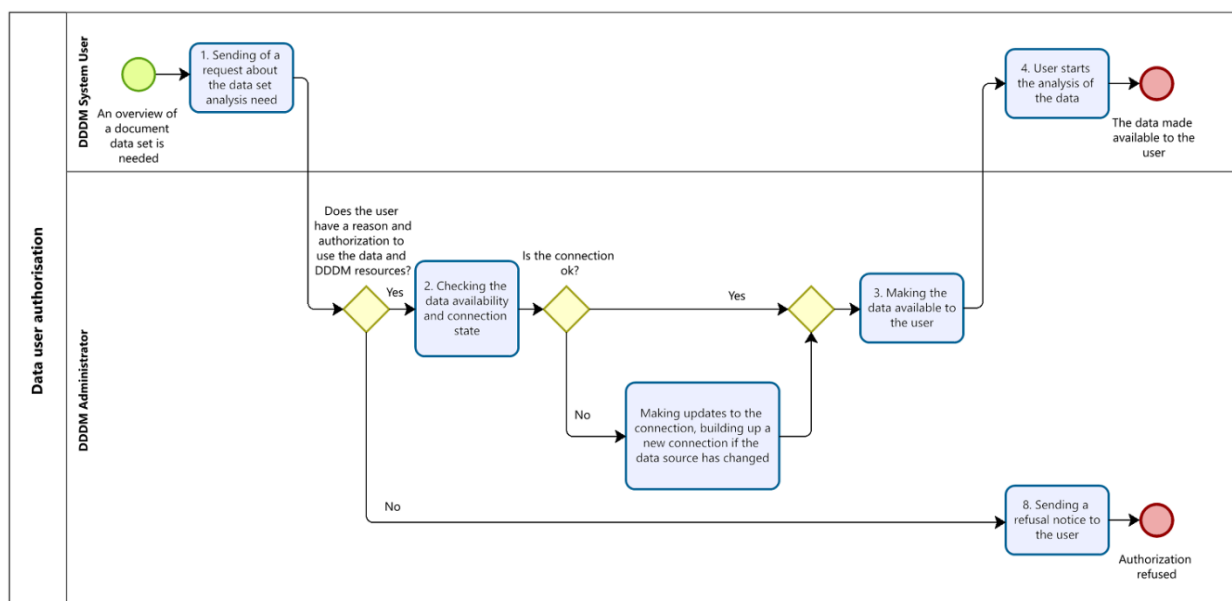


Figure 19. Dataset user authorisation process.

## 7.2 Data Processing Methodology

In the DDDM system, data processing should be automated to the greatest extent possible. There are the following steps of data processing:

1. Gathering and integrating the data from data sources based on analysis needs.
2. Classifying and coding the data.
3. Data quality control.
4. Imputing missing data (values) from additional data sources.
5. Calculating aggregates.
6. Transforming the data into the suitable (fact-dimension) format.

All these tasks should be performed automatically or manually. Some tasks may be performed within the DDDM system and some outside of the DDDM system.

Deciding on a data quality assurance tool is important. The tool may be developed within the DDDM system, or it can be a separate tool. The data quality review process must cover the following:

1. Missing values in the data.
2. Dates are in suitable ranges.
3. Classifications are in place.
4. Relationships' integrity.
5. Duplicates in the data.
6. Start time must be before the finish time; birthdate must be before the date of death.



## 7. Domain-specific business rules on the data.

Domain-specific business rules may be checks based on many data sources data. If the same indicator has different value in different data sources there may be a problem in the one of the data sources. The user must check is it an error or difference between methodologies.

Domain-specific business rules are also multi-record rules where checks are performed on multiple records. It is a data object cross-checking process.

At the time of preparation of this report, the guideline on data quality control is being developed by Statistics Estonia. This task is part of the national action plan of data governance (task no 4.4.)<sup>39</sup>. The aim of that guideline is to ensure that the data quality criteria and requirements for data collection and data analysis are practical and implementable. The manual should also support the developments in data management applications - in particular, possible data quality compliance functionalities in RIHA and RIHAK. Thus, the DDDM system is recommended to coordinate its plans and tasks with key stakeholders in Estonia.

Furthermore, it is important to recognise that according to the regulation on Principles for Managing Services and Governing Information,<sup>40</sup> the public authority is responsible for the information governance and quality thereof. Thus, the DDDM system cannot ensure data quality itself, but should evaluate and inform the users about the data quality as well as the data holders.

## 7.3 Data Analysis Methodology Cornerstones

Cornerstones are:

1. The main data analysis is based on statistical methods.
2. Each dataset and analysis can be reused in the DDDM system.
3. The analysis must be automated to the greatest extent possible.
4. The user of the DDDM system should not be a data analyst.
5. The data analysis is based on the dimensional analysis data model.
6. The DDDM must generate and provide metadata to the user if a dataset has no metadata. Metadata generation is possible if the data quality is sufficient for generation (linking columns are filled with relevant data, etc). Metadata generation is based on an automated machine learning solution.

