Reporting model

Business analysis

Procurement reference number 231015

MAJANDUS- JA
KOMMUNIKATSIOONI-
MINISTEERIUM

Euroopa Liit
Euroopa
Regionaalarengu Fond

Eesti
tuleviku heaks

tietoEVRY

Contents

Introduction

This document describes the methodology of handling reporting data with an of developing guidelines for designing reporting datasets and handling data. The document was developed based on the vision of real-time economy[1].

Real-time economy is a digital ecosystem in which the transactions between different parties occur in real time or with a minimum delay. This means replacement of paper-based economic transactions and administrative operations with automatic data exchange in a digital, structured, machine-processable, and standardised format.

This document is designed for the individuals who are working with the reports and data submitted by businesses and must participate in the respective system and data analysis projects that are concerned with identifying the need for data, designing data structures, and analysing data.

The capability for system analysis is the basis for designing efficient datasets. This primarily applies to the designing of data structures and the requirements set for the system. A specialist without appropriate knowledge may start describing data compositions, but may not be able to finish the process due to the lack of system analysis skills. **Thus, by this document, we recommend using specialists of the field (system analysts) for designing data.** Not using specialists may result in confusion and duplication, which will be difficult to fix late when the real data have already been saved in the system.

This document gives advice and provides a framework for which methods and standards to use in the modelling, collection, and processing of reporting data.

Handling of data begins from the analysis of the composition of the dataset required by the consumer of specific data, identifying which data elements the dataset collected should include. Real data should also be examined in the analysis to determine what can be used and in which format.

This document (model) is compliant with the following requirements described in the public procurement (reference number 231015):

1) The entire reporting obligation placed on the businesses has been mapped and respective documents have been drawn up, identifying the number of the respective state agencies, the number of reports requested by them, and the manner of data collection. The information used in the course of the mapping was taken into consideration in developing the model.
2) The model developed explains to the state agencies requiring reporting from businesses the analysis of data compositions, including the description of how to assess the timeliness and actual necessity of the data fields and suggestions for updating the legislation, standardisation, development of a uniform taxonomy, and the processes for the implementation of the XBRL GL

[1] https://www.mkm.ee/sites/default/files/reaalajamajanduse_visioon_2020-2027_kaskkiri.pdf

standard, and an analysis of the perspective of the business has also been taken into consideration.

3) The model developed allows each state agency to perform detailed operations within the framework of a specific report or dataset, to use the common taxonomy elements that have already been created in the data compositions (information from Statistics Estonia) and to create any missing ones, to implement the XBRL GL standard, and to perform the development required for the receipt and processing of machine-readable data and for the data exchange between state agencies.

4) The model developed will help to prevent situations in which a state agency only settles a problem from the perspective of their own agency, making submission of data to the state more complicated and fractured for businesses. Using the model, the prerequisites required for a situation in which the data submitted by businesses to the state is requested only once and can be reused by different databases and agencies were created.

5) The model developed describes what, in which order, and how to do to enable real-time automatic movement of data between different agencies and businesses. Thereby, thanks to the model created, it will be possible to draw up more cost-efficient and aligned development projects in the future to make the reporting obligation simpler and more automatic and to avoid the need for unnecessary business analyses.

The work was commissioned by the Ministry of Economic Affairs and Communications. The project is co-funded from the European Regional Development Fund.

1. Definitions

Definition	Explanation
Aggregated data	Summarised data, such as the economic indicators of a business in a specific period.
Data processing	Creation, reading, amendment, deletion, transforming, reflecting, or exchange of data (organising of data traffic).
ERD	The Entity Relationship Diagram used for describing data models.
ERP	Enterprise Resource Planning – accounting software used for managing the operations of a business.
Hierarchic data structure	A data structure in which data objects of different types are located within one another so that there is no cross-referencing between different objects. For example, the data structure in a XBRL GL XML file is hierarchical.
Classification	A classification is a list of unique codes and names designed for the classification of phenomena. Each element of a classification must have a code and a name and an element may also have further attributes, such as names in other languages or a feature which makes the classification of the element more easily recognisable. Classifications may be linear or hierarchical. In a linear classification, there are no connections between the elements of the classification. In a hierarchic classification, the elements may in turn be divided into elements, i.e. there may be hierarchic relationships between the elements. In this document, classifications include both national classifications as well as other lists, incl. code lists designed for the classification of data objects nationwide or within one specific report. Thus, the concept of a classification is wider in this document compared to other situations and the law on the system of classifications
Classification element	A unique sub-part of a classification defined by the code name and other data attributes. The codes of different classification elements do not overlap.
List or code list	A list or code list is a list of values used as the value of a data object attribute for the classification of the data object.

Definition	Explanation
Variable	Description of a data element. A variable is tied to a specific indicator characterising a certain object or subject at a certain moment in time. For example, the turnover, profit, or number of employees of a business in a certain period. Indicators form a timeline, which is one of the most important statistics and analysis tools, characterising trends and the situation with respect to an object or subject.
Indicator	Data in the form of a number or characters (text) equivalent to a certain variable described.
Process	A sequence of operations. A process is divided into operations which may be in the linear as well as recursive or branching sequence. Recursions may cause repeated performance of the same operation. A sequence of operations performed based on pre-determined rules can be referred to as a process. Any sequence of unrelated operations is not referred to as a process in this document.
Real-time data	Real-time data are the data made available to the consumer by the creator of the data immediately after creating the data. Real-time does not mean that the data reach the consumer of the data immediately from the temporal perspective. The consumer may not always be prepared to use the data. Thus, data are real-time data if they are made immediately available in an agreed format and data exchange channel.
Relational data structure	Relational data structure contains the data of the same object or subject only once. It differs from the hierarchical structure where the same data may be found in the structure several times in different data objects.
Schematron	*Schematron*[2] is a rule-based data validation language which enables the identification of the presence or lack of patterns in XML data. *Schematroni* is used for defining and implementation of the inspections of XML data to determine the quality of a specific XML file based on the rules defined.
Taxonomy	In this context, taxonomy is the description of a dataset which includes the list and descriptions of the data objects and data

[2] https://schematron.com/

Definition	Explanation
	elements, as well as classifiers and other important information about the structure and content of the dataset. Taxonomy or the content of the description of a report is described in detail in the following section of this document: '2.2.2. Outcomes of describing the composition of data'.
UML	UML[3] or Unified Modelling Language is a wide-ranging language for describing systems and data. UML is used in system models as well as in the European standards on the functioning of data and systems. UML should be used for describing all systems and data which are related to reporting. UML establishes uniform rules for the limitation and form of models.
Correlation table	A correlation table of two classifications in which each row characterises the correlation of two classification elements. Please note! Any one classification element may be correlated with one or several other classification elements.
XBRL GL	XBRL GL or *XBRL Global Ledger* is special structure of XML oriented to the submission of reporting data. While XBRL is mainly oriented to saving aggregated data, XBRL GL also enables saving individual records, for example, the data of economic transactions by individual transactions. XBRL GL enables the submission of data at the level of individual records. It is important to point out that XBRL GL is the most efficient for the real-time submission of data and for cases in which the number of individual records to be sent together is not very high.
XSD	XSD[4] or XML Schema Definition is the description of the XML structure that defines what must be contained in an XML file.
XML	XML[5] is a tool for saving and transporting data and it is independent from hardware and software.
Individual records	A data entry about a specific subject or object. An individual data entry does not include summarised records.

[3] https://www.uml.org/what-is-uml.htm

[4] https://www.w3schools.com/xml/schema_intro.asp

[5] https://www.w3schools.com/xml/xml_whatis.asp

2. Modelling and standardisation of data

Modelling and standardisation of data are required to create the prerequisites for efficient and error-free use of datasets. Standardisation of data facilitates reusing data. One of the examples of a lack of reuse is different agencies using different classifications that are the same in principle, but differ by details, to classify one and the same thing. This means that the classification of the dataset is not uniformly understood by different agencies. Thus, the data are not usable or can only me made usable by making an additional effort in the form of recoding the data manually, programming automatic processing, or using the data with the help of correlation tables.

Repeated submission of the same data to the state may demand double work from the business. For example, if a business submits their turnover data to the Tax and Customs Board in a summarised form and partly in the form of individual records (i.e. invoices), they must, in principle, submit two different types of data. The business should submit all of their invoices or grant access to the invoices. In this case, the state would be able to perform the analysis required and the business would be freed from the obligation to draw up a declaration based on the data. For the state to be able to process the data of the businesses, the data must be in a standardised form, such as in the form of an accounting entry in the case of an economic transaction or in the form of an e-invoice in the case of invoices, and equipped with the required classification data. Based on the latter, the information system of the Tax and Customs Board could machine-read the content of the economic transaction and would be able to process the entries in the analysis properly.

In order to enable reuse of data, the semantics of the data must be defined, i.e. the content of the data must be described. Comparison of the semantic descriptions of different data should enable determining whether the data in question are the same data. If the semantic definition of the data need overlaps that of a dataset which is already being collected, it may be assumed that the data collected enable satisfying the data need. The data may be usable in the direct or transformed form, for example aggregated to a certain level. Transforming the data to a suitable form enables reusing the data in different reporting processes without having to collect and submit them again. As a rule, reuse does, however, call for further processing in the form of making extracts or transforming the data into another form, incl. to another level of detail.

If modelling of data reveals that the data required could be subjected to a standard that has already been established or a commonly acknowledged standard, such standards should be used. This applies to e-invoices and bank transfers, for example, as well as to the data generated by agricultural machinery in the course of operation, in the case of which there are already standards established and introduced. If there are no standards, such standards should be described and enforced. If there will be a constant data need in the future which calls for configuration of the information systems, there must be specific rules established. In this case, a standard is an agreement between several parties that does not necessarily have to be formalised as an official standard.

The drawing below presents the process of modelling and standardisation of data.

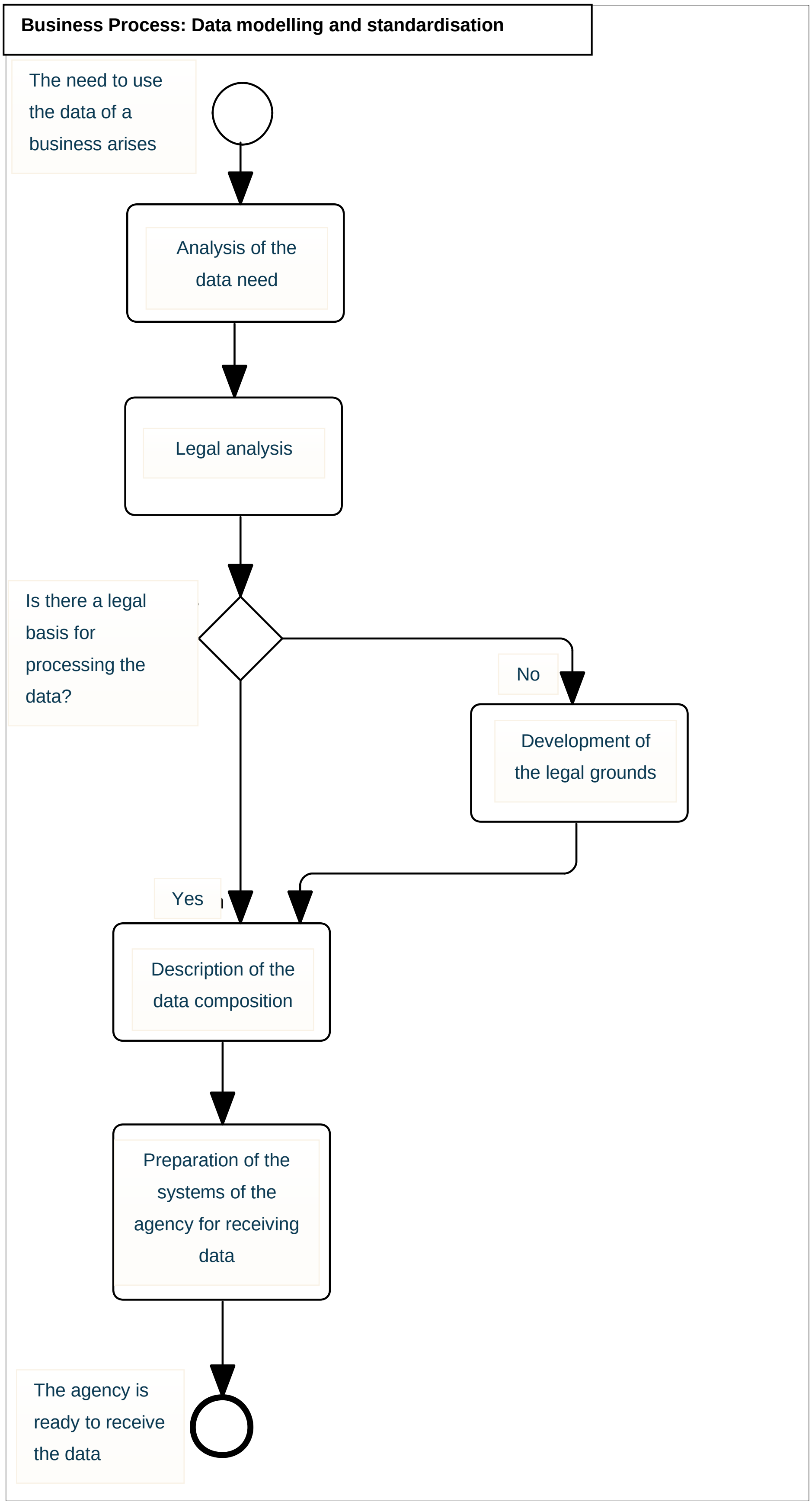
Business Process: Data modelling and standardisation
The need to use the data of a business arises
Analysis of the data need
Legal analysis
Is there a legal basis for processing the data?
No
Development of the legal grounds
Yes
Description of the data composition
Preparation of the systems of the agency for receiving data
The agency is ready to receive the data

Figure 1. Modelling and standardisation of data

Below, we describe in detail how to organise analysing data needs, modelling and standardisation.

2.1. Identification and analysis of the data need

For requesting data from data submitters to be justified, the following must be identified:

1) which data exactly are needed;
2) **which purpose** the data are needed for;
3) how often the data are needed;
4) **is there a legal basis for requesting the data**;
5) **is the data collection justified from the economic perspective**;
6) **is the data already collected by the agency or by some other agency**.

In order to answer these questions, state agencies must conduct a system analysis that involves interviewing relevant agencies about all of the afore-mentioned issues and comparing the data need to the data descriptions that have already been developed or using the central catalogue of taxonomies drawn up by the state that issues the information required to the agencies. The state agency must also conduct a legal analysis on whether the data collection and use are compliant with the legislation. In addition to the above, the economic impact (incl. administrative burden) of the data collection and use on the businesses as well as on the state agency must be assessed.

If there are no specific data descriptions, a **data audit** should be carried out to check whether or not the authority already has the data required. The data audit also involves assessment of the quality of the existing data.

As a result of the analysis of the data need, the economic impact assessment arising from the processing of the datasets (reports) will also be formed, among other things. **The impact assessment of a report is the criterion for the assessment of the level of priority of developing the reports.** The higher the positive impact, the higher the level of priority of the work done with the report. The impact assessment also shows **whether the administrative costs of the investment planned for the transfer to data-based reporting are lower or higher compared to the previous report-based information exchange or remain the same** if the impact of report-based information exchange is also assessed.

2.1.1. Analysis of the need for data

For **registration** and mapping of the **data need**, the agency must establish a register of the variables (see the definition of a variable under 'Definitions') that they intend to use in the future. It should also be kept in mind that variables may form a set. Handling of a set is often more reasonable than collecting and processing data separately based on each variable. For example, individuals and dwellings are data objects of completely different types, but those data are collected as a set in the course of a census, as the datasets are tightly interconnected.

As a rule, the following is required for taking into use each new variable:

1) identification or creation of a legal basis for the processing of the data;
2) an investment into the information systems which collect, transmit, process, and publish new data – if an automatic data collection system must be created, the data submitter must also invest into the development;
3) an investment into hiring and training the staff involved with collection, submission, and processing of data at the state agency as well as private businesses.

Before making any investments, the size of the investment and **whether or not the positive (financial) impact of using the new data outweighs the investments made and the expenses, when will the estimated break-even point be reached, and how high is the return on investment (ROI) index** should also be assessed. The impact should be assessed based on the peculiarities of the sector in the extent of a certain period. A 5-year period is often used. Both the initial investments and expenses, as well as the expenses on administration and maintenance of the systems in the following years should be taken into consideration. The impact assessment must be conducted for the state agency as well as for the private business that will be submitting the data.

If the balance is positive, funding should be found for collecting the data which correlates with the variable. If the benefits are smaller than the estimated costs, the data processing process will be eliminated sooner or later.

The methodology used in the study of the economic impact of real-time economy should be used for the assessment of the impact of data collection[6].

Before taking a specific variable into use, it should be checked whether the respective data are already being collected by Statistics Estonia or another agency. The indicators published are available on the website of Statistics Estonia. The descriptions of the data collected have also been drawn up, but those have not been published and should be requested from Statistics Estonia. When it comes to other agencies, there is currently no solution for determining in one certain manner whether or not a certain dataset is being collected. By using the methodology described in this model, descriptions of data compositions will be formed in the future and these can be used to answer this question. If a variable in use, a legal basis must be created for the cross-use of the data and the existing mechanism must be used for obtaining the data. This approach will enable reducing the administrative burden. At this point, the administration system of the state information system (RIHA)[7] is used for the identification of the datasets processed in Estonia, in which all data collected or created by a dataset should be described. Unfortunately, RIHA is currently not sufficiently updated. The register is not vital for data controllers of datasets and the data are basically only updated in the register when it is unavoidable. An agency-based register of datasets that includes the descriptions of all datasets collected in the form of a taxonomy could provide a solution here. **Please note! It is important that the register of data compositions is updated**

[6] https://www.mkm.ee/sites/default/files/reaalajamajanduse_majandusliku_moju_uuringu_lopparuanne.pdf

[7] https://www.riha.ee/

when new data are added or the structure of the data is changed. Such technology is not presently used in RIHA. Reverse engineering of databases or reading and visualising information about the structure from the database is a well-developed technology that must be taken into use. This can be done by using SPARX Enterprise Architect and other software applications, for example. **A proper register of the datasets of the agency updated in time is the prerequisite for other agencies being able to navigate the datasets of the agency in order to reuse them.** It is also important to highlight that automatic scanning of a data structure only works if the reports are saved in a database in which the exposed logical data structure is exactly the same as the physical data structure. **XBRL GL is a data structure in which the logical structure of the data has not been exposed.** If the data is kept in the XBRL GL format, automatic scanning of the data will not provide an overview of the logical structure of the data. Thus, the physical data structures must be designed so that they could be used to determine the logical data structure by using specific rules.

2.1.2. Legal analysis

When it comes to the legal basis for requesting data, it should be checked whether processing of the data by the party requesting the data is lawful. If processing of the data is prohibited or is not enabled by the law, the data and the data processing process should be modelled at the general level and respective legislative acts should be developed and enforced to legalise the processing of the data. Without the permission and support arising from the law, there is no point in continuing to build the systems for data processing. If the legislator does not approve the right to process the data, it is not permitted to collect or use real-time data (or so-called live data). **The legislator forms its opinion based on the rights and needs of all stakeholders, thus proceeding from whether or not collection of the data is justified. This position must be taken into consideration.** The legislation of the European Union (incl. GDPR)[8] establishes strict rules for data processing and violation of those rules results in big penalty payments. State agencies can also only place obligations (to submit data) on businesses directly based on law.

The right to process data has so far mainly been established in Estonia by:

1) a respective act of law;

2) a regulation of the minister of the respective sector;

3) a regulation of the Government of the Republic of Estonia.

In addition to those, there are other options arising from the law to obtain the right for data processing. Those options should be identified separately in each specific case. In addition to the Estonian law, the right to request and collect data may also arise from the European Union law. The statutes of the database must also be developed and enforced to create a permanent data register. If a database is created by statutes which name the data controller, the respective data controller may authorise another person to process the data. This person is referred to as the data processor.

[8] https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en

It is advisable to perform a separate legal analysis for the processing of each dataset and to update the legal basis for the activity based on the analysis. Without the legal basis, it is not possible to request or process data. The development of the databases which do not have a legal basis is not funded from the state budget and they also do not have access to grants. In most cases, the funding party checks whether or not a specific database is described in RIHA and whether its operations are legal.

A consent service[9] is also being developed in Estonia and it will enable the owner of data to grant access to their data. Thus, if an agency is able to obtain the consent of the required circle of persons through the consent service, this will also settle the legal issues concerning data processing. The consent service is suitable for obtaining permission for smaller-scale (related to a limited number of persons) data processing.

In the near future, the principles of data management, incl. processing, will also be regulated at the EU level. In order to take into consideration the data sharing and reuse regulations planned, the proposed data governance act of the European Commission should also be read: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020PC0767&from=ET. For example, the proposal in question includes a principle for the conditions for data sharing: 'Conditions for re-use shall be non-discriminatory, proportionate and objectively justified with regard to categories of data and purposes of re-use and the nature of the data for which re-use is allowed. These conditions shall not be used to restrict competition'.

2.2. Description of a data composition

A description of data composition or the data model must be designed based on the actual need for the data in the public sector, taking into consideration which data can be collected and received from the data submitter at reasonable expenses. In a simpler case, a description of the data composition is a list of the data elements required. In most cases, however, the models are more complex, involving groups (tables) formed by data elements with hierarchical or other types of connections between the groups. This means that data compositions must be described in the form of a data model. Otherwise, the description will not be usable by software specialists and may cause considerable miscommunication and misunderstandings with respect to the data.

One data composition (or report) should include one type of data, for example, the data of purchase transactions, sales transactions, or wages. If is not reasonable to request this data in one 'report', as this would make the logic of the data verification which will be used to verify the data received very complex. The reporting in which each taxonomy reflects a specific field or type of data is more reliable and easier to introduce. Thus, not only the structure and content of the data in one data composition is important, but also the link between the reports and the division of the data between reports.

[9] https://www.sm.ee/sites/default/files/content-editors/Ministeerium_kontaktid/Uuringu_ja_analuusid/nt_aruanne_final_3.pdf

2.2.1. Process of describing data composition

A description of data composition is a data model which includes the descriptions of data objects, their attributes, and the links between the objects. In order to compile such model, the works described below must be performed. The specialist in charge for detailed modelling of data must have appropriate education or must have passed a data modelling training. This is often disregarded in different sectors, which results in duplicating data or even a confusion which prevents using the data. A one-dimensional list of variables cannot be treated as a description of data composition as, as a rule, it does not enable identifying the structure of data objects, incl. the links between data objects. It is also important to describe the data objects and the links between the data objects or the hierarchy for the model to be comprehensive from the logical perspective and clear for the consumer of the data.

The figure below (see figure 2) presents the conceptual process of describing data compositions. It is important to highlight that it may be necessary to repeatedly return to the previous step in the course of describing data composition. This is most likely to happen after clarifying the links between data objects, but may also occur in other stages.

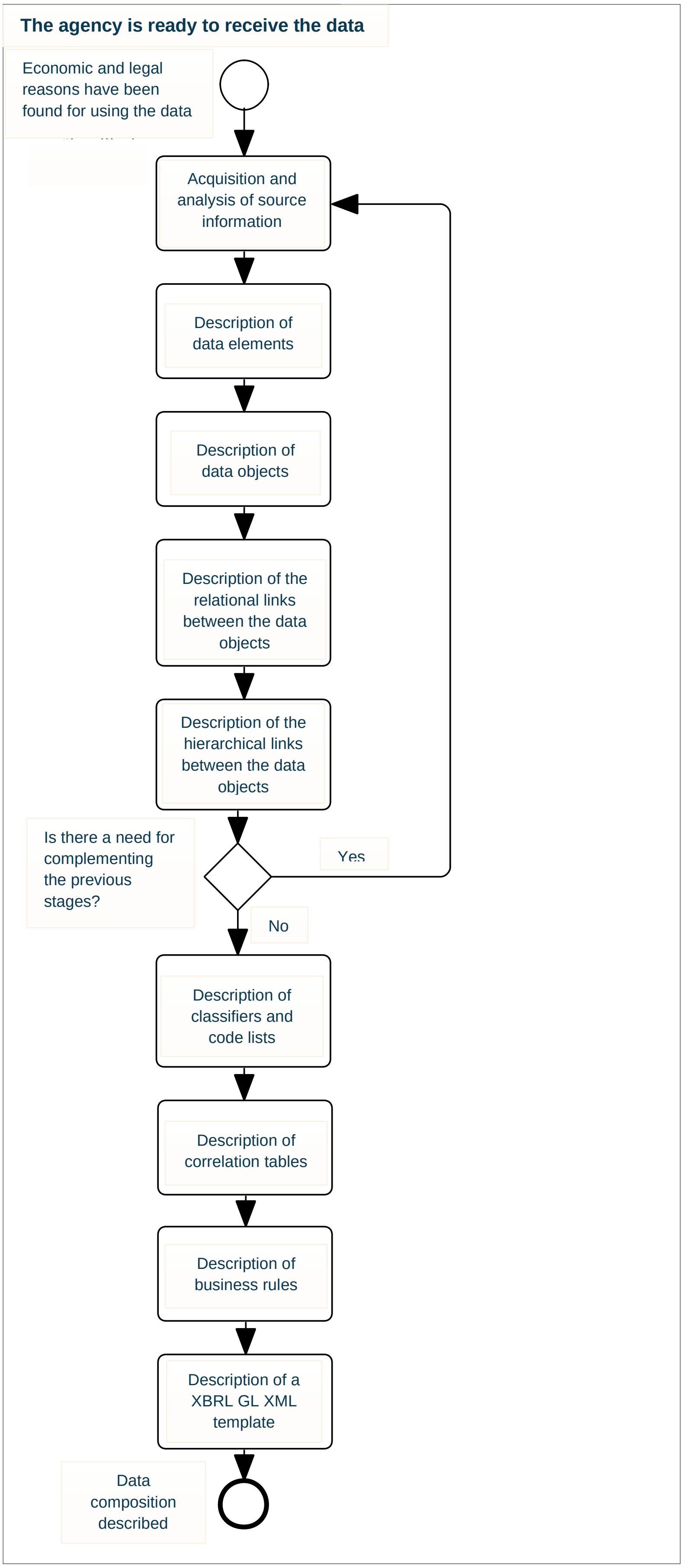
The agency is ready to receive the data
Economic and legal reasons have been found for using the data
Acquisition and analysis of source information
Description of data elements
Description of data objects
Description of the relational links between the data objects
Description of the hierarchical links between the data objects
Is there a need for complementing the previous stages?
Yes
No
Description of classifiers and code lists
Description of correlation tables
Description of business rules
Description of a XBRL GL XML template
Data composition described

Figure 2. Description of a data composition

2.2.1.1. Acquisition and analysis of source information

Before describing data composition, it is necessary to examine the potential data sources or the most important sources where the data will be originating from. The existing data sources should be submitted to the following assessments:

1) Are the descriptions of the data of a sufficient quality (incl. objects, attributes, and links)?
2) For which data objects there are actually data available in the data source?
3) Is there information on the data collection process as a result of which the data were placed in the data source? Is it possible to determine the origin and reliability of the data?
4) How many data objects are there (number of entries); which data need can they satisfy? The consumer of the data should possess the computing capacity for processing the data arriving from the source.
5) What is the quality of the data:
    a. How big is the share of missing values among important data elements?
    b. How frequently do the values occur, i.e. what types of data are available?
    c. Which classifications are used in the data source? Are the data classified in this manner usable?
    d. Is it known when the data were created?
    e. If the values are quantitative, is the unit of the values known?
    f. Are the dates specified in the data reasonable, incl. are the dates of birth and death compatible?

2.2.1.2. Description of data elements (variables)

In order to describe a data element, information must be obtained on the potential semantics and form of the element in the data source. If the data source exists, the description of the element should be aligned as much as possible with what can be obtained from the data source in the original or transformed format. If there is no data source, the elements may be described exactly based on the data need. This applies if the data source is created. In describing a data element, the actual possibilities of what can be achieved by the data source must also be taken into consideration. For example, if the data source is activated at a certain point in time, the data source cannot be expected to be usable for obtaining the data which were created before the data source in the case of somewhat more voluminous data.

2.2.1.3. Description of data objects

Data objects are described based on data needs. A data object is a collection of data elements. Data consumers are not required to use the same data groups (objects) which can be found in the data source. For example, if the data source is citizen data, incl. births, deaths, places of residence, etc., the consumer of the data may only be interested in the places of residence of the citizens. In this case, the data are transformed for the consumer so that the place of residence is the data group and the names of the citizens, etc. are not transferred at all.

Thus, keeping in mind the data need, data objects should be described based on the needs of the consumer of the data.

2.2.1.4. Description of the relational links between the data objects

A data object is a concept that refers to the grouping of data elements and there are, as a rule, links between different groups in databases. The links may be hierarchical or relational. The difference between the two links is that in the case of hierarchical links, the ‘child’ cannot be related to several ‘parents’. Relational datasets, however, allow this. In the case of the data with hierarchical links, it is always necessary to specify whether the correlations are relational at the logical level, i.e. whether there are cross-references between the objects. The presence of such cross-relating must always be described, explaining what should be done with any duplicates if the data thereof differ. Duplicates are created in a hierarchical data structure if one data object is related to several branches of the hierarchical tree. All duplicates must be identified in the course of data analysis and they may not be counted as several objects, especially in the case of compiling statistical data on objects.

Thus, the connections between objects must be described clearly, including specifying whether one object is related to one or several other objects or if several objects are related to several other objects.

2.2.1.5. Description of the hierarchical correlations between the data objects

In the case of handling data as XML, the level in the hierarchy of an element or object and the object in which the element or object is located are very important. In the case of a list without connections, the structure of elements often remains unclear. It is important to know this to be able to identify what the element provides data about. An element of the same type and name may be found at several different levels and the meaning thereof depends on the level which it is at. For example, let us examine two elements of id:

1) Invoice/cac:TaxRepresentativeParty/cac:PartyIdentification/cbc:ID.

2) Invoice/cac:PayeeParty/cac:PartyIdentification/cbc:ID.

If cbc:ID is referred to in the data composition, the consumer of the data may not understand to which object the id is related. In this case, it may be related to the tax representative or the recipient of the payment. In the example above, the description of the element path is as precise as possible. In the case of an element referred to in this manner, the place of the element in the hierarchy shows clearly what it is. All hierarchical connections must be described in a similar manner.

2.2.1.6. Description of classifications and code lists

In describing classifications and code lists, it should be presumed that the list in question is already in use somewhere. If the list exists, the same list must also be used in the new dataset created and, if there is a need to change the list, for example, to remove or add elements, a new version of the list must be created. The new version must include as much data as possible from the existing list (same element codes, same names).

2.2.1.7. *Description of the correlation tables of classifications and code lists*

In order to use an existing classification or list in a reduced or changed form, the new version of the list and the link of the version with the existing classifier must be described. The correlations must be indicated as element code pairs between the elements of the most detailed levels of the new list and the previous version of the list. If the link of the elements is 1 : 1, indicating one pair suffices. If the link is n : n or n : 1 or 1 : n, the correlation algorithm must also be described, i.e. the conditions on which to use one or another of the links. The conditions must be defined against the attributes of the classification element. This also means that in the case of more complex correlation tables, it is also necessary to define additional attributes of the elements of the classifications and lists.

For example, a correlation table must be complied if one agency wishes to use the classification of counties so that the objects of Tallinn are under Harju County, while another agency wishes to treat Tallinn as a separate administrative unit. In the latter case, there is one extra element in the classification of counties (Tallinn). In the correlation table, it must be indicated that Tallinn = Harju County if the exact address is within Tallinn.

2.2.1.8. *Description of the business rules*

Business rules must be described based on the needs of the consumer of the data. Business rules must be defined for the data which are important for the consumer of the data. Business rules are often also referred to as data checks or simply as checks. The steps for defining business rules include:

1) describing in the rules checks of what is mandatory, indicating which data elements must always be included in the data;
2) describing the checks of the value ranges and the scale;
3) describing the checks of the use of classifications (is the value of the element among the elements of the classification);
4) describing the cross-checks involving several data elements or objects;
5) creating the schematron of data checks;
6) testing the checks on actual data;
7) supplementing the descriptions of the checks based on the results of testing.

2.2.1.9. *Drawing up of XBRL GL XML templates*

Describing data model is always accompanied by the issue of the data consumer not getting enough information from an abstract model. Thus, sample data are also required in addition to the model. Sample data may be added to the descriptions of attributes, but an XML template provides the most expressive overview, especially in the case of the data in the XML format. The template must describe the data of a phenomenon which is likely to actually occur. XBRL GL XML samples can be found from the documentation of ‘Aruandlus 3.0’[10].

2.2.2. Outcomes of describing the composition of data

Below, we describe the structure of a description of data composition.

2.2.2.1. Description of data objects and the links between data objects

The model of data objects or links of data objects is drawn up as a conceptual XSD or ERD **model**. It is advisable to use the XSD model if the data can be placed in a data structure with hierarchical links. In the more complicated cases when there are cross-links between data objects, i.e. the model in question is a relational model, an ERD-type data model must be used. In the case of more complex models, this is essential to ensure that different parties have a common understanding.

2.2.2.2. Description (attributes) of data elements

Data elements are described so that the following data is provided about the attribute:

1) the code;
2) the name;
3) a verbal description which explains the meaning and content of the attribute;
4) the sub-type of the attribute;
5) the mandatory nature of the attribute which shows whether the attribute must always have a value or the value field may be left blank;
6) if the attribute must have a value from a classification or list, the description must specify which classification or list must be used. The code of the classification or list must be specified;
7) the name of the XBRL GL element matching the element described.

2.2.2.3. Description of classifications and lists

All classifications and lists used in the model (referred to under the attributes) must be described by the model. If a classification is exclusively only characteristic to the specific model, all elements of the classification must also be provided in the description. In the case of commonly used classifications and lists, the description must indicate which source the classification is from and where a machine-readable file of the content of the classifier could be obtained from or the online service issuing it.

The data structure of a classification or list must be aligned with the data structure of the classification provided in the inter-agency taxonomy (see the taxonomies of ‘Aruandlus 3.0’[11]).

[10] https://www.stat.ee/et/aruandlus-30

[11] https://www.stat.ee/et/aruandlus-30

2.2.2.4. Description of correlation tables

A correlation table consists of entries which include:

1) the classification code and element code;

2) the code of the respective classification and the code of the respective element;

3) description of the correlation rule.

2.2.2.5. Classification repository and the online services thereof

For the data submitter to be able to keep themselves posted about any changes in the classifications and lists, i.e. to ensure the quality of the data submitted, the information in question must be accessible by the system of the data submitter in a machine-readable format. A suitable sample solution is the classification system of Statistics Estonia which can be used by anyone interested.

2.2.2.6. Description of business rules

For a data submitter to be able to tell which rules the data must comply with, the party requesting the data must also describe respective business rules and this information must be made available to the system of the data submitter in a machine-readable format.

The descriptions of all business rules must include the following details:

1) the code of the business rule;

2) the control formula or the description of the algorithm;

3) the name of the data element checked (x-path) in XML;

4) the message which the system should issue of the data are not compliant with the business rule.

2.2.2.7. Samples of the data

Assuming that the data submitter is submitting data in the XBRL GL XML format, the data samples must also be in the format of an XBRL GL XML file. The samples must be validated against XBRL GL XSD files. If the content of the data varies greatly and the hierarchy of the XML content may have mutually exclusive branches, more than one sample file should be drawn up. All sample files must be compliant with the business rules. It is not advisable to include mutually exclusive data in one file.

2.2.2.8. Machine-readable taxonomy

Machine-readable taxonomy consists of the following:

1) details of classifications and lists in the form of a machine-readable data repository. The data repository must be accessible via online services with REST API-based access without authorisation or authentication of the user;

2) a list of XSD data elements (XBRL GL or an extension thereof) and the elements of its sub-part used in the specific dataset;

3) the description of the business rules in the schematron format which is human and machine-readable and executable;

4) semantic descriptions of the data or the data model, preferably in the form of Enterprise Architect EAP or XMI 2.1 XML;

5) links between data elements and classifications.

2.3. Preparation of the systems of an agency for receiving data

For data transmission and receipt to function, the agency must configure and, if necessary, develop the systems based on the description of the data composition and the system architecture described in the roadmap. All components referred to must be developed and configured.

3. Process of submission and receipt of data in a state agency

Today, the majority of the data is received from state agencies via different portals, such as eSTAT, KOTKAS, ePRIA, e-MTA.

In order to automatise data submission, an electronic channel must be defined for submission of data. Defining the channel alone, however, is not sufficient. The process of data transmission must also be defined. For example, are the data submitted to the state agency or can the state agency request the data from the submitter? When do the data move and when and how is the first confirmation received about the arrival of the data? In addition, how to obtain the secondary confirmation of the arrival of the data, i.e. the validation confirmation? Is the entire process asynchronous or synchronous?

We will find answers to these questions in the following sub-chapters.

3.1. Process of data submission in the case of the push method

Taking into consideration the goals of real-time economy, the data submission process in the case of which the data are transferred from the software of the data submitter to the software of the recipient of the data by an electronic channel is discussed next. This section describes the process in which the data submitter sends the data to the state, i.e. the so-called push data transmission. There is also a process in the opposite direction where the state agency requests the data from the system of the data submitter. This process is referred to as pull data transmission. In this case, the state agency can control the data transmission, which may be the best option for the agency. The agency can control its own load.

The steps of the process of submission and receipt of data include the following:

1) the system of the data submitter puts together the dataset to be sent;

2) the system of the data submitter checks the dataset against the business logic checks which have been defined for this dataset;

3) if there are any shortcomings, the data submitter will fix the shortcomings of the data;

4) the system of the data submitter sends the data to the state agency via an electronic channel (X-Road);

5) the system of the state agency accepts the data and saves them in the form in which they arrive;

6) the system of the state agency automatically checks the compliance of the structure of the data;

7) the system of the state agency provides initial feedback about the appropriateness of the data to the system of the data submitter;

8) if the structure was appropriate, the business logic check of the data is performed at the state agency by the system as well as by an official, if necessary;

9) the system of the state agency sends the result of the data check to the system of the data submitter;

10) if there were any deficiencies in the data, the data submitter will make the necessary changes in the data and the data submission process is repeated.

3.2. General data submission process by using the pull method

If necessary, the so-called pull method may also be used for submitting data, in the case of which the data are not submitted by the data submitter, but the state agency request the data from the data submitter's respective online service. Such reporting may also be referred to as real-time monitoring which basically enables real-time assessment of the content of the data, as well as the quality indicators. This method may also be referred to as monitoring where the data submitter must only prepare the systems for the data to be submitted to the state. This method is, for example, used by the Agricultural Registers and Information Board (ARIB) to check mowing of fields, with information from satellite photos obtained from EstHUB (a satellite information database) and processed by the ARIB which enables identifying whether or not a field has been mowed by the required date. The recipient of mowing support is only contacted for further information if there are any doubts.

The steps of the data transmission process based on the pull method include the following:

1) the systems are put together in the system of the data submitter;

2) the system of the state agency requests data from the system of the data submitter as and when needed;

3) the system of the state agency accepts the data and saves them in the form in which they were requested;

4) the system of the state agency checks the compliance of the structure of the data;

5) if the structure was appropriate, the business logic check of the data is performed at the state agency by the system as well as by an official, if necessary;

6) the system of the state agency sends the information about the quality of the data to the system of the data submitter;

7) if there were any deficiencies in the data, the data submitter will make the necessary changes in the data in their system.

4. Process of data verification and processing at a state agency

When the data has arrived at a state agency and the structure of the data has been deemed compliant, the state agency must check the content of the data. Among other things, the check involves verifying whether the data have been classified by using the classifications included in the taxonomy, whether the data include all the necessary values, whether the values are mutually aligned and compliant with the business rules, and whether the values are within a proper range. Another important fact to determine is whether the quality level of the data calls for the data to be resubmitted. Any issues detected must be recorded at the state agency in a manner which enables reproduction thereof (electronically in a database) so that the data submitter can be notified of the issues automatically and repeatedly, if necessary, to achieve better quality of the data.

The need to update data by the sender or the recipient may also arise in the course of checking and processing data or the recipient of the data may adjust the data, having, for example, received additional data from further sources without requesting them from the data submitter. An opposite case (the data recipient transfers a change to the data submitter) occurs if the Tax and Customs Board changes the amount of the tax payable so that it is different from what was declared by the data submitter.

4.1. Verification of the data

The process of verifying data at a state agency could, in principle, consist of the following:

1) transformation of the data received in the system to a format suitable for use at the agency;
2) subjecting the data to business logic checks;
3) analysing the data and creating a test output of the data;
4) drawing up further advice and questions to the data submitter based on the results of the analysis;
5) making the results of the checks of the data by the state agency accessible to the data submitter, including indicating whether or not the data submitter must resubmit the data;
6) if the state agency requests resubmission of the data, the data submission process is relaunched.

4.2. Data processing and analysis

State agencies have specific sector-based information systems for processing data which are designed to satisfy specific needs. Data processing, incl. any datasets submitted by the data submitter at a later date, may be used in various very different processes (tax collection, inspection of compliance with the requirements of an environmental permit, etc.). It is not possible to align all those processes and information systems. Only the input and output which are related to those systems are important for the data submitter. Thus, this document only discusses the part of those systems which is visible to the public and mainly designed for interfacing.

If the data submitted by a data submitter are only used for data analysis or for statistical purposes, the Generic Statistical Business Process Model (GSBPM[12]) should be used.

[12] https://statswiki.unece.org/display/GSBPM

5. Process of reflecting data

It is important to provide feedback to the business about the manner of using the data received from them and the new indicators calculated about the business at the state agency. This is especially the case if the data were processed further, i.e. classified further, reclassified, or if some facts have been changed, having obtained further information from other data sources. Reclassification means, for example, that if a business submits the data on their transactions taxable with 0% value added tax, but the Tax and Customs Board finds that the transaction should be taxable at 20%, this information must be communicated to the business. This determines the amount of the tax payable. It is very important to reflect the tax balances of a business from the perspective of tax returns as well as payment of taxes. In the case of the Tax and Customs Board, this has currently been solved in the e-MTA environment, which users or persons authorised by them can use to view the balances, as well as individual records.

5.1. Types of reflected data

The following types of data are reflected:

1) Issuing of information about the outcomes of the data format check. In the case of the data submitted in the XML format, data format checks are performed with the help of XSD. The XSD which XML must be compliant with includes a defined data structure and certain restrictions (e.g. the mandatory nature of an element).

2) Issuing of information of the business logic checks of the data. The business rules against which data are checked must be known in advance and clearly described based on the structure indicated in the chapter about data composition. In this case, the data can be submitted in a structured and machine-readable format. The structure and machine-readability of data is one of the important foundations of real-time economy.

3) Issuing information about the expert assessment based on detailed quality analysis of the data. This step should only be performed if previous automatic checks have revealed a need for more detailed quality assessment.

4) Issuing of information about changes in the data made by the state agency. An example of this is a situation in which the Tax and Customs Board reclassifies the amount taxable and issues respective information about changes made in the tax return.

5) Issuing of information about the aggregated indicators formed based on the data. For example, issuing of the value added tax balance of a business which is currently solved in the e-MTA environment, but there is no machine interface.

The data reflecting process must be built so that it is possible for the business to request feedback data at any time. Thus, it cannot function so that the information is sent somewhere by the state agency. A respective X-Road service must be created for the business, through which inquiries can be made for all types of feedback specified above.

5.2. Process of reflecting

The general process of reflecting consists of the following:

1) The data submitter requests information from the state agency through their system, providing the following data as input:

    a. details of the person submitting the request;

    b. details of the business which the request is concerned with;

    c. the type of feedback requested;

    d. the data for which the feedback is requested.

2) The state agency accepts the request and verifies the rights of the person who submitted the request to receive the data.

3) If the person is entitled to the data, the system of the state agency draws up a response and sends the response automatically to the person who submitted the request.

6. Reuse of data by different state agencies

Reuse of data can be divided in two:

1) use of numerical data;
2) use of personalised data.

This document only discusses the reuse of personalised data, as a numerical data portal is already in use in Estonia and all agencies and businesses can make their numerical data accessible via the portal.

If data processing involves personal data, the General Data Protection Regulation of the EU must also be taken into consideration[13]. The principles of data protection must be applied to any information concerning an identified or identifiable natural person (GDPR does not apply to the data of legal persons or agencies, i.e. it apples to personal data in the data-based reporting model). It should also be kept in mind that data exchange between state agencies is subject to a provision of the Public Information Act, which requires the use of the data exchange layer of the state information system (X-Road) for the data exchange between the databases included in the state information system.

6.1. Prerequisites for sharing personalised data

The following conditions must be met for reuse of data by another agency:

1) there must be a legal basis or the permission of the owner of the data for using data, unless the data are public data;
2) the party sharing data must be capable of making the data available and keeping the data available for the required period of time;
3) the party sharing data must ensure the authorisation and authentication of the consumers of the data;
4) the party sharing data must ensure logging of the use of the data. The log must contain information about who, when, and where form requested the data and which data were requested;
5) further consumers of the data must refer to the original source of the data in all of the analysis published by them and in other works in which the data are used.

When sharing data, it should be kept in mind that the law establishes specific restrictions for data sharing. For example, Statistics Estonia may only disclose personalised data to other agencies or persons for statistical or scientific purposes.

[13] https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en

6.2. Temporary sharing of data

The process of temporary reuse of personalised data, which may include personal data or business secrets, includes the following steps:

1) a state agency or business submits to the data controller (a state agency) a request to obtain the data collected by them;

2) the data controller verifies the rights of the person who submitted the request to use the data;

   - if the person who submitted the request is not entitled to use the data, the data controller rejects the request and the process is closed;

   - if the person who submitted the request is entitled to use the data, the data controller enters into a confidentiality agreement with the person who requested the data for using the data;

3) the data controller will make the data available to the person who requested the data, including granting a licence for using the data (the terms and conditions for use and deletion) in their computer network by configuring a secure connection for the person who requested the data or by ensuring a possibility to use the data in the premises and in a computer of a data controller or in another secure manner which protects the data from leaks.

businesses as well as state agencies may be entitled to use data. The right to use data may arise on different grounds, including, for example, for a business if the business is performing a task received from the state within the framework of public procurement which is related to using the respective data.

6.3. Permanent need to share data

If the reuse of data is of a permanent nature, the data submitter should also be placed under an obligation to submit data to the other agency which will be using the data. If the composition of the data remains the same, this will not cause a significant additional burden for the data submitter, as no further system development is needed for compiling the data, but the data submitter can send to the additional agency a copy of the data or part of the data, if needed.

7. Technological environment for processing and modelling of data

Data exchange:

1) X-Road, XML (for sending personalised data);
2) REST API, JSON (for sending open data).

Reporting service providers/environments:

1) ERP systems;
2) transmission service providers (such as MCDS);
3) other specific systems in which reporting data may be recorded.

Data modelling:

1) XML and XSD redactors (Altova XML Spy, Oxygen or another tool with the capability to draw up and validate XML and XSD files);
2) UML, ERD (Sparx Enterprise Architect which has *de facto* became the standard or another tool which supports model exchange in the XMI 2.1 format).

Process and algorithm modelling:

1) BPMN (Sparx Enterprise Architect, Bizagi).[14]

Data processing:

1) The tools used in the information system of the agency.

[14] https://www.mkm.ee/sites/default/files/protsessianaluusi_kasiraamat.pdf

Summary

This document provides a conceptual model of the structure of the reposting system used by businesses to report to the state. The model is primarily based on the assumption that the same approach should be used in the case of all reports submitted to the state regardless of the sector. All sectors differ by the logic of their operations, but they could look the same from the perspective of reporting. This would make it easier for businesses to develop interfaces and automatise their reporting. In the future, the system of submitting reports may be replaced by a monitoring (or screening) system in which the data only move in real time without any manual work from the data submitter or the recipient in the processing of the data. For this to become possible, the technical as well as business rules must be established in detail and agreed on. This will allow for the movement of data without any further input from an employee of the data submitter.

The authors of this model hope that they have made a small contribution – but with a great positive effect – towards lowering administrative expenses and promoting real-time economy in the Estonian reporting landscape by creating this description of a potential model.