# Harmonized structural metadata for faster and more coherent output

Piret Piiroja, Statistics Estonia, piret.piiroja@stat.ee

Tuuli Mei, Statistics Estonia, tuuli.mei@stat.ee

**Abstract**

*In the modern world, people expect information to come to them easily and fast. If we want official statistics to be able to compete with all other sources of information, it has to meet the needs of the consumers.*

*In order to make our products more user-friendly we should, in addition to implementing better visualization techniques and disseminating our products via channels that are readily accessible to users, make the information easy to comprehend and compare. Meta-information plays a vital role in this. By using harmonized metadata in our output, we make our statistical products more coherent and easy to use for all consumers while optimizing our own work processes.*

*In Statistics Estonia, we have started converting from SAS to R. While we create R-scripts to produce output based on the data in the central database, we also implement Eurostat’s standard code lists and SDMX standard in our output metadata whenever possible. The aim of this is to make our statistical products comparable to each other as well as to the statistics found in the Eurostat database. Having the standard code lists inserted into our meta-information system makes it possible for us to easily reuse them when necessary. This makes the production of new statistical products faster while making the new product coherent with the rest of our products.*

*The more countries that use standard code lists for their output, the more convenient it is for the consumer to compare statistics published by different countries across the world. Although the harmonization of metadata may seem unnecessarily resource consuming, once it is done, the gain for the user is obvious.*

**Keywords**: Structural metadata, standard code lists, output, harmonization

## Introduction

In the modern world with its abundance of information, the ability to find relevant and reliable data is becoming increasingly important. While information is available from many different sources, few are as reliable as the agencies that produce official statistics. The Marketing and Dissemination Department at Statistics Estonia has found that while the users of official statistics appreciate the scope of the information offered by Statistics Estonia, they frequently consider the relevant information difficult to find. There are several ways to make information easier to access and understand, including the use of different dissemination channels and better visualization techniques. These methods have received more attention in recent years with the role of social media becoming increasingly vital in commerce as well as in the dissemination of any information and undoubtedly help to make large amounts of data less intimidating and more user-friendly. However, next to the modern methods of dissemination, the importance of metadata sometimes seems to be underestimated. Yet, metadata is one of the most important means that helps to make sense of the large amount of information surrounding us every day.

In order to make the most of metadata, the metadata should be harmonized within one agency as well as between different agencies. Structural metadata needs to be clear and concise to ensure that the user understands the data presented to him. Yet, even clear and concise metadata might not be sufficient if the user needs to compare different datasets. Whether the datasets are published by one agency or different national agencies, or national agencies and one or more international organizations, comparing data can be a serious challenge. The user has to ascertain that the variables or categories that are named the same are indeed the same. It may take extensive knowledge of the subject matter to realize that categories named differently refer in effect to the same concept. Such discrepancies in the presentation of data can make comparisons between datasets extremely problematic and cause misunderstandings and mistakes in interpreting the data. This in turn could damage how the users perceive the reliability of the data provided by the agencies and the reputation of the agencies as sources of reliable data.

While the need for standards for reporting metadata has been recognized for a long time and steps have been taken to create those standards, the process is slow and concentrated more on how to report data to Eurostat and other international organizations. Yet, it is just as important that the data published by a national agency were coherent, easy to understand and compare. For that reason, there also needs to be harmonization of metadata within the national agency. The difficulties that users face when trying to compare statistics between countries or even within one country have been acknowledged before (OECD 2007, p. 10-11). Regardless of the efforts made to solve the problem, the situation is still far from ideal.

## Metadata management in Statistics Estonia

### Centralization

Statistics Estonia has approached the challenge of harmonizing metadata by centralizing metadata management. In the Statistics Design Department of Statistics Estonia there are two teams dealing mainly with metadata management. The metadata team handles register variables (variables created based on the data collected) and classifications while the output team concentrates on cube variables (variables published in data cubes or sent to international organizations) and standard code lists.

By centralizing metadata management, we gain a better overview of which classifications and standard code lists are used within our organization, which variables are used and how the codes are structured and the variables and elements named. The disadvantage of this approach is that while the people responsible for metadata are well informed of the metadata used across all domains within the organization, they are not experts in the particular fields of study. This calls for close cooperation between the statisticians who know their field and the metadata specialists.

### The output team

It is the task of the output team to ascertain that the metadata used in different datasets published by Statistics Estonia is harmonized. Also, when Statistics Estonia publishes the same data that is collected and published by Eurostat or other international organizations, the output team ensures that the codes and names for variables and the standard code lists used match those used by Eurostat. This makes it possible for the user to more easily compare different datasets published by Statistics Estonia as well as allow the comparison of datasets published by Statistics Estonia with datasets published by an international organization.

While metadata for the data collected has been described in Statistics Estonia’s meta-information system for years, the describing of output metadata only seriously started in 2017 with the creation of the output team. Until that time, output metadata existed in Excel files of the statisticians and on Statistics Estonia’s website. It was not harmonized across domains and sometimes not even across different datasets within one domain. By 2018, the team had established its work processes and started the task of describing structural metadata for output in the meta-information system and creating R scripts for producing the output data cubes. However, a lot of work still remains to be done.

When Statistics Estonia decided to replace PC Axis software with .Stat to publish statistics, the decision was also made to redesign the data cubes. Along with the redesign of the cube structure, we started to harmonize the metadata and implemented standard code lists. All redesigned output data cubes are also described in the meta-information system. While currently the metadata for the published data cubes is manually retrieved from the meta-information system, in the future this process is intended to be automatic.

Statistics Estonia’s development plan for 2018-2022 (Statistics Estonia 2018, p. 19) stipulates a decrease in the number of software used in the organization as well as reduction in the license cost by replacing expensive or outdated software with open source solutions. As part of this process, previously used analysis software such as FoxPro and SAS are being replaced with R. For all redesigned data cubes there is also an R script created, which compiles the cube using the metadata retrieved from the meta-information system and calculates the data based on the data files in the Final Observation Registry. In order for the R scripts to work, the metadata for the data cubes must be described in the meta-information system.

### Principles of harmonization

In its effort to harmonize metadata, Statistics Estonia attempts to use the SDMX standard whenever possible. Sponsored by seven international organizations, including Eurostat, the SDMX standard was created specifically for the description and exchange of statistical data and metadata. Data structure definitions (DSDs) which describe the structure of a dataset and the standard code lists used in it have already been developed for several domains and the work is ongoing. Whenever Statistics Estonia publishes data belonging to a domain that has a DSD, we use the DSD to describe the metadata in our meta-information system. The existence of the DSD implies that an international organization collects data about that domain from national agencies and the DSD has been developed in order to standardize the structure of the exchanged data. Since the data has to be sent in a format compliant with the DSD, the metadata will be described in that format. If the same data is published by Statistics Estonia, the metadata only needs to be described once, reducing the time and effort spent on the description of the metadata.

However, since DSDs are mainly compiled with data exchange in mind, the international organizations collecting the data from national agencies do not always use the same codes to publish the collected data. While the metadata prescribed in a DSD are compliant with the SDMX standard, the code lists used to publish the data do not always follow the standard. This creates a dilemma for the metadata describers in Statistics Estonia. On the one hand, we would like to use the SDMX standard to describe our metadata, while, on the other hand, we want our metadata to be harmonized with the metadata published by international organizations to make it easy for the user to compare.

National needs differ from international ones and a national agency often has to publish either more detailed information than it sends to international organizations or publish data that has been structured in a different way. In that case, harmonization with metadata published by international organizations is not possible or can be done only partially. In the latter case, the datasets published by the national agency and the international organization are not identical but the overlapping part is still harmonized. Thus, on the webpage of the national agency, the user might encounter datasets that contain some variables familiar from the datasets published by Eurostat while other variables look different. Like variables, the standard code lists used by the international organization are also adopted for national purposes when possible, with additional elements, if necessary, added for national use only.

Some of the harmonization takes place only within Statistics Estonia. When the data published by the national agency differs considerably from that published by Eurostat or other organizations there are variables and standard code lists created for national use only. Occasionally such code lists are loosely based on standard code lists used internationally, in other cases they may be completely unique, with no similar code lists available internationally. However, even when a code list or a variable is created for national use only it should be used consistently in different data sets, where applicable. In accordance with the intention to use the SDMX standard whenever possible, when creating code lists for national use, we follow the guidelines established by the SDMX community (SDMX 2018).

In conclusion, while the ideal is clarity and accuracy, the actual state of affairs is a lot more complicated. There are similar but slightly differing code lists for national and international use, code lists containing identical categories with different codes for the SDMX-compliant and non-SDMX-compliant code lists, identical categories with different codes for data transmission and publication, etc. The task of harmonization requires constant negotiation between national and international needs. When the necessity for using different metadata for national publications is well justified, it overrides the need for harmonization. On the other hand, if the main reason for doing things differently on national level is that it has always been done that way before then the need for harmonization should prevail. Regardless of the difficulties, the harmonization of metadata is a worthwhile endeavour and with DSDs slowly but surely becoming available for additional domains, the situation is improving and metadata harmonization is becoming easier.

## The benefits of harmonization

### Coherence of output

The task of metadata harmonization may seem daunting but there is much to gain if it is done well. As mentioned before, the first and most obvious benefit is to the user of statistics. It is noted in the OECD Data and Metadata Reporting and Presentation Handbook (OECD 2007, p. 10) that even data that is essentially the same may look different when published by different organizations. The same is true for data published by Statistics Estonia. Much of the data published on our website is also published by Eurostat, but the data cubes often look different. While Statistics Estonia does not endeavour to structure their data cubes exactly the same way as Eurostat, the names of the variables and the elements of standard code lists used in the data cubes should match with those published by Eurostat if the data is indeed the same.

Without someone to oversee the harmonization of metadata across different subject matter domains, even datasets published by the same national agency may contain code lists that are identical in nature while looking different. For instance, if each statistician uses a slightly different wording to refer to the same economic activity, it might not be easy for the user to realize that it is in fact the same. It takes someone with a wider view of metadata used across all domains to notice and question such differences between different datasets.

The ultimate goal of metadata harmonization is to achieve a more coherent output. By managing all structural metadata used in output by one team, we can ensure that the datasets the user sees on our webpage are provided with coherent metadata and easily comparable to each other, as well as to similar datasets published by international organizations such as Eurostat. By doing so, we improve the user’s understanding of the statistics we publish and increase both our reliability and popularity among the users.

### Optimized work processes

While metadata harmonization is mainly motivated by the needs of the user there are also benefits for the process of preparing output. With standard code lists inserted into the meta-information system and readily available for further use, the preparation of output for new data cubes becomes faster. Instead of having to create new code lists for each new dataset, the existing code lists can be used. Even the exact code sets used in one dataset can sometimes be re-used. All subject matter domains naturally also have specific features where cross-domain standard code lists are of no use. However, related fields have previously often used slightly different classifications in their data cubes. With the implementation of standard code lists and centralized handling of metadata, such duplication of work will be avoided, reducing the time spent on the preparation of output.

## Conclusion

Harmonization of metadata is unquestionably a difficult and time-consuming task that requires extensive knowledge of both the subject matter domain as well as metadata description standards. In Statistics Estonia, we have approached the challenge by dividing the work between statisticians who are experts in their particular domains, and metadata specialists who have a better overview of metadata standards and existing metadata descriptions across all the domains. By combining the two specific fields of knowledge in the cooperation between the statisticians and the metadata specialists, we hope to achieve the best results in metadata harmonization and provide the user of official statistics with data that is easy to understand and compare. As an added benefit, once the standard code lists have been added to the meta-information system, they are easy to re-use, thus making the preparation of output faster.

It is the case with standards that the more extensively one is used, the greater the benefits. The same can be said about metadata harmonization. The more countries that use standard code lists and harmonize their output metadata with that published by international organizations and other countries, the easier it will be for users to find what they need and compare data published by different organizations and agencies. Since the needs of the users are undoubtedly important for all statistical organizations we should all strive to achieve better results at harmonizing the metadata we publish.

## References

OECD (2007), Data and Metadata Reporting and Presentation Handbook, pp. 10-11

SDMX (Statistical Data and Metadata Exchange) (2018), Guidelines for the Creation and Management of SDMX Code Lists. Available at: <https://sdmx.org/?page_id=4345/> (Accessed 15 May 2019)

Statistics Estonia (2018), Development Plan of Statistics Estonia. Available at: <https://www.stat.ee/strategy> (Accessed 14 May 2019)