## 7.4 Data Usage of Data Sources Data

In order to use data from the data source, it is important to understand how the analysis relates to the data source and what kind of data the data source provides. Chapter 6 describes the aspects to be evaluated before starting the analysis. Detected issues and problems must be resolved beforehand.

## 7.5 Sticking Visualisations into a Document

There must be able to visualise data analysis in a document. This can be done in several ways:

1. Sticking an image of visualisation to the document. It is a static image in the document.
2. Linking the document to a visualisation located on the analytical web server. The reader of the document can click the link and view the visualisation on the analytics server web page. It is a dynamic way to inspect the data analysis.
3. Placing the object of data analysis with visualisation in the document. It means that both the analysis results and the dataset are parts of the document. It can only be done with a smaller dataset. The positive aspect of this case is that the analysis result will remain the same as it was at the time of drafting the document.

---

<sup>39</sup> <https://digiriik.ee/index.php/andmehalduse-tegevuskava/>

<sup>40</sup> <https://www.riigiteataja.ee/en/eli/ee/502062021006/consolide/current>

# 8. Appendices

## 8.1 Extended Legal Analysis in Estonian

An extended version of the legal analysis of the DDDM system in Estonian is attached to Deliverable 1.4. Please see the separate document named D1.4\_DDDM\_system\_legal\_analysis\_EST.

## 8.2 List of conducted Interviews

Table 4. List of Interviews conducted

Date	Organisation	Participants, organisation/role
11.10.2022	Government Office	<b>Dmitri Burnašev</b> , Deputy Strategy Director <b>Erik Ernits</b> , Head of Data
18.10.2022	Government Office	<b>Dmitri Burnašev</b> , Deputy Strategy Director <b>Erik Ernits</b> , Head of Data
25.10.2022	Government Office	<b>Dmitri Burnašev</b> , Deputy Strategy Director <b>Erik Ernits</b> , Head of Data
01.11.2022	Government Office	<b>Dmitri Burnašev</b> , Deputy Strategy Director <b>Erik Ernits</b> , Head of Data <b>Ivar Hendla</b> , Strategy Adviser
03.11.2022	Government Office	<b>Erik Ernits</b> , Head of Data
07.11.2022	Government Office	<b>Erik Ernits</b> , Head of Data
08.11.2022	Government Office	<b>Dmitri Burnašev</b> , Deputy Strategy Director <b>Erik Ernits</b> , Head of Data
11.11.2022	Government Office	<b>Erik Ernits</b> , Head of Data
11.11.2022	Ministry of Environment, Republic of Estonia Environment Agency – Meeting to discuss data availability for Proof of Concept	<b>Rauno Künnapuu</b> , Ministry of Environment <b>Kristel Kund</b> , Ministry of Environment <b>Kertu Sapelkov</b> , Ministry of Environment <b>Anneli Averin</b> , Republic of Estonia Environment Agency
15.11.2022	Government Office	<b>Erik Ernits</b> , Head of Data
17.11.2022	Government Office	<b>Erik Ernits</b> , Head of Data
22.11.2022	Government Office	<b>Dmitri Burnašev</b> , Deputy Strategy Director <b>Erik Ernits</b> , Head of Data <b>Ivar Hendla</b> , Strategy Adviser
22.11.2022	Government Office	<b>Erik Ernits</b> , Head of Data
25.11.2022	Ministry of Environment	<b>Rauno Künnapuu</b> , Ministry of Environment
29.11.2022	Government Office	<b>Dmitri Burnašev</b> , Deputy Strategy Director <b>Erik Ernits</b> , Head of Data <b>Ivar Hendla</b> , Strategy Adviser
1.12.2022	Government Office, Ministry of Environment	<b>Erik Ernits</b> , Head of Data of Government Office <b>Rauno Künnapuu</b> , Ministry of Environment
6.12.2022	Government Office	<b>Erik Ernits</b> , Head of Data <b>Ivar Hendla</b> , Strategy Adviser
8.12.2022	Government Office, Ministry of Environment	<b>Erik Ernits</b> , Head of Data of Government Office <b>Rauno Künnapuu</b> , Ministry of Environment

Date	Organisation	Participants, organisation/role
9.12.2022	Government Office, Statistics Estonia	<b>Erik Ernits</b> , Head of Data of Government Office <b>Veiko Berendsen</b> , Statistics Estonia <b>Maiki Ilves</b> , Statistics Estonia <b>Kaja Sõstra</b> , Statistics Estonia <b>Ott Karp</b> , Statistics Estonia
05.01.2023	Government Office	<b>Erik Ernits</b> , Head of Data
09.01.2023	Government Office	<b>Erik Ernits</b> , Head of Data
10.01.2023	Statistics Estonia	<b>Maiki Ilves</b> , Statistics Estonia



Funded by  
the European Union

Visit our website:



Find out more  
about the Technical  
Support Instrument:

